バイオインフォマティクス人材育成講座

平成 24 年度

スタンダードコース講義集

株式会社リバネス

目次

1　はじめに　5

2　カリキュラム一覧　6

※バイオインフォマティクス・スタンダードコース（ＷＥＢサイト）　8

3　講義集　11

3.1　バイオインフォマティクスの歴史、生命科学の基礎　11

3.2　生命科学実習①：ゲノム DNA 抽出、総 RNA 抽出　16

3.3　生命科学実習②：PCR、DNA シークエンス　22

3.4　生命科学実習③：トランスクリプトーム解析　29

3.5　生命科学実習④　35

3.6　情報科学基礎　42

3.7　類似性発見　49

3.8　配列と機能　54

3.9　タンパク質の機能と構造　59

1　はじめに

バイオインフォマティクス人材育成講座 平成 24 年度スタンダードコース講義集は、公益財団法人沖縄科学技術振興センターが、平成 22 年度より 3 年計画で実施した「バイオインフォマティクス人材育成推進事業」において開発したスタンダードコースカリキュラムをもとに、平成 24 年度に国立学校法人琉球大学の農学部食品機能科学特別講義Ⅰおよび同じく工学部情報工学科特別講義Ⅴにおいて実施された講義を集めたものです。

本講義集では、バイオインフォマティクス人材育成推進事業が目的とする高等教育機関における育成基盤構築の一助として、実際に行われた 30 講義（1 講義あたり 90 分授業）に関する情報を、講義スライドや実習プロトコルなど講義で用いられた資料も合わせて収録しました。新たなバイオインフォマティクス人材の育成に活用いただければ幸いです。

なお、本講義集に収録されている講義スライド等資料の著作権は、各講義を担当した講師に帰属します。許可なく転載するなどの教育目的以外の二次利用はご遠慮ください。

2　カリキュラム一覧

食品機能科学特別講義Ⅰ　カリキュラム　2012年9月3日（月）～9月7日（金）

第1部　バイオインフォマティクスの歴史、生命科学の基礎

講師　要匡　准教授　琉球大学大学院　医学研究科　遺伝医学講座

日時　9月3日（月）10:20-11:50、12:50-14:20、14:40-16:10
場所　琉球大学農学部207講義室

概要　生化学、遺伝学、分子生物学など生命科学の発展に計算機が活用され、計算機の発展に伴ってバイオインフォマティクス分野が重要な役割を担うようになってきた歴史を学ぶとともに、細胞の構造や細胞内器官の働き、生体分子（DNA、RNA、タンパク質など）に関する生命科学の基礎的な知識と遺伝学の基本的な法則などを学ぶ。

第2部・第3部　生命科学実習①・②

講師　池松真也　教授　沖縄工業高等専門学校　生物資源工学科

日時　9月4日（火）10:20-11:50、12:50-14:20、14:40-16:10
9月5日（水）10:20-11:50、12:50-14:20、14:40-16:10
場所　沖縄工業高等専門学校　創造・実践棟　生物化学工学実習室

概要　生命科学実習①：生命科学研究において利用される基本的な実験技術として、細胞からのDNA抽出および制限酵素処理を加えて電気泳動で分離する処理について講義および実習を交えて学ぶ。生命科学実習②：ポリメラーゼ連鎖反応（PCR）と、DNAシーケンサによる配列解析について、その原理の理解を中心に実習を交えて学ぶ。

第4部　生命科学実習③

講師　門田幸二　特任准教授　東京大学大学院農学生命科学研究科

日時　9月6日（木）10:20-11:50、12:50-14:20、14:40-16:10
場所　琉球大学総合情報処理センター

概要　細胞における遺伝子発現の全体像を調べるトランスクリプトーム解析について、その実験技術であるマイクロアレイや次世代型シーケンサの活用について学び、サンプル間の発現変動遺伝子の抽出等、そのデータ解析の実例を実習する。

第5部　生命科学実習④

講師　田邊俊朗　准教授　沖縄工業高等専門学校　生物資源工学科

日時　9月7日（金）10:20-11:50、12:50-14:20、14:40-16:10
場所　沖縄工業高等専門学校　創造・実践棟　生物化学工学実習室

概要　機能を担うタンパク質の総体を取り扱うプロテオーム解析として、目的とするタンパク質の検出を行う活性染色法や、二次元電気泳動および質量分析によるタンパク質発現のサンプル間比較を実習し、タンパク質同定の技術について学ぶ。

情報工学科 特別講義Ⅴ　カリキュラム　9 月 10 日（月）～9 月 13 日（木）

第１部　情報科学基礎

講師　名嘉村盛和　教授　琉球大学　工学部情報工学科

日時　9 月 10 日（月）8：30-10:00、10:20-11:50、12:50-14:20、14:40-16:10、16:30-18:00
場所　琉球大学総合情報処理センター
※8:30-10:00および16:30-18:00は、岡崎威生先生が担当する。

概要　情報科学分野の基礎として、生命科学分野の研究者／技術者が修得すべきデータ構造とアルゴリズム、最適化計算の基本を解説と演習を通して学ぶ。

第２部　類似性発見

講師　岡崎威生　講師　琉球大学　工学部情報工学科

日時　9 月 11 日（火）10:20-11:50、12:50-14:20、14:40-16:10、16:30-18:00
場所　琉球大学総合情報処理センター

概要　データ解析の基礎となる確率論と統計的推論からスタートし、データ解析言語Rを利用した実習を行う。次に基本的な配列解析であるアラインメントの原理と方法を演習を交えて理解する。最後に類似性に基づいたクラスタリング手法について、適用ノウハウと供に学ぶ。

第３部　配列と機能

講師　有田正規　准教授　東京大学　理学部生物情報科学科

日時　9 月 12 日（水）10:20-11:50、12:50-14:20、14:40-16:10
場所　琉球大学総合情報処理センター

概要　DNAやタンパク質など生体分子の配列の相同性と機能について、BLASTによる相同配列の検索や構造ドメイン、二次構造予想など配列の二次構造予測の実習を行う。また、酵素や代謝、タンパク質の相互作用など、KEGGやSTRINGのWEBツールを活用して、それらのネットワークについて実習する。

第４部　タンパク質の機能と構造

講師　広川貴次　創薬分子設計チーム長　産業技術総合研究所　生命情報工学研究センター

日時　9 月 13 日（木）10:20-11:50、12:50-14:20、14:40-16:10
場所　琉球大学総合情報処理センター

概要　立体構造やドメインのデータベースについて学び、PDBでみる立体構造を実習する。またINTERPROによるモチーフ予測、膜貫通領域予測、触媒部位の予測等を実習する他、AUTDOCKによりリガンドのドッキングシミュレーション実習も行う。

※バイオインフォマティクス・スタンダードコース（ＷＥＢサイト）

農学部食品機能科学特別講義Ｉと工学部情報工学科特別講義Ｖの両方の講義を受講することで、

バイオインフォマティクススに関連する基礎的な知識と技術を身につけるコースとなります。

平成 24 年度沖縄県産業振興基金事業「バイオインフォマティクス人材育成推進事業」 の一環として実施されます。

講座概要 pdf

農学部　食品機能科学特別講義Ｉ

- 科目番号：資 535
- 単位数：2
- 開講日時：平成 24 年 9 月 3 日(月)～7 日(金)
- 亜熱帯生物資源科学科の学生：選択科目
- 他学科・他学部の学生：自由科目
- 登録方法：農学部学務係

工学部情報工学科　特別講義Ｖ

- 科目番号：情 495
- 単位数：2
- 開講日時：平成 24 年 9 月 10 日(月)～13 日(木)
- 教室：総合情報処理センター2 階第 3 実習室
- 情報工学科の学生：選択科目
- 他学科・他学部の学生：自由科目
- 登録方法：岡崎威生(okazaki@ie.u-ryukyu.ac.jp)まで下記の情報をメールで

  Subject：【特別講義Ｖ】受講登録、 学部学科：、学籍番号：、氏名：、農学部受講：登録/未登録
- 登録受付学生リスト(学内限定アクセス)

授業評価アンケート(農・工共通)

受講生は、担当教員毎に記入して下さい。

授業評価アンケート回答状況

- 農学部・食品機能科学特別講義Ｉ　回答状況
- 工学部・特別講義Ｖ　回答状況

バイオインフォマティクス技術者認定試験

- 平成 24 年度試験案内ポスター
- 平成 23 年度問題と解答
- 平成 23 年度解説
- 平成 22 年度問題と解答
- 平成 21 年度問題と解答
- 平成 20 年度問題と解答

- 平成 19 年度問題と解答

講義関連情報

農学部・食品機能科学特別講義 I

- 9 月 3 日 バイオインフォマティクスの歴史、生命科学の基礎(要匡)
  - バイオインフォマティクス pdf
  - 生命科学の基礎 pdf
  - 遺伝学の基礎 pdf
- 9 月 4,5 日 生命科学実習 1,2(池松真也)
  - ゲノム DNA の抽出 pdf
  - ゲノム DNA の調整 pdf
  - 総 RNA の抽出 pdf
  - 課題 pdf
  - 課題解答例 pdf
  - 電気泳動 doc
  - PCR による DNA 増幅 pptx
  - 遺伝子の増幅と配列分析法 docx
  - 乳酸菌のシークエンス doc
  - プライマー位置 ppt
- 9 月 6 日 生命科学実習 3(門田幸二)
  - 門田幸二のホームページ リンク
  - (R で)塩基配列解析ウェブページ リンク
  - (R で)マイクロアレイデータ解析ウェブページ リンク
  - 講義資料（公開版） pdf
  - 課題 pdf
  - 実習用データ zip
- 9 月 7 日 生命科学実習 4(田邊俊朗)
  - 講義スライド pdf
  - 配布資料 pdf

工学部・特別講義 V

- 課題提出状況(9/21 現在)
- 9 月 10 日 1 限 イントロダクション(岡崎威生)
  - 講義スライド pdf
- 9 月 10 日 2 限-4 限 情報科学基礎(名嘉村盛和)
  - アルゴリズムの基礎 pdf
  - コンピュータの基礎 pdf

- グラフ・アルゴリズム pdf
- 探索 pdf
- 整列 pdf
- ヒープ pdf

- 9 月 10 日 5 限-11 日 類似性発見(岡崎威生)
  - 講義スライド pdf
  - 演習問題 pdf
  - R 実行メモ pdf
  - R 概説(総合情報処理センター広報) pdf
  - BLOSUM50 表 pdf
  - PAM250 表 pdf
  - 正規分布数表 pdf
  - t 分布数表 pdf
  - F 分布数表 pdf
  - 遺伝子発現データ csv
- 9 月 12 日 配列と機能(有田正規)
  - 講義スライド pdf
- 9 月 13 日 タンパク質の機能と構造(広川貴次)
  - 講義スライド pdf
  - 課題 pdf
  - 関連リンク集

3　講義集

3.1　バイオインフォマティクスの歴史、生命科学の基礎

3.1.1 概要

講師：要 匡／琉球大学大学院医学研究科遺伝医学講座准教授

日時：2012 年 9 月 3 日　10:20～／12:50～／14:40 ～

場所：琉球大学農学部 207 講義室

内容：生化学、遺伝学、分子生物学など生命科学の発展に計算機が活用され、計算機の発展に伴ってバイオインフォマティクス分野が重要な役割を担うようになってきた歴史を学ぶとともに、細胞の構造や細胞内器官の働き、生体分子（DNA、RNA、タンパク質など）に関する生命科学の基礎的な知識と遺伝学の基本的な法則などを学ぶ。

3.1.2 準備

特筆事項なし

3.1.3 配布物

講義スライド

・バイオインフォマティクス概要基礎

・生命科学の基礎

・遺伝の基礎

3.1.4 課題

・レポート（記述 2 題）

3.1.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.1.6 時間割

時間	内容
10:40～ 10:55～	<u>講師自己紹介</u> 略歴紹介 研究内容紹介 遺伝子工学手法による疾患遺伝子の解析 TG マウス、KO マウスによる病態モデル研究 ヒト人工染色体ベクターによる遺伝子治療法 <u>バイオインフォマティクスの歴史</u> ・ Bioinformatics＝Bio(生物)+informatics(情報学) ・ 遺伝情報、核酸情報、タンパク質情報などを扱う ・ 従来の実験スタイルと新しいスタイル →ウェット＋ハイスループット解析＋情報処理 <u>ゲノム解析の革新</u> ・ シークエンス技術とヒトゲノムプロジェクト ・ 次世代シーケンサの誕生 ・ 次次世代シーケンサの誕生 ・ 1,000 人ゲノムプロジェクト ・ シークエンサーによるデータ量の変化 ・ ゲノム解読費用 ・ パーソナルゲノム時代を拓く ・ 生命科学研究とバイオインフォマティクス
11:50～	昼食
13:00～	<u>生命科学の基礎</u> ・ 生命とは（自己保存、系統保存、進化） ・ 細胞内器官 ・ ミトコンドリアの機能と局在タンパク ・ クローン羊　ドリー <u>ゲノム</u> ・ DNA＝設計図 ・ DNA の性質、相補性 ・ 分子遺伝学の基本ツールと用語解説 ・ DNA と遺伝子と染色体（とタンパク質）

	・ ゲノム構造と遺伝子 ・ コドン表 ・ 遺伝子スプライシング ・ 転写と翻訳 <u>遺伝子組換え技術</u> ・ ベクター ・ プラスミド ・ 制限酵素 ・ Ligase ・ DNA ライブラリー ・ PCR の原理 ・ シークエンス解析、使用機器 ・ 遺伝子組換え事例 遺伝子組換え医薬品、青いバラ、遺伝子改変マウス等 ・ キメラの作成 ・ 染色体について
14:20～	休憩
14:40～	<u>遺伝学の基礎</u> ・ 遺伝＝親から子へ受け継がれる形質 ・ メンデルの遺伝粒子説 ・ 病気と遺伝子 ・ 遺伝性疾患と対立遺伝子 ・ 常染色体遺伝、ミトコンドリア遺伝 ・ Hardy-Weinberg 平衡 ・ 血液型（ABO） 糖鎖について、血液型物質について、特殊な血液型とその遺伝 ・ OMIA、色々な形質の遺伝 肌の色、髪の毛の色、眼の色、耳あか型など ・ SNP とその種類 ・ 味覚の遺伝 PTC 味覚感受性遺伝、味覚受容体遺伝子 ・ その他の形質遺伝 アスパラガス遺伝、匂い受容体遺伝、トマト遺伝 ・ 多様な形質の遺伝を調べる方法 家系図、パーソナルシークエンス

バイオインフォマティクス

バイオインフォマティクスの歴史（概要）

2012年9月3日

琉球大学大学院医学研究科遺伝医学講座
要　　匡

研究内容

・遺伝子工学的手法を用いた疾患責任遺伝子の解析
（病気の原因を見つける、遺伝子で診断するなど）

・トランスジェニックマウス・ノックアウトマウス
を用いた病態モデル研究

・ヒト人工染色体ベクターによる遺伝子治療法の研究

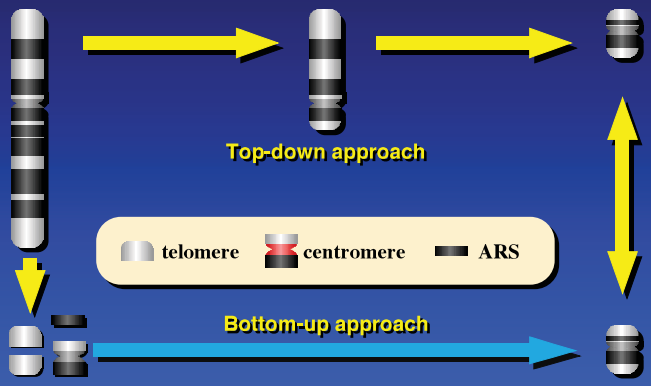
・巨大DNA，染色体操作の技術開発

発生工学，染色体工学，遺伝性疾患，ゲノム解析，
遺伝子治療，再生医療，胚幹細胞（ES細胞）など

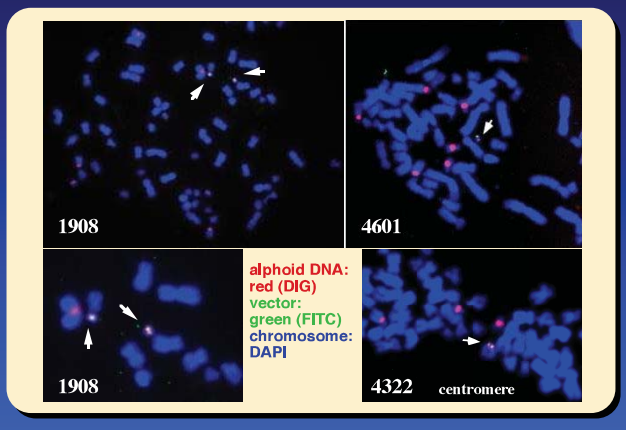
895-3331 (x2708)

医学部 第２基礎研究棟

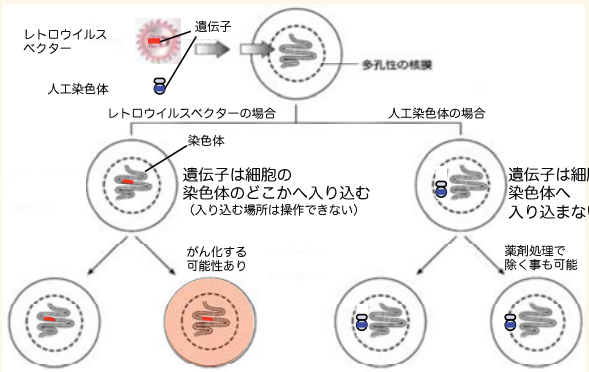
人工染色体をつくる



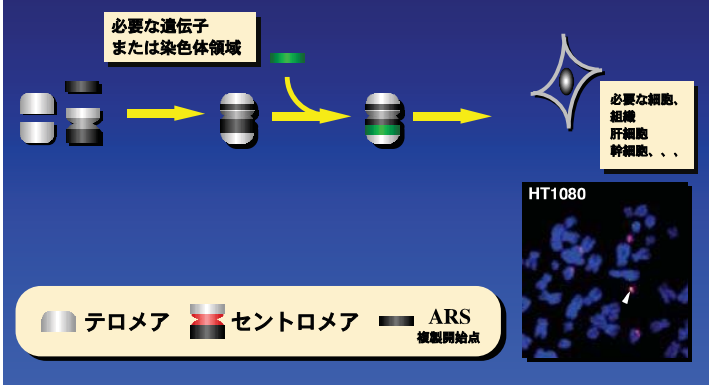
MAC formation in HT1080s



遺伝子を細胞へ導入する



人工染色体の応用



バイオインフォマティクスとは？

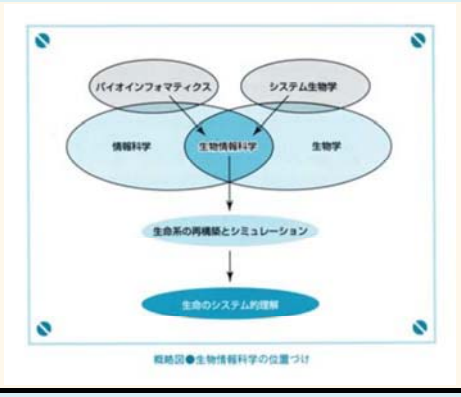
「Bio」＋「infomatics」

「バイオインフォマティクス」は「生命情報学」、「生物情報学」、「情報生物学」とも訳され、生物に関連する情報を包括的に扱う学問分野である。

ここでいう「情報」は狭い意味では遺伝情報（DNAの塩基配列やタンパク質のアミノ酸配列）から広く細胞、組織、器官、生物個体レベルでみられるさまざまな生命現象も含まれており、
バイオインフォマティクスは、地球上に存在する生命の現象を統合的に理解しようという学問ともいえる。

バイオインフォマティクスによって得られる知識は、生物学の基礎研究ばかりでなく、人間の医学や健康管理にも活用されている。

生物情報科学



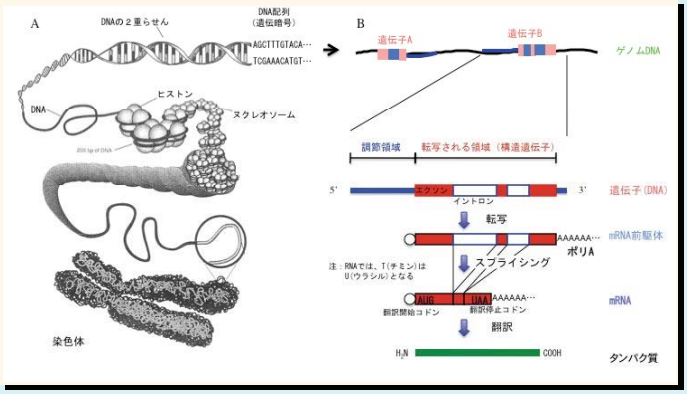
概略図●生物情報科学の位置づけ

バイオインフォマティクス

バイオインフォマティクスであつかう情報

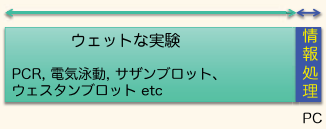
遺伝情報：核酸情報（DNA, RNAの塩基配列、
タンパク質のアミノ酸配列情報）

DNAと遺伝子と染色体とタンパク質

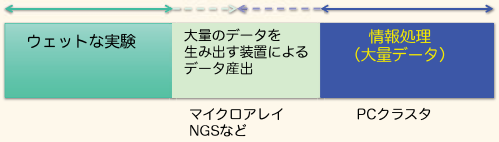


従来の実験スタイルと新しい実験スタイル

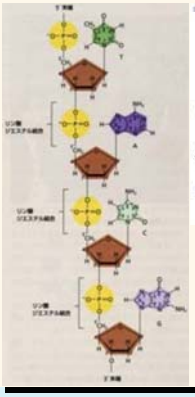
従来の実験スタイル

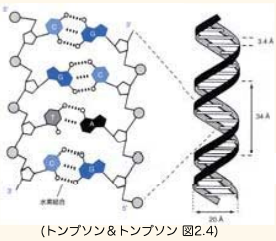


新しい実験スタイル



DNA





（トンプソン＆トンプソン 図2.4）

ワトソンとクリック

大きさ

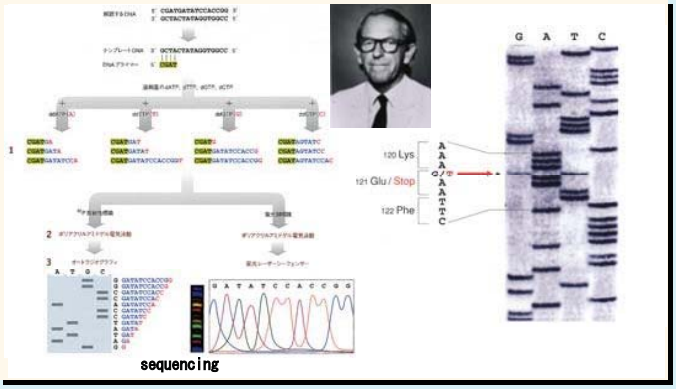
1 bp（base pair:塩基対）

1000 bp= 1kb, 1000kb=1Mb,
1000 Mb=1Gb, 1000Gb=1Tb

5′-AGCATGAAGTAG-3′
3′-TCGTACTTCATC-5′

塩基長:12 bp

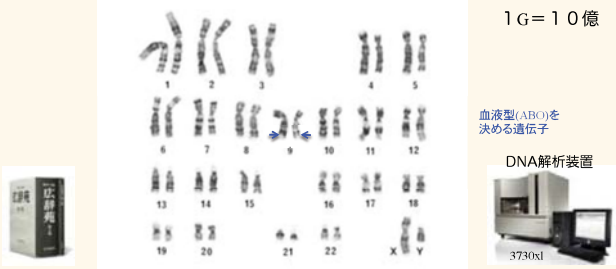
シークエンス解析



ヒトゲノム

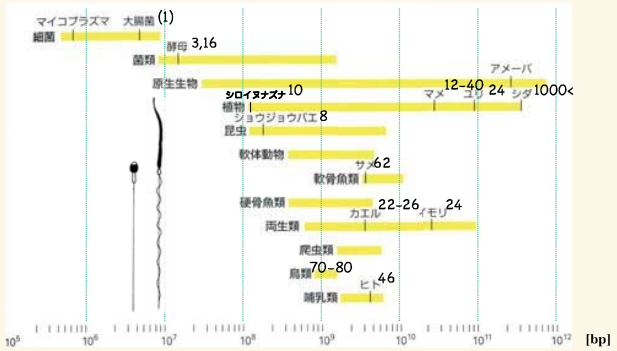
ヒトゲノム（2n）＝６０億塩基対 (6Gb: 6ギガベース)

１G＝１０億



広辞苑の文字をATGCで置き換えると 約４５０冊分

各生物のゲノムの長さ



大量のデータを生み出す機器と情報処理

マイクロアレイ → 発現プロファイル解析プログラム コピー数解析プログラム

質量分析装置 → マススペクトル解析プログラム

次世代シーケンサ → タグアセンブル タグマッピング プログラムなど

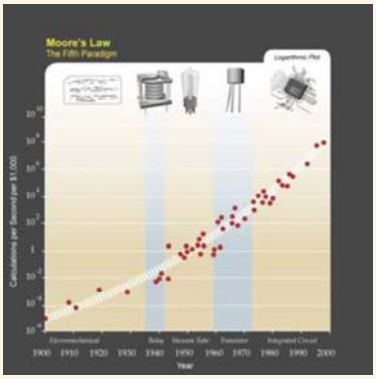
ムーアの法則

米Intel社の創業者ゴードン・ムーア（Gordon E. Moore）が1965年に提唱した、半導体技術の進歩に関する経験則。

『半導体チップの集積度は、およそ18カ月で2倍になる』というもの。

（＝指数関数的に進歩する）

マイクロプロセッサの性能は18カ月で2倍になる、などといわれることもある。



コンピュータシステム

PCクラスタ

1．超並列・分散計算

数十個~のCPU

2．大容量メモリ

数十~百GB, XXTB

3．大容量ストレージ

数十TB, XXPB

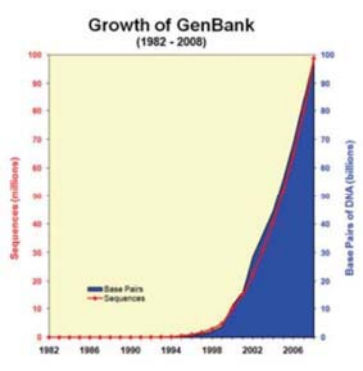
なぜ生命科学の解析にコンピュータが必要か？

1. 近年の生命科学データの爆発的な増大！
「過去３年に生成されたデータの量は，過去４万年の間に蓄積されたデータの総量よりも多い」

2. コンピュータを用いる２つのアプローチ：
・網羅的（ハイスループット）なウエット実験から生成される大量データの処理
・実験前にコンピュータによる情報処理を徹底的に行うことにより，ウエット実験を絞り込み効率化する
→インシリコスクリーニング

GenBankに登録されたDNA配列の推移

2012年9月3日



人と人の違い

個体差（表現型の違い）

辻先生

要

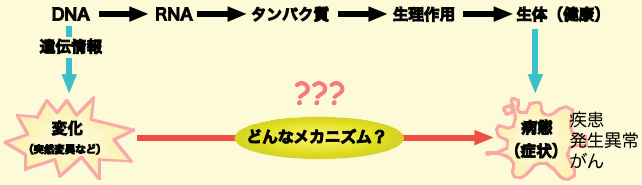
0.12%の相違

1%異なると
(1.2%)

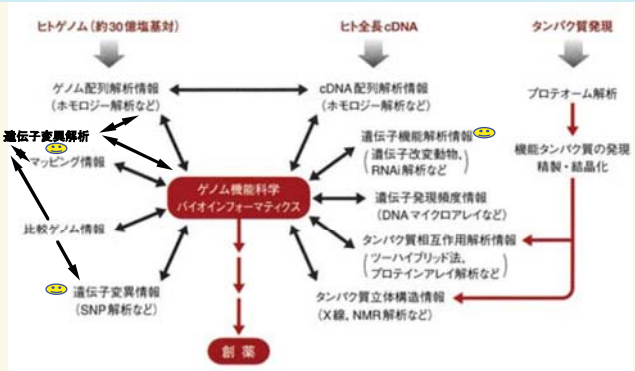
この違いに重要な情報が潜んでいる

遺伝情報と表現型を結びつける

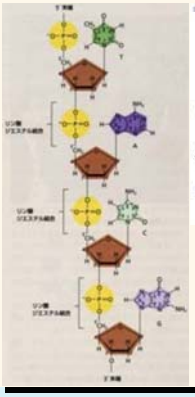
・差を観察することによって様々な現象の違いがわかる
・差を比較することによって様々な現象の重要な点がみつかる
・重要な分子（標的分子）を見いだす→病態解明・治療へ

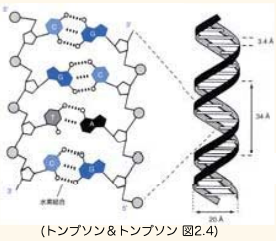


バイオインフォマティクス（ヒトゲノム）

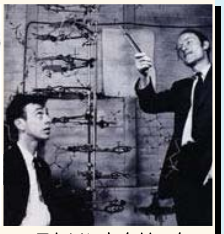


DNA





（トンプソン＆トンプソン 図2.4）



ワトソンとクリック

大きさ

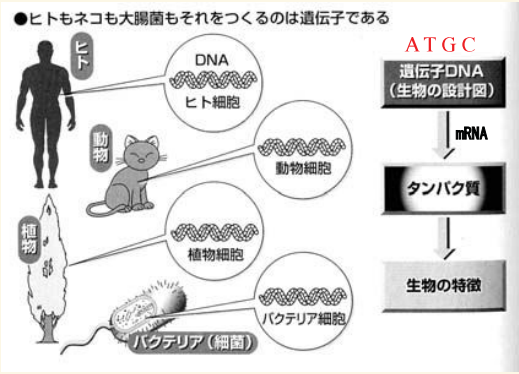
1 bp（base pair:塩基対）

1000 bp= 1kb, 1000kb=1Mb,
1000 Mb=1Gb, 1000Gb=1Tb

5′-AGCATGAAGTAG-3′
3′-TCGTACTTCATC-5′

塩基長:12 bp

DNA=設計図



従来のシーケンサーのデータ量

最高性能のシーケンサー

広辞苑約17ページ分

シーケンサー	メーカー	タイプ	塩基/read	read/run	塩基/run
3730xl	AB	キャピラリ	900	96	86,400

ヒトゲノムプロジェクト

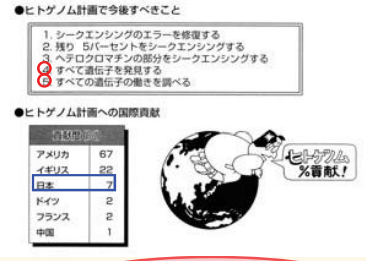
1990年　米国からスタート

フランシス　コリンズ　　クレイグ　ベンター
（セレラ社）

参加国

アメリカ合衆国
EU
（英国，フランス，
ドイツ）
日本
中国

2003年　終了宣言

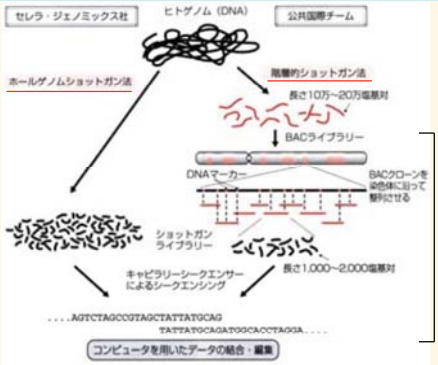


→ ２１世紀はゲノム機能科学の時代？

遺伝子の機能
疾患との関わり

遺伝子治療
創薬　etc

セレラ社とHGPのゲノム配列決定法



原則
非公開
（有料）

公開
（無料）

ヒトゲノムプロジェクト

1990年　米国からスタート

フランシス　コリンズ　　クレイグ　ベンター
（セレラ社）

参加国

アメリカ合衆国
EU
（英国，フランス，
ドイツ）
日本
中国

2003年　終了宣言

2003年

終了宣言！？？

ヒトゲノム解読終了？

2003年，2004年頃

「ゲノムは全部わかったんですよね。
じゃあ、私のを全部調べてください。」

「エッ！？」

その頃の答え

「ゲノムを読んだとしても配列だけで、
それにどんな意味があるのかはこれからなんですよ。」

「確かに全部調べようと思えばできますが、そもそも
ゲノムを読んだときどのくらいかかったかご存知ですか？」

ヒトゲノム解析

ヒトゲノムプロジェクト

参加国：米国、EU（英国、フランス、ドイツ）、日本、中国

1990 → 2001 → 2003

ヒトゲノムプロジェクトスタート → ヒトゲノムドラフト発表 → ヒトゲノムプロジェクト終了宣言

1998

クレイグ・ベンター
ゲノム解析開始
セレラ社
（3年以内に終了すると宣言）

13年　約3000億円

5年　約100億円

終了宣言！？

ABI 3730xl

キャピラリー
シークエンサー
2Mb/day（200万塩基／日）

原因未同定の疾患は意外と多い

OMIM(Online Mendelian Inheritance in Man)での統計

Number of Entries in OMIM (Updated 10 February 2012) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	13,077	640	48	35	13,800
+ Gene and phe	155	6	0	2	163
# Phenotype des	3,107	259	4	28	3,398
% Phenotype de	1,648	137	5	0	1,790
Other, mainly t	1,794	128	2	0	1,924
Totals	19,781	1,170	59	65	21,075

原因が特定されたもの　計3561

原因が特定されていないもの　計3714

合計7275

http://omim.org

ゲノム解析の革新

米国　もっと安くで読めるような技術を開発する人（グループ）にお金を出します

目標：１人1000ドル（１人　9万円）

1990 → 2001 → 2003 → 2004 → 2005

ヒトゲノムプロジェクトスタート → ヒトゲノムドラフト発表 → ヒトゲノムプロジェクト終了宣言 → 1000ドルゲノムプロジェクト(NIH, 米国) → 新型シークエンサー（Gシークエンサー）400Mb/day

1998

クレイグ・ベンター
ゲノム解析開始
セレラ社
（3年以内に終了すると宣言）

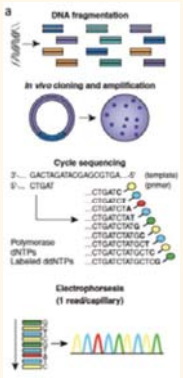
NHGRI

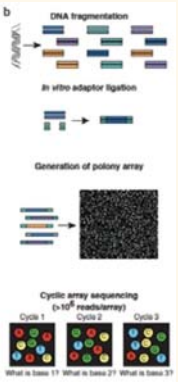
ABI 3730xl

キャピラリー
シークエンサー
2Mb/day

次世代シーケンサ

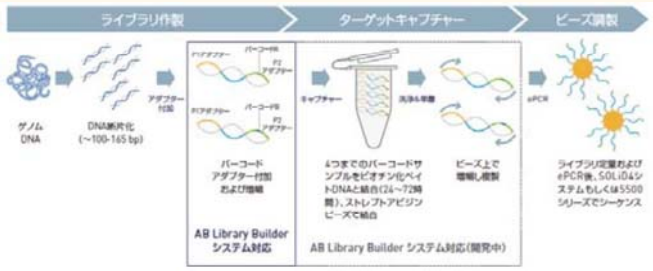
サンガーシーケンス





次世代シーケンス

次世代シーケンサ ワークフロー



フラグメントライブラリ　→フラグメント解析、ペアエンド解析

メイトペアライブラリ　→メイトペア解析

ジェームス・ワトソンのゲノム

2007年5

使用機種

所要日数

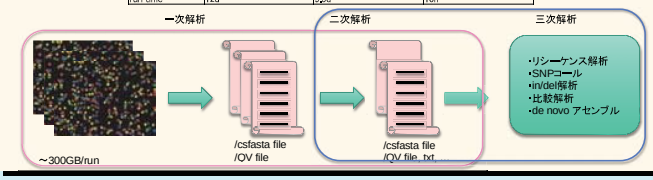
費用　1億2千万円

やっぱり遺伝子に変異がある!!

次世代シーケンサ

2008~2009

	SOLiD3	GA IIx	FLX
Read length	50bp x2 (mate-paired)	100bp x2 (paired-end)	400bp
# of reads	1G	300M	400K
output/run	60Gb	30Gb	160Mb
run time	12d	9.5d	10h



ゲノム解析の革新

ヒトゲノムプロジェクトの変遷

1990 ヒトゲノムプロジェクトスタート → 2001 ヒトゲノムドラフト発表 → 2003 ヒトゲノムプロジェクト終了宣言 → 2004 1000ドルゲノムプロジェクト(NIH, 米国) → 2008 1000人ゲノムプロジェクト（ヒトゲノム再解析）(国際コンソーシアム)

1998
クレイグ・ベンター
ゲノム解析開始
セレラ社
（３年以内に終了すると宣言）

キャピラリーシークエンサー
2Mb/day

ポストゲノム
cDNA
SNP
プロテオーム

癌ゲノムプロジェクト
(NIH, 米国)

G-シークエンサー
400Mb/day

知的財産

2009
アジア100人ゲノムプロジェクト(韓国)

1000人ゲノムプロジェクト

1000 Personal Genome Project　(2008~)

提唱　フランシス・コリンズ
Francis Collins, MIT, NIH, (ex-Director NHGRI)

Knome社
(ボストン)
35万ドル／1人

ギガシークエンサー
SOLiD (アプライド バイオシステムズ社)
Solexa（イルミナ社）
FLX　（ロシュ社）

組織
サンガー研究所（英国）　２６台～
ブロード研究所（米国）　２０台～
北京ゲノム研究所@深圳（中国）　６台～

2009：SOLiD 2, GA IIx 29, FLX 3

炎黄計画
(YanHuang Project)
2007.10~
２人以上ゲノム解読

ゲノム読みますよ
3000万円／１人
但し中国人優先

1000人ゲノムプロジェクト

1000 Personal Genome Project (2008年スタート)

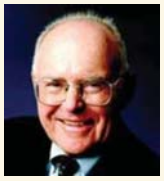
民間助成（懸賞金）

2006年末～　Archron X Prize for Genomics

100人のゲノムを10日以内にシーケンスしたチーム
（１人当たり１万ドル以下）

→賞金1000万ドル

ムーアの法則

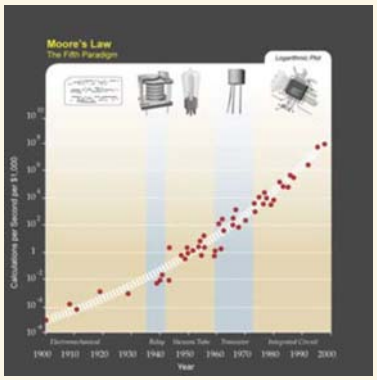


米Intel社の創業者ゴードン・ムーア（Gordon E. Moore）が1965年に提唱した、半導体技術の進歩に関する経験則。

『半導体チップの集積度は、およそ18カ月で2倍になる』というもの。

（＝指数関数的に進歩する）

マイクロプロセッサの性能は18カ月で2倍になる、などといわれることもある。



次世代シーケンサ

次世代シーケンサ
5500xl SOLiD

ライフテクノロジーズ社

Hi-Seq2000

イルミナ社

FLX +

ロシュ社

	SOLiD 3Plus システム	SOLiD 4Q システム	5500xl SOLiD システム
コスト/ラン	約2,700,000円	約600,000円	約1,400,000円
ギガ(10億)ベース/ラン	80	100	180
コスト/GB	約40,000円	約6,000円	約7,800円

(2010年11月現在)

処理能力（解析読取塩基数）	約700M（メガ）塩基　（1ランあたり）
リード長	最長1,000塩基
リード数	約1,000,000リード
ラン時間	23時間
サンプル	ゲノムDNA、PCR産物、cDNAなど
本体サイズ	74.3cm（W）×69.8cm（D）×36.1cm（H）
本体重量	242kg

リード長	ラン時間（日数）	データ量（Gb）
1 x 35bp	~2	95-105
2 x 50bp	~5.5	270-300
2 x 100bp	~11	540-600

次世代シーケンサ

パーソナルタイプ

Ion PGM

Ion proton

MiSeq

イルミナ社

GS Junior

ロシュ社

Read Length	Total Time for Amplification and Sequencing*	Output**
1 x 35 bp	~4 hours	175-245 Mb
2 x 25 bp	~5 hours	250-350 Mb
2 x 100 bp	~10 hours	[illegible]
2 x 150 bp	~27 hours	7Gb
Reads	5.0-7.0 million clusters pass and 10.0-14.0 million paired-end reads	

処理能力（解析読取塩基数）	平均35 M（メガ）塩基　（1ランあたり）
リード長	平均 400塩基
リード数	平均　100,000リード
ラン時間	10時間（シークエンシング） 2時間（データ処理）
サンプル	ゲノムDNA、PCR産物、cDNAなど
本体サイズ	34（W）×53（D）×41（H）cm
本体重量	約25kg

次世代・次次世代シーケンサ

超高速シークエンス　2~4Gb/run →600Gb/run

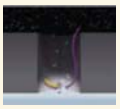
次世代シーケンサ
(5500xl SOLiD)

Hi-Seq2000

Roche社
(454)
FLX

超超高速シークエンス　20Gb/run~2.4Tb/day?

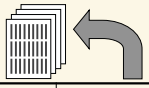
Pacific Bio社
2011年発売



SMRT sequencing

次世代シーケンサのデータ量

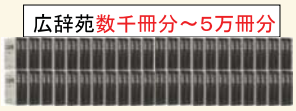
従来のシーケンサー



広辞苑約17ページ分

シーケンサー	メーカー	タイプ	塩基/read	read/run	塩基/run
3730xl	AB	キャピラリ	900	96	86,400
SOLiD 5500	ABI	次世代	100	1,800,000,000	180,000,000,000

5500xl SOLiD



広辞苑数千冊分～5万冊分

数百億文字～

次世代シーケンサー

ゲノム解読費用の価格破壊？！

		個人ゲノム解読の費用	
1990年～2003年	ヒトゲノムプロジェクト	13年	約3000億円
2003年	クレイグベンター独自に読む	５年	約300億円
2007年	次世代シーケンサー(454)使用	67日	約1.2億円
2008年	高性能次世代シーケンサー	約1か月	約3000万円
2009年	超高性能次世代シーケンサー	約1か月	約1500万円～800万円
2010年	新超高性能次世代シーケンサー	数週間	100万円～
20XX年		数時間？	10万円？！

全ゲノム解析プロジェクト

種々の生物の全ゲノム解析が
進行中である

BMC Genomics

RESEARCH ARTICLE　Open Access

Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data

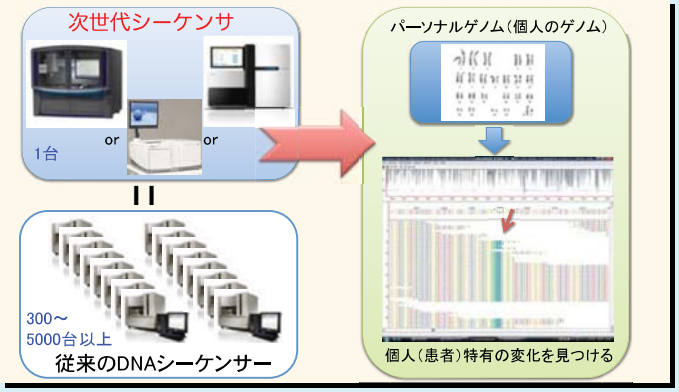
納豆菌ゲノム解読

納豆を作る納豆菌＝写真、慶応大提供＝のゲノム（全遺伝情報）を、慶応大の榊原康文教授らの研究チームが初めて解読した。納豆のネバネバの秘密は、塩基配列のわずかな違いにあった。

納豆菌は枯草菌の一種でネバネバ成分を作るが、ネバネバを作らない枯草菌もいる。研究チームは、納豆菌のネバネバ作りに関する主な遺伝子二つと、ネバネバを作らない枯草菌の同じ場所にあたる遺伝子を比べた。すると遺伝子がそれぞれ１塩基しか違っていなかった。

納豆菌のゲノムサイズは約409万塩基対。遺伝子数は推定約４千個だった。

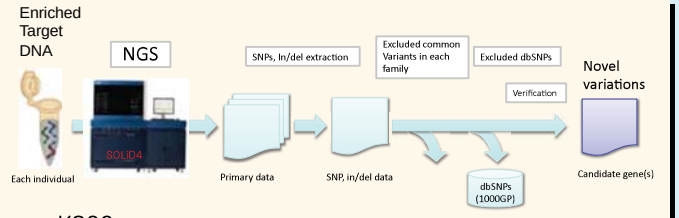
パーソナルゲノム時代を拓く次世代シーケンサ



SNPs in KS06

ID#	homoSNP	heteroSNP	totalSNP
KMS001	66,471	62,234	128,705
KMS002	66,682	57,270	123,952
KMS003	67,589	58,393	125,982

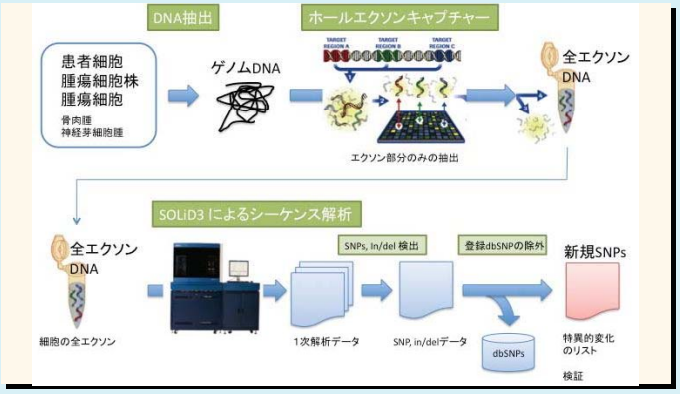
NGSデータ解析



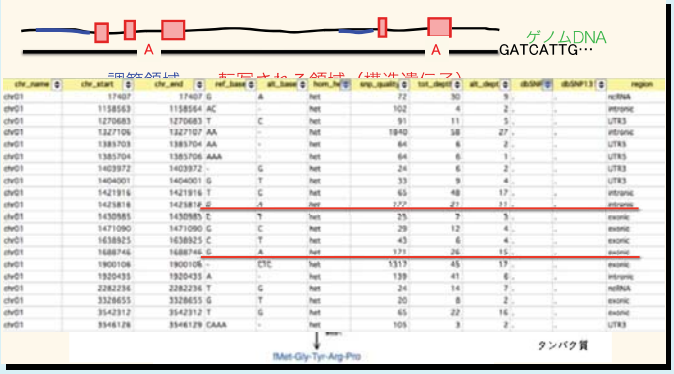
KS06

KS06	KMS001	KMS001-002	KMS001-002 -003	KMS001-002 -003-dbSNPs
heteroSNPs	62,234	33,792	19,762	625

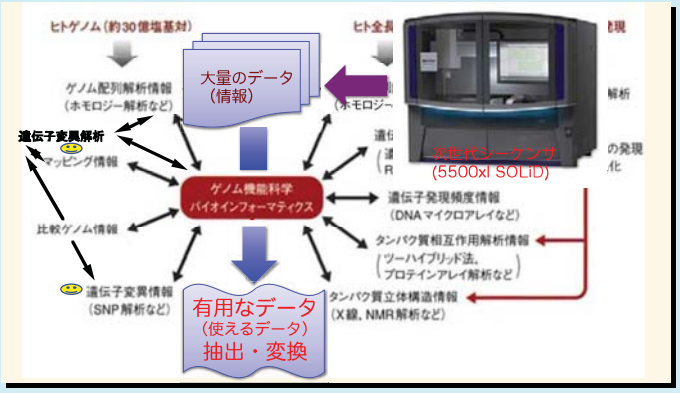
Exome解析



ゲノムの構造と遺伝子3



生物研究におけるバイオインフォマティクス



生命科学研究とバイオインフォマティクス

今や
生命科学研究に
↕
バイオインフォマティクスは

必要不可欠と
なりつつある

生命科学研究とバイオインフォマティクス

生命科学研究
＋
バイオインフォマティクス

＝最先端

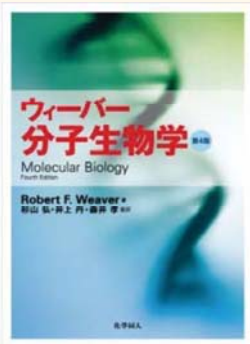
バイオインフォマティクス集中講義

生命科学の基礎
（分子生物学）

2012年9月3日

琉球大学大学院医学研究科遺伝医学講座
要　　匡

参考本



¥23,415　　¥12,600

生命

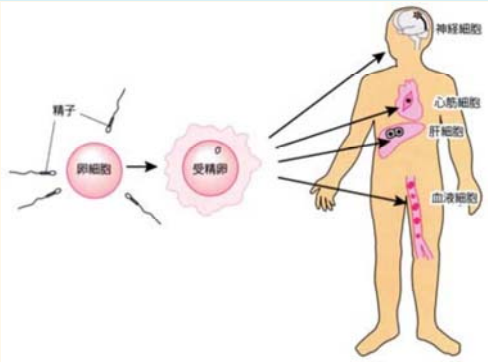
・自己保存の法則
生命は自らその生命を維持することができる

・系統保存の法則
生命は自らの系統を維持することができる

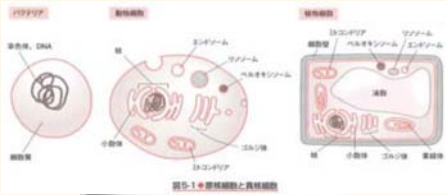
・進化の法則
生物は時代とともに進化し、その系統を分化させていく

全ての生物は細胞から成り立っている
自己保存、自己増殖

受精卵から個体へ



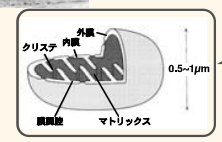
細胞内器官



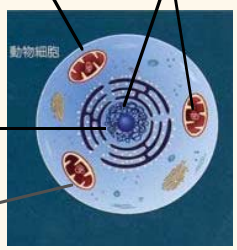
区画	主な機能
細胞質	物質輸送と多数の代謝経路をもつ
核	ゲノムがあり，DNAとRNAを合成
小胞体	脂質や膜タンパク質の合成
ゴルジ体	タンパク質と脂質の修飾と選別輸送
エンドソーム	細胞内への物質取り込みと選別
リソソーム	細胞内の物質消化
ペルオキシソーム	有毒分子の酸化
ミトコンドリア	酸化的リン酸化によるATP合成
葉緑体	光合成によるATP合成と炭素固定

ミトコンドリア

光学顕微鏡像

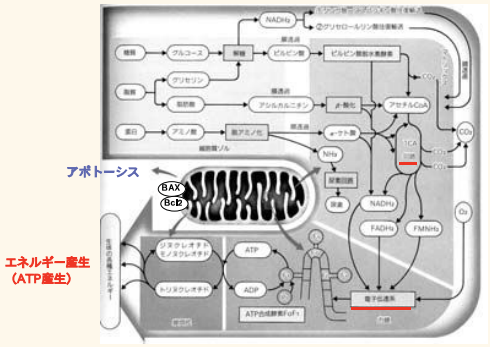


ミトコンドリアの模式図

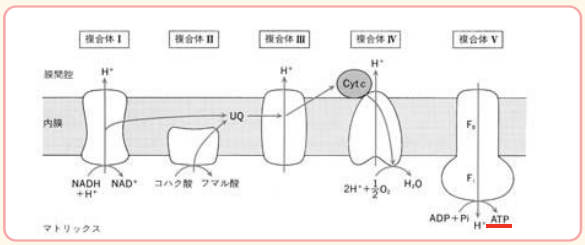


細胞内小器官の模式図

ミトコンドリア機能



ミトコンドリア局在タンパク



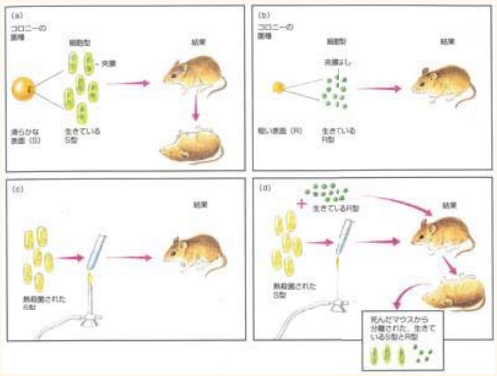
電子伝達系
エネルギー産生
(ATP産生)
構成タンパク 80<

由来

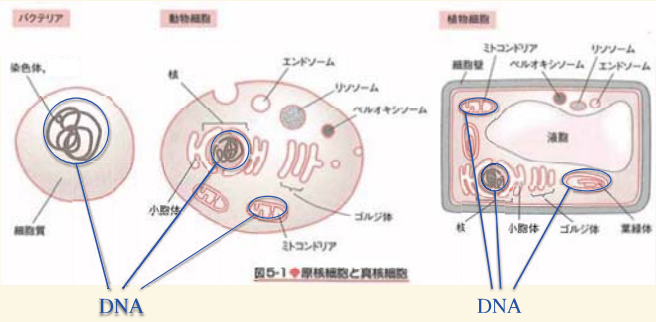
呼吸活性の高い組織（脳，骨格筋など）
にミトコンドリアは多い

遺伝子局在	ポリペプチド数					合計
	電子伝達系				ATP 合成酵素	
	複合体 I	II	III	IV	V	
ミトコンドリア	7	0	1	3	2	13
核	≥33	4	10	10	10	≥67
合　計	≥40	4	11	13	12	≥80

グリフィスの実験

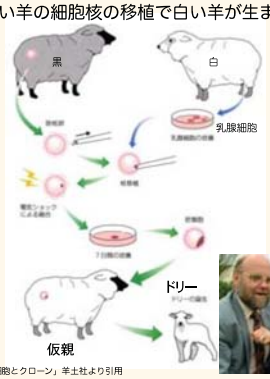


細胞内器官



クローン羊　ドリー

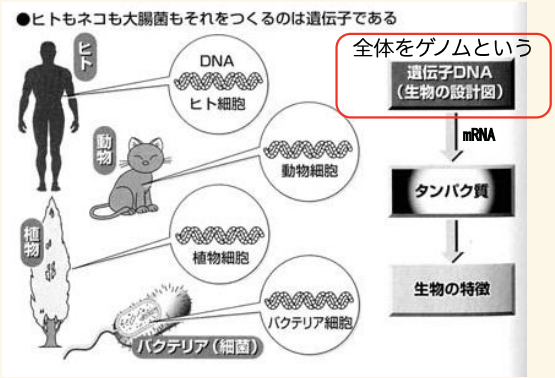
白い羊の細胞核の移植で白い羊が生まれる



「幹細胞とクローン」羊土社より引用

1997年

DNA=設計図



DNA(デオキシリボ核酸)

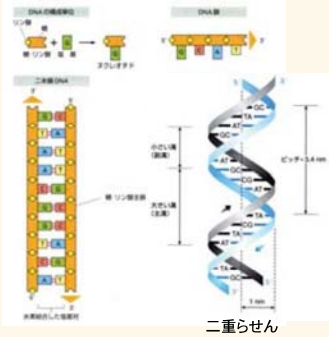
DNAの性質

- 4種類の塩基
  (A, G, C, T)
- 二重らせん構造
- 相補的水素結合
- 陰性に帯電

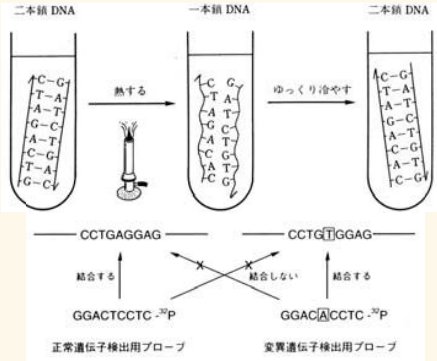
大きさ

1 bp(base pair:塩基対)

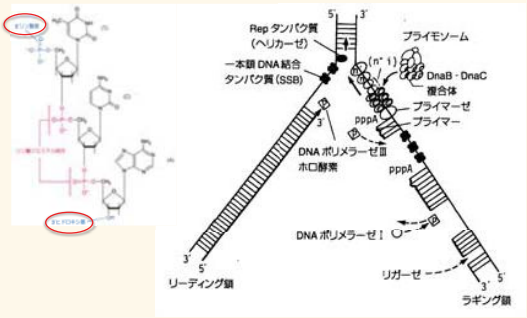
1000 bp= 1kb, 1000kb=1Mb
1000 Mb=1Gb, 1000Gb=1Tb



DNAの相補性



DNA複製に関与するタンパク複合体



岡崎令治

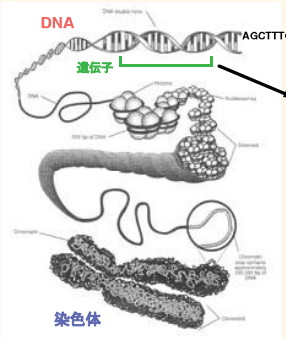
分子遺伝学の基本的ツールと用語

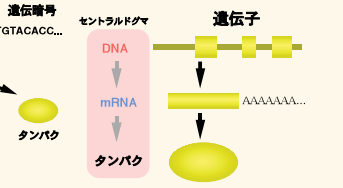
DNAハイブリダイゼーション
ゲル電気泳動
プローブ
Southern blotting
northern blotting
western blotting
ポリメラーゼ連鎖反応(PCR)
ゲノムPCRとRT-PCR
定量PCR
プライマー
シークエンス解析
FISH
aCGH

制限酵素
ライゲーション
ベクター
プラスミド
ライブラリー
宿主
挿入断片
ゲノムデータベース
解析ソフトウェア

トンプソン＆トンプソン遺伝医学
(第3章、第4章)
ヒトの分子遺伝学
(Chapter 1, 5, 6, 7-1)

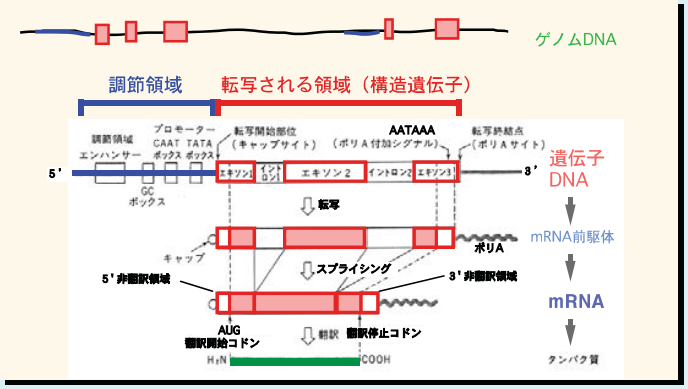
DNAと遺伝子と染色体とタンパク質





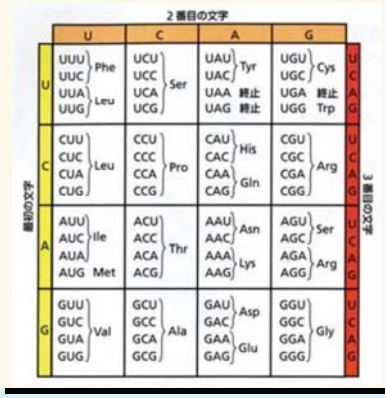
DNA は体の設計図であり、
それは 染色体 に収納されている。
DNA にはタンパクを作るための情報が書いてあるところがあり、
それを 遺伝子 という。
染色体は、細胞の中の 核にある。

ゲノムの構造と遺伝子3



コラーゲン遺伝子

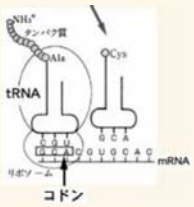
コドン表



最初の文字	2番目の文字 U	2番目の文字 C	2番目の文字 A	2番目の文字 G	3番目の文字
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA 終止	UGA 終止	A
	UUG Leu	UCG Ser	UAG 終止	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

・３塩基で１つのアミノ酸をコードしている

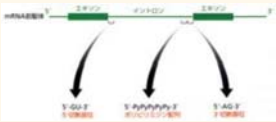
・３塩基単位の読み枠をフレームという



遺伝子スプライシング

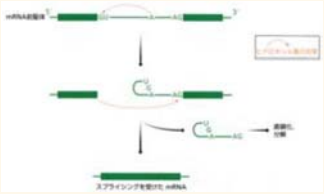
コンセンサス配列

5'-AG↓GUAAGU-3'

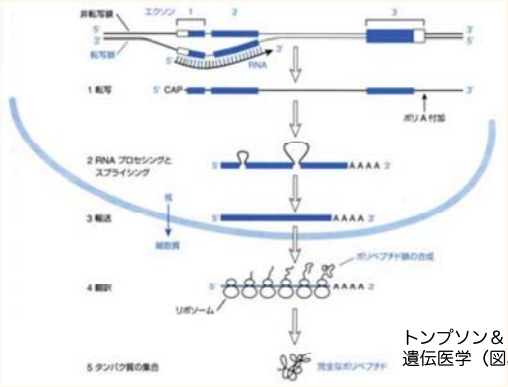


コンセンサス配列

5'-(Py)6NCAG ↓ -3'

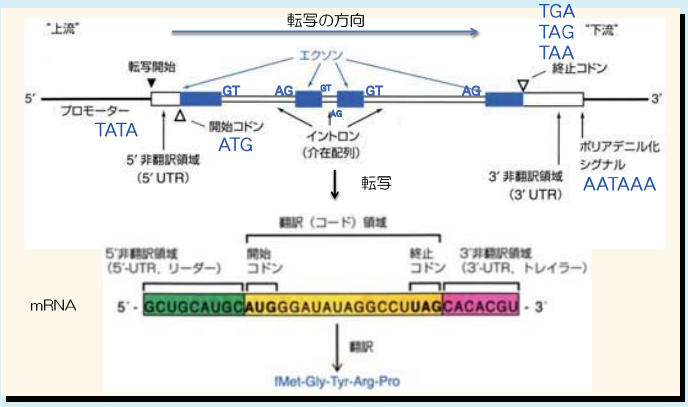


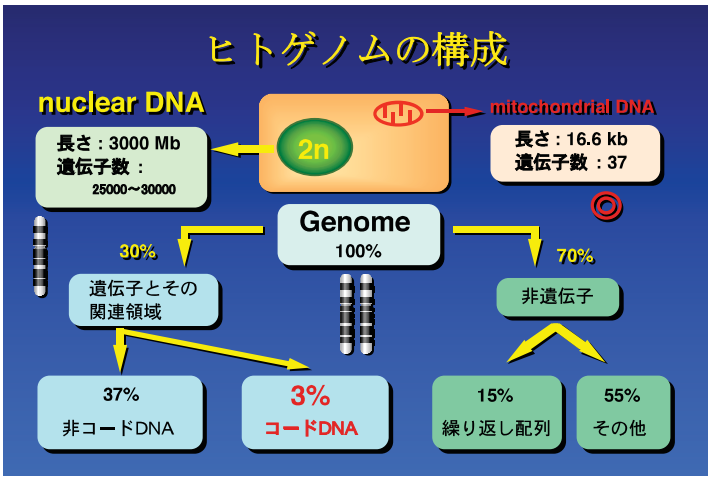
DNAからRNA, タンパク質への情報の流れと局在



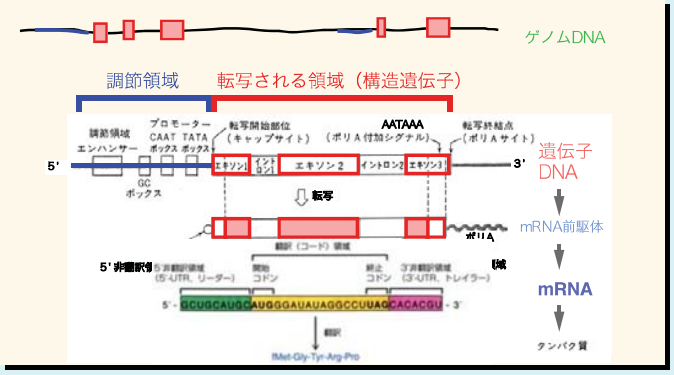
トンプソン＆トンプソン遺伝医学（図3.5）

遺伝子の構造

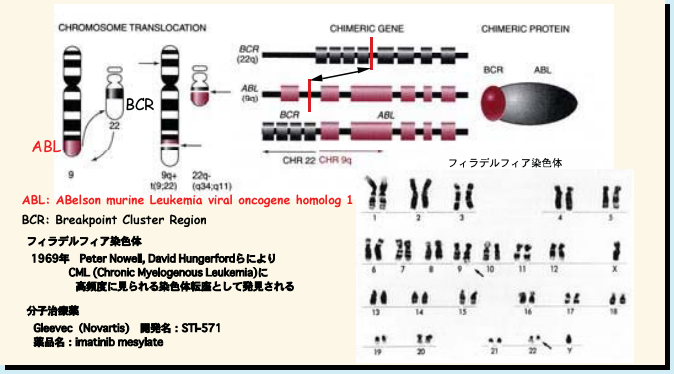




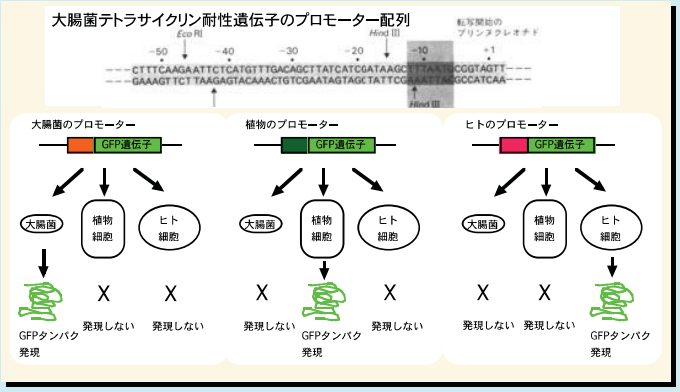
ゲノムの構造と遺伝子3



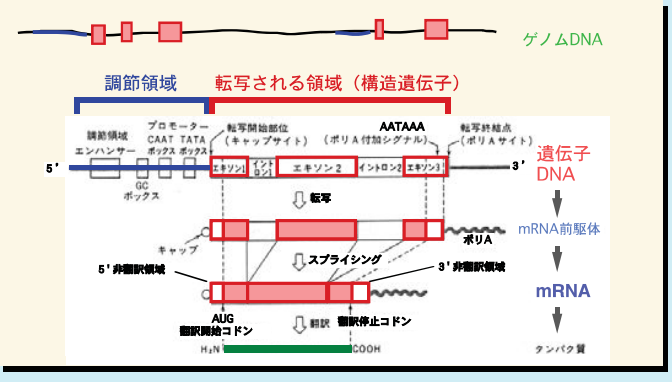
白血病における遺伝子融合



プロモーター



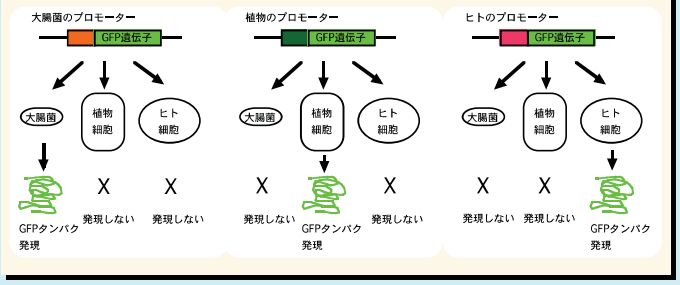
ゲノムの構造と遺伝子3



遺伝子操作技術

DNAを　切って　つないで　増やして　みて　改変する。

組換えDNA技術



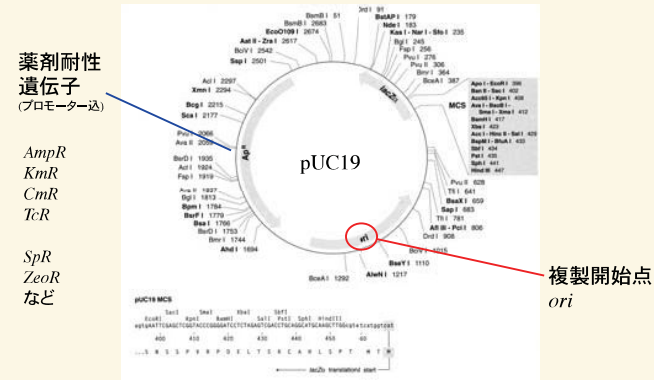
組換えDNA技術

ベクター

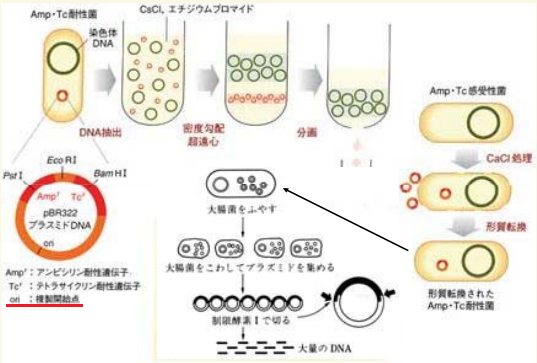
酵素（のりとはさみ）
　制限酵素（restriction enzyme）
　連結酵素（DNA ligase）

形質転換（transformation）

プラスミド



ベクター



保持可能な
DNAの長さ

プラスミド
　～15kb
ファージ
　9~20kb
コスミド
　40kb
PAC
　～200kb
BAC
　～300kb
YAC
　100kb
　～3Mb

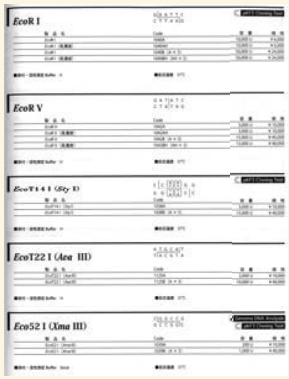
制限酵素

酵素名	切断部位の塩基配列	発見された微生物
Eco R I	5′ G|AATTC 3′ 3′ CTTAA|G 5′	*Escherichia coli* RY13
Bam H I	5′ G|GATCC 3′ 3′ CCTAG|G 5′	*Bacillus amyloliquefaciens* H
Hind III	5′ A|AGCTT 3′ 3′ TTCGA|A 5′	*Haemophilis influenzae* R_d
Pst I	5′ CTGCA|G 3′ 3′ G|ACGTC 5′	*Providenica stuartii* 164
Sac I	5′ GAGCT|C 3′ 3′ C|TCGAG 5′	*Streptomyces achromogens*
Hae III	5′ GG|CC 3′ 3′ CC|GG 5′	*Haemophilus aegyptius*
Pvu II	5′ CAG|CTG 3′ 3′ GTC|GAC 5′	*Proteus vulgaris*
Sma I	5′ CCC|GGG 3′ 3′ GGG|CCC 5′	*Serratia marcescens* S_b

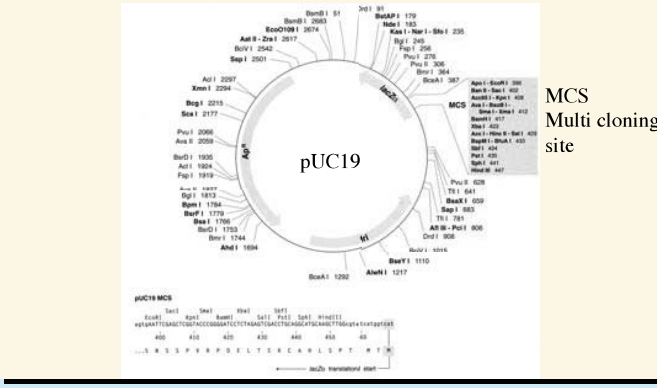
いろいろな制限酵素

4塩基認識酵素 1/4x4x4x4= 1/250塩基対
6塩基認識酵素 1/4x4x4x4x4x4= 1/4000塩基対

I-Sce I 認識部位 0箇所/ヒトゲノムDNA

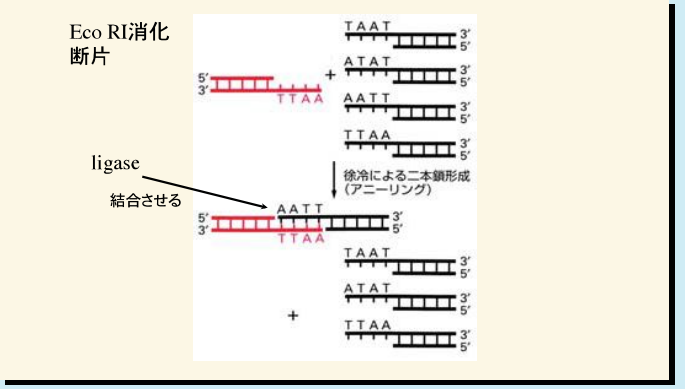


制限酵素地図

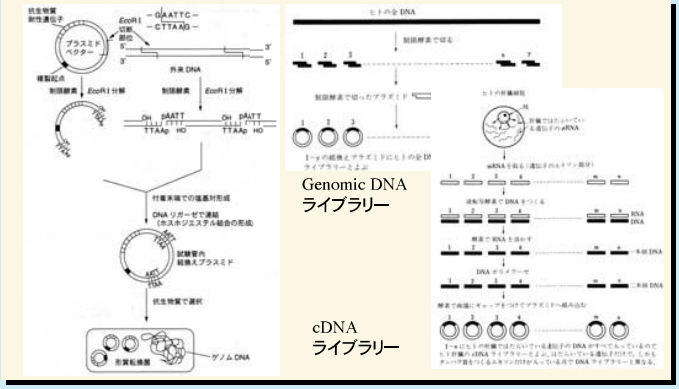


MCS
Multi cloning
site

Ligase

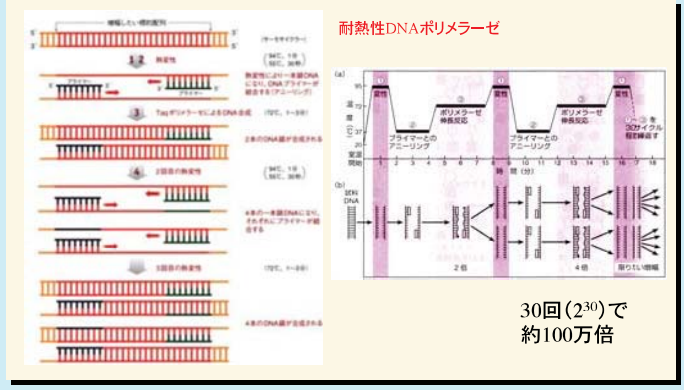


DNA library

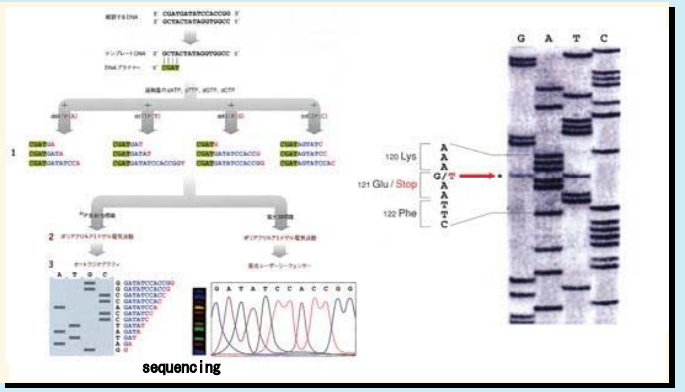


組換えDNA操作

PCRの原理

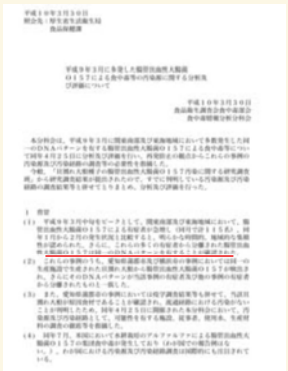


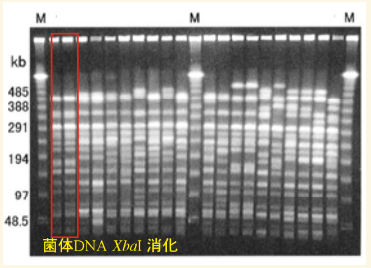
シークエンス解析

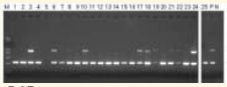


使用機器

腸管出血性大腸菌（0157）







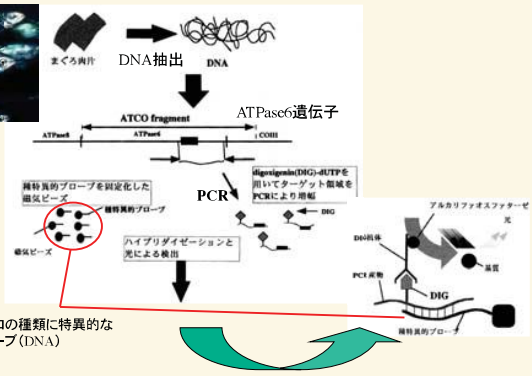
PCR

マグロ類の判別（DNA鑑定）

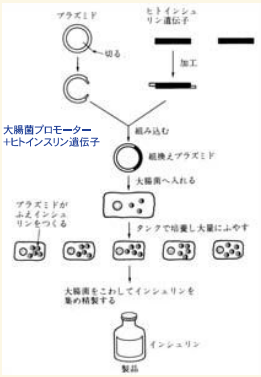
マグロ

クロマグロ

トンボマグロ



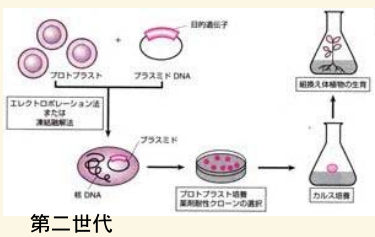
遺伝子組換え医薬品



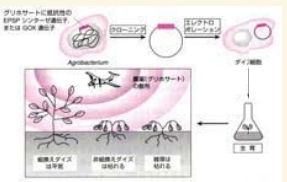
バイオ医薬品	研究開発*1	外国発売（年）	日本発売（年）	製造技術	おもな対象疾患	アミノ酸残基数
ウロキナーゼ			84	組織培養	血栓溶解剤	
インターフェロ					がん，肝炎*2	166
ヒト・インシュ					糖尿病	51
ヒト・成長ホル					小人症	192
インターフェロ（CC）*3					がん，肝炎*2	
インターフェロ（GE）*3					がん，肝炎*2	165
B型肝炎ワクチ					B型肝炎予防	
ヒト・成長ホル					小人症	191
エリスロポエチ					腎性貧血	166
インターフェロ					腎臓がん	143
t-PA（CC）					急性心筋梗塞	
t-PA（GE）					急性心筋梗塞	527
G-CSF	1986	91	91	組換え DNA	好中球減少症	174
マウス・モノクローナル抗体		86	91	組織培養	拒絶反応抑止	
インターロイキン 2	1982	89	92	組換え DNA	血管肉腫	133
プロウロキナーゼ			92	組織培養	急性心筋梗塞	411
血液凝固第Ⅷ因子		92	93	組換え DNA	血友病 A	
改変型 G-CSF			94	組換え DNA	好中球減少症	174
ヒト・ソマトメジン C（IGF-1）			95	組換え DNA	糖尿病など	70
ナトリウム利尿ペプチド（hANP）			95	組換え DNA	急性心不全	28

*1 cDNA のクローン化または組換え産物の生成，*2 B型肝炎とC型肝炎，*3 非天然型，*4 組織培養製品，*5 組換え製品，*6 天然型

遺伝子組換え植物



第二世代



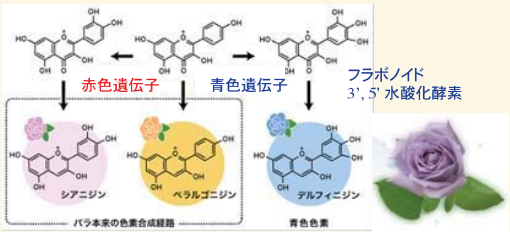
植物による医薬品の産生

植 物	医 薬 品
タバコ	モノクローナル抗体（アメリカ）
	B型肝炎表面抗原（アメリカ）
	トリコサンチン（アメリカ）
ナタネ	エンケファリン（ベルギー）
トマト	ACE2 ペプチド（日本）
ジャガイモ	ヒト血清アルブミン（アメリカ）

青いバラ

青い（?紫）カーネーション

ムーンダスト（発売中）



遺伝子導入・改変マウス

トランスジェニックマウス

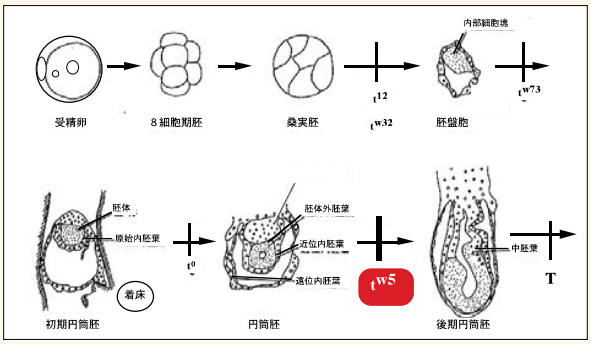
遺伝子導入マウス

ノックアウトマウス

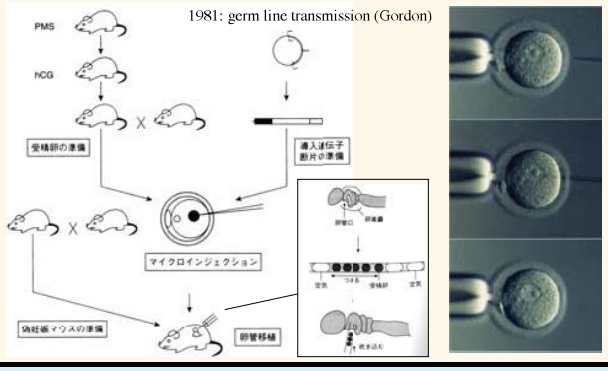
相同遺伝子組換え（ジーンターゲティング）マウス

トラップマウス

初期発生



Transgenic mouse 作製



遺伝子導入（注入）装置

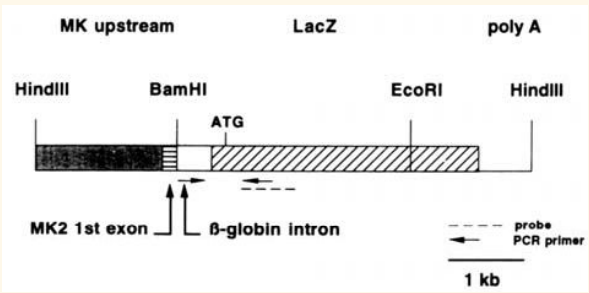
Injection 装置

倒立顕微鏡，ノマルスキー微分干渉装置

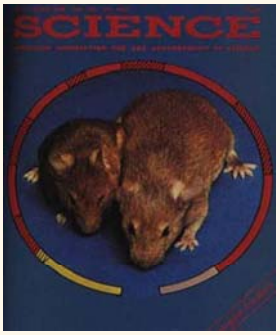
マイクロマニピュレーター

三次元

導入遺伝子の基本構造



スーパーマウス

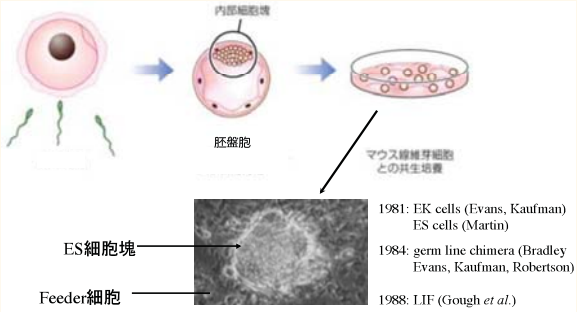


（左）
対照マウス

（右）
ヒト成長ホルモン
遺伝子導入マウス

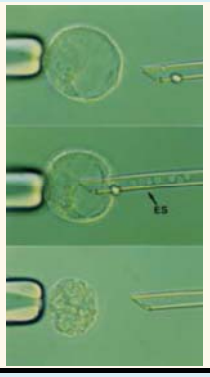
Palmier, Brinster, Evans *et al.*
Science 222:809-(1983)

胚性幹細胞(ES細胞)

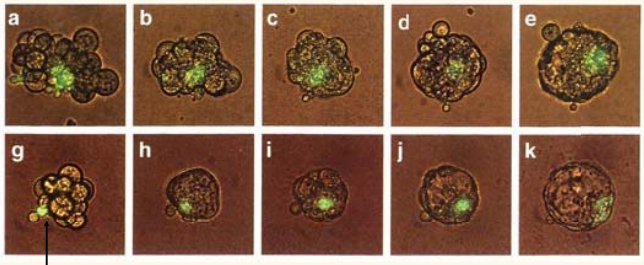


ES細胞（マウス）

Blastocyst injection



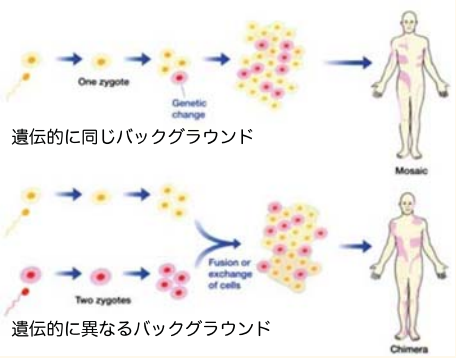
GFP標識ES細胞



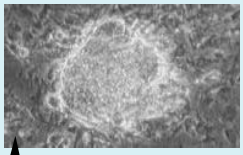
ES細胞 モルラ胚 → 胚盤胞

GFP発現キメラマウス

モザイクとキメラ



Embryonic stem (ES) cells and chimeras

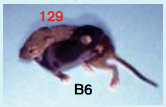


ES colony

(D3, CCE, E14-1, TT2 ...)

Feeder cells :
primary embryo fibroblasts
STO (SL10)

chimeras



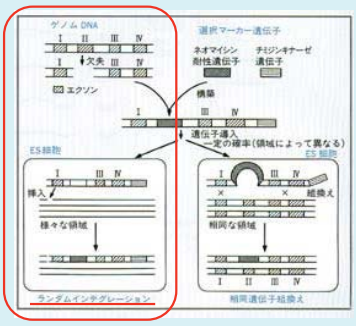
Homologous recombination

ゲノムDNA単離

系統
10kb<
遺伝子構造
制限酵素地図

ES細胞での発現

ノーザンブロット
RT-PCR



マーカー遺伝子の決定

neo^r
HPRT
puro^r
brs^r
tk
DTA

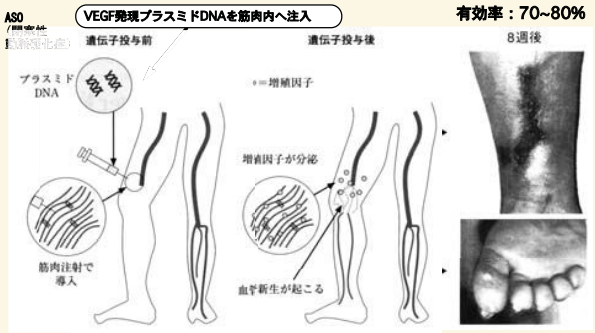
相同組換えクローンの確認

サザン法
PCR法

Naked DNAによる遺伝子治療

VEGF発現プラスミドDNAを筋肉内へ注入

有効率：70~80%



プラスミド注入法によるASO性難治性虚血性潰瘍の治療　（タフツ大学）

バイオインフォマティクス

遺伝学の基礎

2012年9月3日(3)

参考本

¥10,500

¥9,975

参考本

￥1,575

遺伝

親から子へ受け継がれる形質

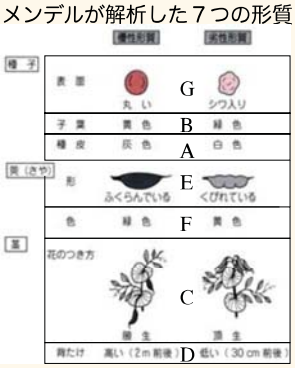
ヒポクラテス
(Ἱπποκράτης)
紀元前460年頃 - 紀元前370年頃

父親の身体の各部分が、遺伝を支配するある物をつくりだし、それが精液の中にまとめられて生まれてくる子供の形質を決定する

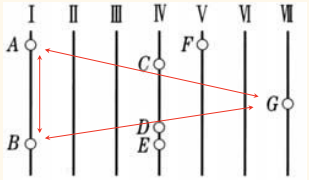
アリストテレスは黒人と白人の混血児を調査して隔世遺伝を指摘し、 また体の一部を失った父親の子供が同じように人体を欠損して生まれてくるわけでないことなどから、 ヒポクラテスの説を批判

メンデルの遺伝粒子説

メンデルが解析した７つの形質



エンドウの染色体と遺伝子の位置



A-B, A-G, B-Gの組合せに注目した交雑実験から「遺伝粒子の概念」を導き出した。（1865年）

メンデルの法則

Gregor J. Mendel
(1822-1884)

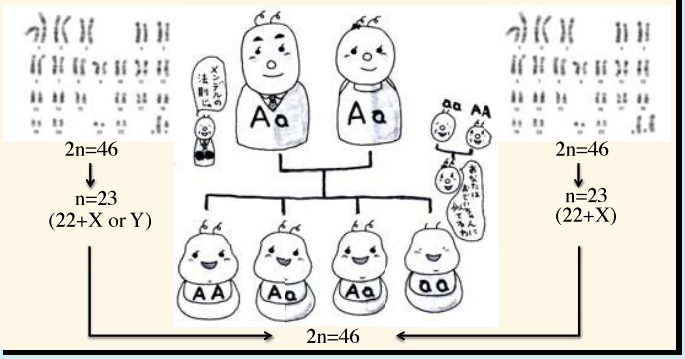
法則	内容
分離の法則	アリルの伝達は等確率
優劣の法則	表現型は遺伝子型の関数
独立の法則	アリルごとの伝達は独立

AA x aa
↓
Aa x Aa
↓
AA Aa Aa aa
1 : 1 : 1 : 1
(A) 3 : (a) 1

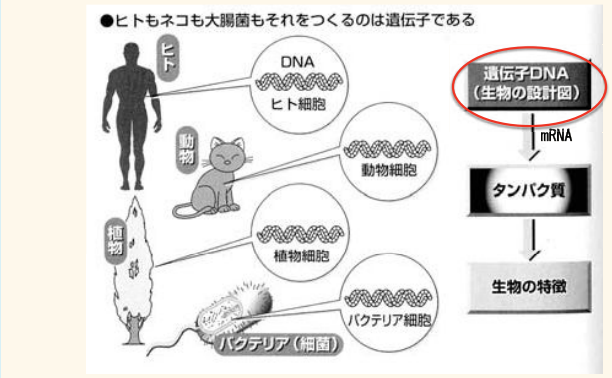
AaBb x AaBb
↓
AB Ab aB ab
9 : 3 : 3 : 1

遺伝

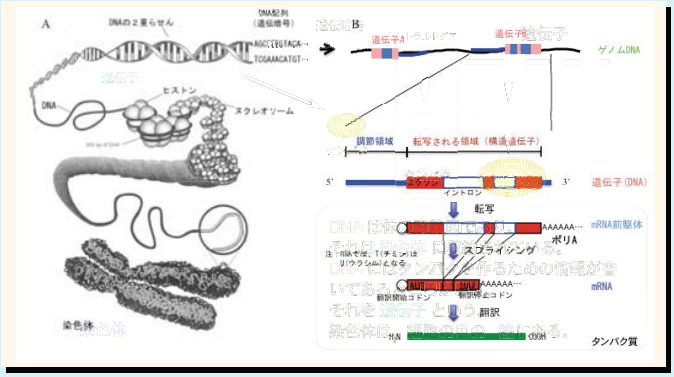
親から子へ



DNA=設計図



DNAと遺伝子と染色体とタンパク質

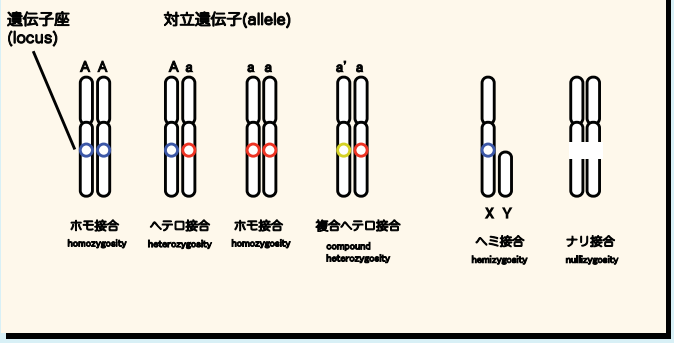


病気と遺伝子

病気と原因遺伝子との関係が解明されているものは、
遺伝子検査（解析）が可能

疾患	原因遺伝子
Aarskog-Scott症候群	*FGD1*
ARC症候群	*VPS33B*
コステロ症候群	*K-RAS,*
筋強直性ジストロフィー	*DMPK*
Rubinstein-Taybi症候群	*CBP, P300*
プロゲリア症候群	*Lamin A/C, ZMPSTE24*
	Filamin A
骨形成不全症候群	*COL1A1, COL1A2,*
点状軟骨異形成症	*CDPX1*
列手裂足異常症	*TP73*

遺伝性疾患と対立形質（名称）



常染色体性遺伝

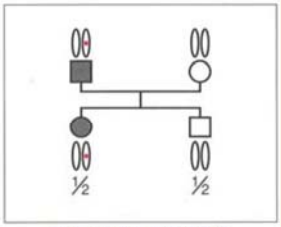


図６　常染色体優性遺伝

所謂一般的なイメージの遺伝性難聴
家族内発症を認める
進行性難聴が多い

図５　常染色体劣性遺伝

遺伝子が関与する難聴の約70%
子は1/4の確率で難聴になる
両親に難聴がない場合が多い

X連鎖性遺伝　ミトコンドリア遺伝

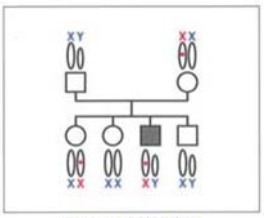


図７　X連鎖性遺伝

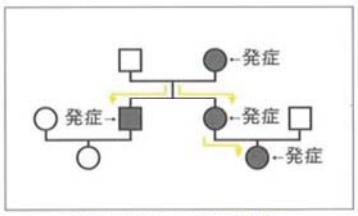


図８　ミトコンドリア遺伝

X連鎖劣性遺伝では男子のみに
難聴が認められる
難聴＋先天性アブミ骨固着＋gusher
DFN３（POU3A4遺伝子）

ミトコンドリアDNAの変異
母系遺伝

1555A>G（アミノ配糖体高感受性）
感音性難聴の約3%

メンデルの法則

Gregor J. Mendel
(1822-1884)

法則	内容
分離の法則	アリルの伝達は等確率
優劣の法則	表現型は遺伝子型の関数
独立の法則	アリルごとの伝達は独立

AA x aa
↓
Aa　x Aa
↓
AA　Aa　Aa　aa
1 : 1 : 1 : 1
(A)　3　: (a) 1

AaBb x AaBb
↓
AB　Ab　aB　ab
9 : 3 : 3 : 1

Hardy-Weinberg平衡

定理

大きな任意婚集団で、特定の遺伝子型に対する淘汰がなく、突然変異率が一定であると各遺伝子型の割合は世代を通じて変化しない。

二つの対立遺伝子A, a
遺伝子（型）A の頻度をp，遺伝子（型）a の頻度をqとすると
$p+q=1$である。

常染色体劣性遺伝の場合

$(p+q)^2=1$

頻度　$p^2+2pq+q^2=1$

遺伝子型　AA　Aa　aa

表現型　非患者　患者

♂＼♀	p(A)	q(a)
p(A)	p^2(AA)	pq(Aa)
q(a)	pq(Aa)	q^2(aa)

フェニルケトン尿症の頻度

フェニルケトン尿症の英国での発症頻度は１万出生に対し
約１人である。（常染色体劣性遺伝）
保因者の頻度はどのくらいと推定されるか。

常染色体性遺伝

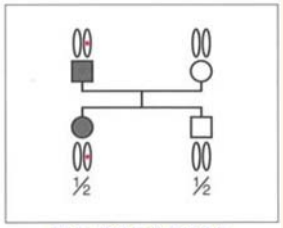


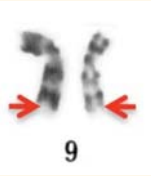
図６　常染色体優性遺伝

図５　常染色体劣性遺伝

所謂一般的なイメージの遺伝性難聴
家族内発症を認める
進行性難聴が多い

遺伝子が関与する難聴の約70%
子は1/4の確率で難聴になる
両親に難聴がない場合が多い

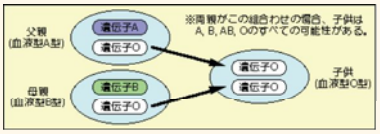
血液型（ABO）の遺伝



血液型	遺伝子型
O	OO
A	AA、AO
B	BB,BO
AB	AB

AはOに対して優性
BはOに対して優性

AとBは同じように
形質（性質）が現れる



AO→A型
BO→B型
AB→AB型
OO→O型

血液型の発見

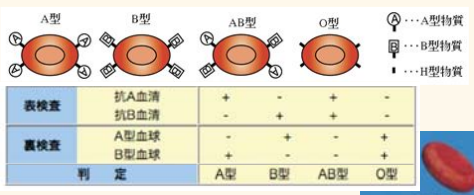
1901年　血液型の発見　A, B, C

Karl Landsteiner
(1868-1943:墺)

1902年　AB型の追加　A, B, C, AB
1910年　A, B, AB, Oに変更

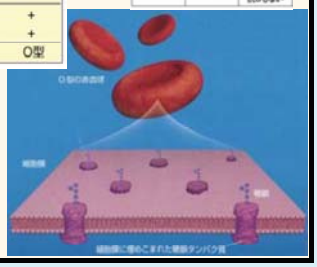
1990年　ABO式血液型決定遺伝子の特定
９番染色体上の糖転移酵素

血液型物質



表検査	抗A血清	+	-	+	-
	抗B血清	-	+	+	-
裏検査	A型血球	-	+	-	+
	B型血球	+	-	-	+
判　定		A型	B型	AB型	O型

血液型	赤血球の抗原	血清中の抗体
A	A	抗B
O	A も B も ない	抗A と 抗B
B	B	抗A
AB	A と B	抗A も 抗B も ない

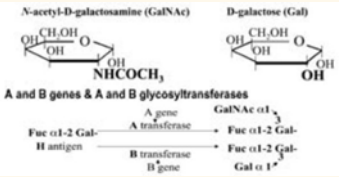


血液型物質は糖鎖である

糖鎖：糖がつながったもの

糖：ブドウ糖、果糖、ガラクトースなど

単糖類



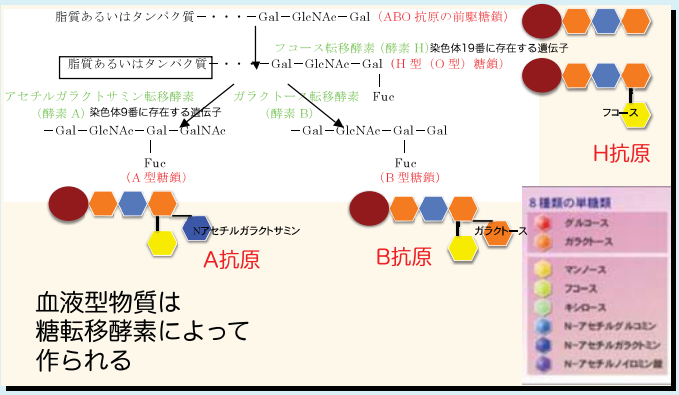
血液型物質は糖鎖である

糖鎖：糖がつながったもの

糖：ブドウ糖、果糖、ガラクトースなど

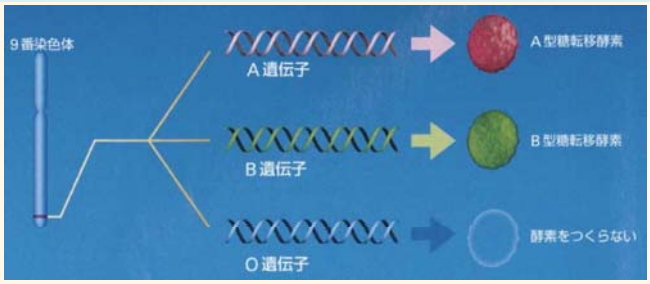
グルコース（穀物類など）
ガラクトース（乳製品など）
マンノース（アロエ、サボテンなど）
フコース（海藻類など）
キシロース（樹枝の皮など）
N－アセチルグルコミン（エビ、カニの甲羅など）
N－アセチルガラクトミン（牛乳、鮫軟骨など）
N－アセチルノイロミン酸（燕窩(ツバメの巣)、母乳など）

ABO式血液型物質の産生



血液型物質は
糖転移酵素によって
作られる

ABO式血液型決定遺伝子



O->A->B->AB

ABO式血液型決定遺伝子

Various Alleles at the ABO Locus

Exon Number	6								7									
Nucleotide Position	224	256	266	267	268	271	297	298	771	796	802	803	901	902	703	804	946	1060
A alleles																		
A101	G	A	C	C	T	C	G	G	C	C	G	G	G	G	G	C	C	
A102	+	+	T	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
A201	+	+	T	+	+	+	+	+	+	+	+	+	+	+	+	+	A	
A301	+	+	+	+	+	+	+	+	+	+	+	+	A	+	+	+	+	
Ax01	+	+	+	A	+	+	+	+	+	+	+	+	+	+	+	+	+	
Cis-AB01	+	+	T	+	+	+	+	+	+	+	+	C	+	+	+	+	+	
B alleles																		
B101	+	G	+	G	+	T	+	A	+	A	+	C	+	+	A	+	+	
B301	+	G	+	G	+	T	+	A	+	A	+	C	+	+	A	T	+	
BCA101	+	G	+	G	+	+	+	A	+	A	+	C	+	+	A	+	+	
O alleles																		
O1	A	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
O2	A	G	+	+	A	+	A	+	T	+	+	+	A	+	+	+	+	
O3	+	G	+	G	+	+	+	+	+	+	+	A	+	+	+	+	+	
Possible Amino Acid Change	Frameshift	No change	P156L	R176G	F216I	No change	No change	G235S	No change	L266M	G268R	G268A	V277M	D291N	No change	R352W	Frameshift	

A → → → B
→ O

O->A->B->AB

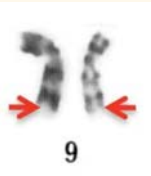
ABO式血液型決定遺伝子

同じ遺伝子の配列の違いによって、基質特異性、活性が変わっている

O->A->B->AB

A → → → B
→ O

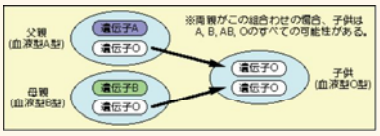
血液型（ABO）の遺伝



血液型	遺伝子型
O	OO
A	AA、AO
B	BB,BO
AB	AB

AはOに対して優性
BはOに対して優性

AとBは同じように
形質（性質）が現れる



AO→A型
BO→B型
AB→AB型
OO→O型

血液型の分布

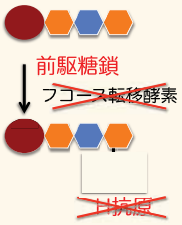
表3　ABOとRh血液型の割合

ABO血液型	Rh血液型	日本人の割合	
A	Rh＋	39.80%	約10人に4人
	Rh－	0.20%	約500人に1人
O	Rh＋	29.85%	約10人に3人
	Rh－	0.15%	約670人に1人
B	Rh＋	19.90%	約10人に2人
	Rh－	0.10%	約1,000人に1人
AB	Rh＋	9.95%	約10人に1人
	Rh－	0.05%	約2,000人に1人

□O型 □A型 □B型 ■AB型

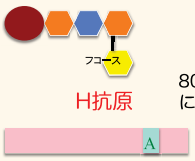
特殊な血液型

ボンベイ型



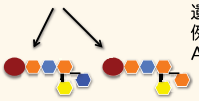
H抗原がないので
見かけ上O型のように見えるが
抗H抗体をもつ
（非常に少人数）

cisAB型



80万～ 100万人
に一人

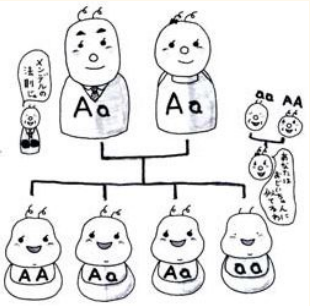
基質特異性：GalNAc、Galの両方



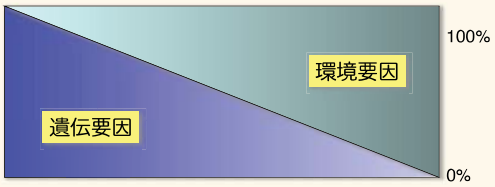
遺伝子型は
例えば
AB/Oとなる

遺伝

親から子へ受け継がれる形質



遺伝は多くの病気に関わっている



単一遺伝子病 ⇒ 多因子遺伝病（体質、心疾患、癌、高血圧、動脈硬化、糖尿病）、感染症、中毒・外傷

ゲノムの違いを調べることが重要

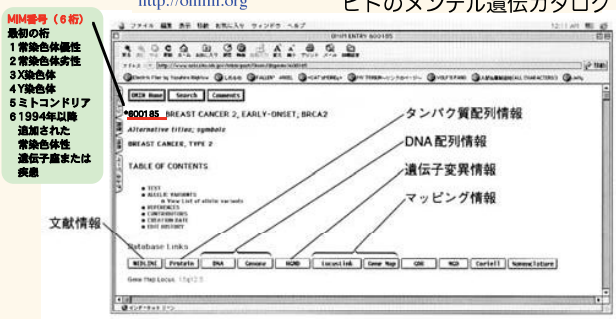
OMIM

私が作りました

OMIM(Online Mendelian Inheritance in Man)

http://omim.org

ヒトのメンデル遺伝カタログ



OMIA

http://omia.angis.org.au/home/

http://www.ncbi.nlm.nih.gov/omia

いろいろな正常形質

優性　劣性　　優性　劣性

アセトアルデヒド脱水素酵素（ALDH）

お酒に強い　　お酒に弱い

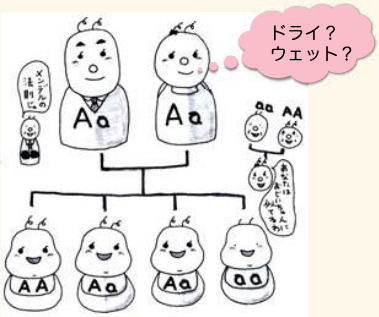
人類遺伝学上の重要な形質

- 肌の色
- 髪の毛の色
- 目の色（虹彩の色）
- 血液型
- 耳あか型
- 味覚感受性（PTC感受性）

- 嗅覚感受性
- その他

遺伝

親から子へ受け継がれる形質



耳あか型

耳あか型は、湿型と乾型があり、湿型が優性である

形質はW,wで表す

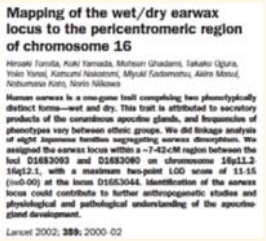
WW　湿型（しっとりタイプ）
Ww　湿型（しっとりタイプ）
ww　乾型（かさかさタイプ）

世界的には湿型がほとんど
（９０％以上）
乾型はアジア特有

耳あか型の原因の発見

長崎大学医学部（新川詔夫教授）で学生実習
をしていた時に、偶然原因の糸口が見つかる

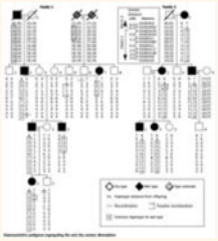
学生実習：正常形質について家族、親戚を調べて家系図を作る


Mapping of the wet/dry earwax locus to the pericentromeric region of chromosome 16

Hiroaki Tanida, Koki Yamada, Mohsen Ghadami, Takaku Ogura, Yoko Yanai, Katsumi Nakatomi, Miyuki Kadomatsu, Akira Masui, Nobumasa Kato, Norio Niikawa

Human earwax is a one-gene trait comprising two phenotypically distinct forms—wet and dry. This trait is attributed to secretory products of the conundrous apocrine glands, and frequencies of phenotypes vary between ethnic groups. We did linkage analysis of eight Japanese families segregating earwax dimorphism. We assigned the earwax locus within a ~7-42-cM region between the loci D16S3093 and D16S3080 on chromosome 16q11.2-16q12.1, with a maximum two-point LOD score of 11.15 (θ=0.00) at the locus D16S3044. Identification of the earwax locus could contribute to further anthropogenetic studies and physiological and pathological understanding of the apocrine-gland development.

Lancet 2002; 359: 2000-02



PKC（発作性運動誘発性ディスキネジア）と耳あか型が一致

耳あか型決定遺伝子

耳あかタイプ（湿型，乾型）は
６０億個の塩基のうち、たった１つの違いで決まる。

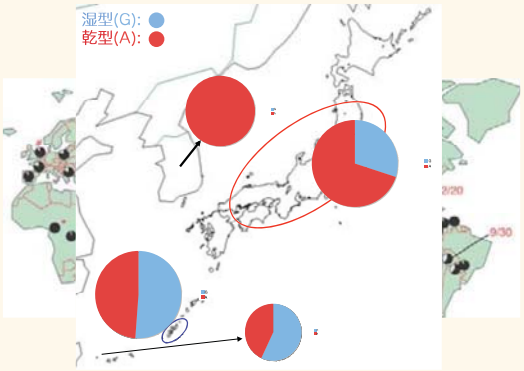
このタイプは、体臭やある種の薬の代謝に関係している可能性がある。

```
                =R=
     ATTGCCAGTGTACTCAGGCCAATATTGATTATACCAA

601  ATTGCCAGTGTACTCGGGCCAATATTGATTATACCAA   DNA
175  -I--A--S--V--L-=G=-P--I--L--I--I--P--   アミノ酸
```

*ABCC11*遺伝子@16q12.1

Gタイプ（G/G, G/A）→ 湿型
Aタイプ（A/A）　→ 乾型

耳あかタイプの分布



SNP

SNP: Single Nucleotide Polymorphism　１塩基多型

- **１塩基の多様性（個人差）**　人口の１％以上で異なる
- **約1kbに一つ存在すると推定されている（＝ヒトゲノム全体で約300万箇所）**
- **ジーンチップやマイクロプレートと合わせて大量のタイピングが可能（⇨コンピューターによる解析）**
- **個別医療（オーダーメイド医療，テーラーメイド医療）の判断材料**

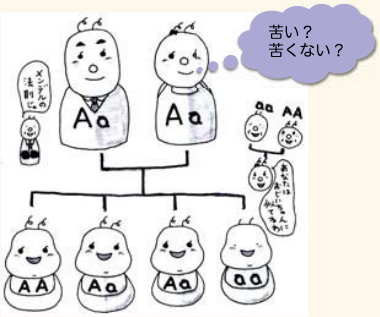
SNPの種類

種　類	説　明	ゲノムDNA上の部位
cSNP（coding SNP）	一塩基置換によりアミノ酸の種類が変化する	翻訳領域（アミノ酸変換領域）
sSNP（silent SNP）	一塩基置換は起こるが，アミノ酸の種類は変化しない	
rSNP（regulatory SNP）	一塩基置換により遺伝子の発現量に影響を与える可能性あり	調節領域（プロモーター領域）
iSNP（intronic SNP）		調節領域（イントロン領域）
uSNP（untranslated SNP）		非翻訳領域（非アミノ酸変換領域）
gSNP（genomic SNP）	一塩基置換は起こるが，遺伝子の発現量に影響を与える可能性は低い	他の領域（非遺伝子領域）

遺伝

親から子へ受け継がれる形質

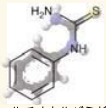
味覚



味覚の遺伝

Du Pont (1929)

1931年、アメリカのデュポン社の
アーサー・フォックスが
フェニルチオカルバミド(PTC)という
化合物を瓶に移そうとしたところ、
一部が飛び散ってしまった。



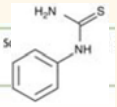
フェニルチオカルバミド(PTC)

すると研究室内の同僚が「にがっぽいほこりだ」と苦情を述べた。
彼は、今まで全く感じたことがなかったので、PTCをなめて全然味がしないと主張した。
ところが同僚は「苦いに違いない」といって実際になめ、「とんでもなく苦い」と言い出した。
論争で決着がつかないので、別の研究室の人を呼んできて、どちらの味覚がおかしいのか、決めてもらおうとしたが、全く味がしないというひとと、苦いと言う人に分かれた。
（苦いという人が多かった）

PTCの味を感じない形質を便宜的に「味盲」とよんだ

PTC味覚感受性遺伝子

メンデル遺伝カタログ (http://www.omim.org)



#171200

THIOUREA TASTING

PTCの味を感じる形質

Other entities represented in this entry:

PHENYLTHIOCARBAMIDE TASTING, INCLUDED

PTC TASTING, INCLUDED
PROPYLTHIOURACIL TASTING, INCLUDED
PROP TASTING, INCLUDED

T：苦いと感じる　優性
t：感じない（味盲）劣性

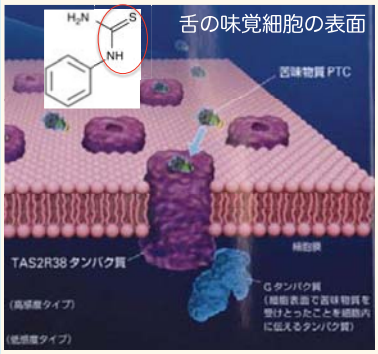
TT　有味者
Tt　有味者
tt　味盲

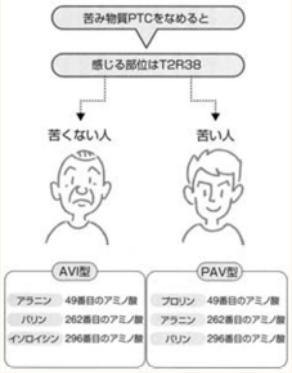
Phenotype Gene Relationships

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
7q34	[Phenylthiocarbamide tasting]	171200	TAS2R38	607751

Clinical Synopsis

PTC味覚感受性タンパク質





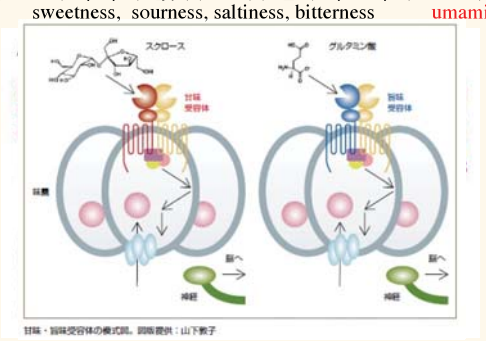
PTCに対する味盲の頻度

人種によって味盲の割合が異なっている

人種	味盲の頻度（%）
日本人	8
韓国人	3
中国人	6～11
米国白人	25～35
ドイツ人	38
米国黒人	9
アフリカ黒人	3～9
米国先住民	2

味覚受容体遺伝子

基本五味：甘味（1）、酸味、塩味、苦味（25）、うま味（1）
sweetness, sourness, saltiness, bitterness　umami

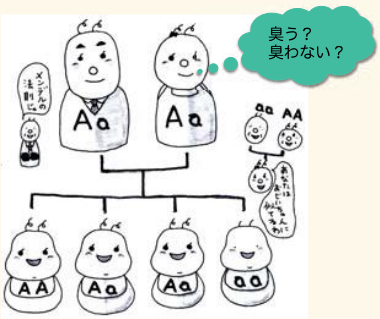


甘味・旨味受容体の模式図。図版提供：山下敦子

遺伝

親から子へ受け継がれる形質

味覚



アスパラガス遺伝？

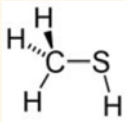
変な話ですが、、、

アスパラガスを食べるとおしっこが臭う方はいませんか？

アスパラガス特異的におい高感受性遺伝

メンデル遺伝カタログ OMIM

アスパラガス



メタンチオール

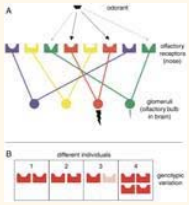
およそ５人に一人いるようです（人種差あり）

におう？
におわない？

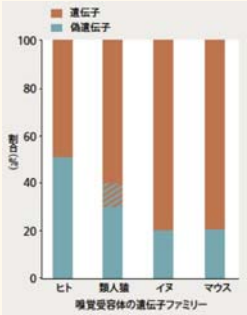
におい受容体遺伝子

およそ900〜1500種類

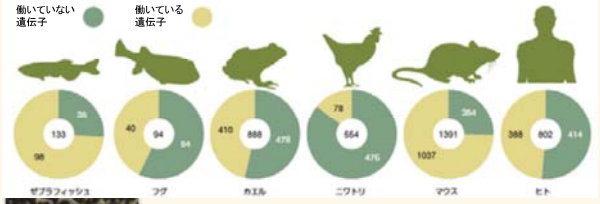
Olfactory receptor (OR) proteins interact with odorant molecules in the nose, initiating a neuronal response that triggers the perception of a smell. The OR family is one of the largest known mammalian gene families, with around 900 genes in human and 1500 in mouse.



嗅神経



におい受容体遺伝子



ヒトでは、およそ800の遺伝子のうち、約半分が働いている

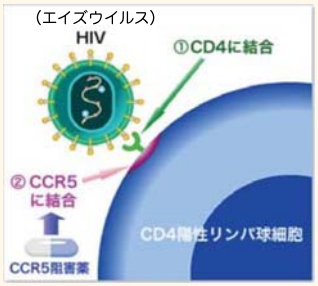
トマト遺伝？

またまた、変な話ですが、、、

Tomato tasting
(OMIM ???)

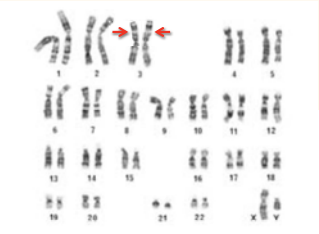
もしおられたら
ぜひご一報ください

HIV感染症（AIDS）



CCR5:ケモカイン受容体５

*CCR5*遺伝子 (3p21.31)



*CCR5*遺伝子に変異がある人（デルタ32）

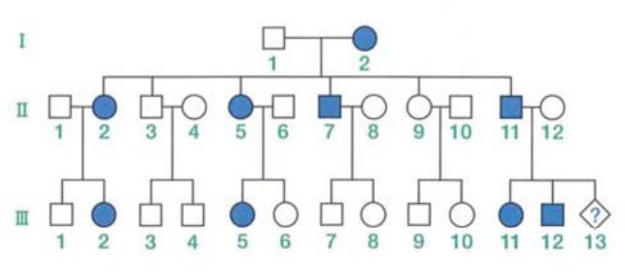
白人の約5~14%

CCR5遺伝子検査（エイズにかかりにくいか診断）

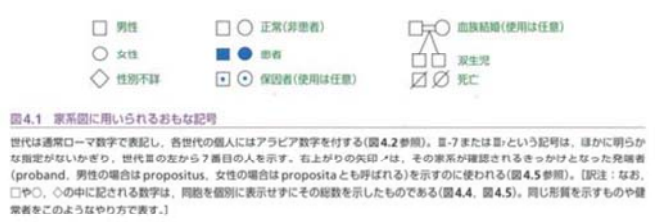
その他の正常形質の遺伝

「これって遺伝？」を調べるには？

家系図が大事です

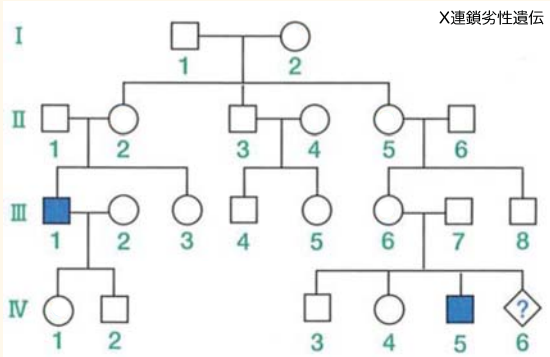


家系図の記号

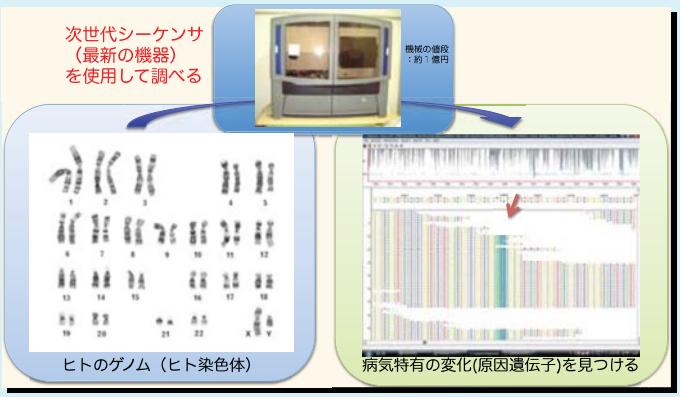


・男性は□、女性は○、分からないときは◇
・出生の早い順に左から書く

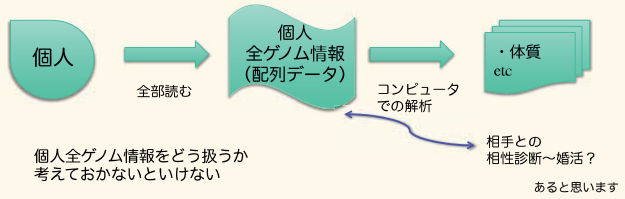
家系図の一例



パーソナルゲノム時代を拓く次世代シーケンサ



パーソナルゲノムの時代

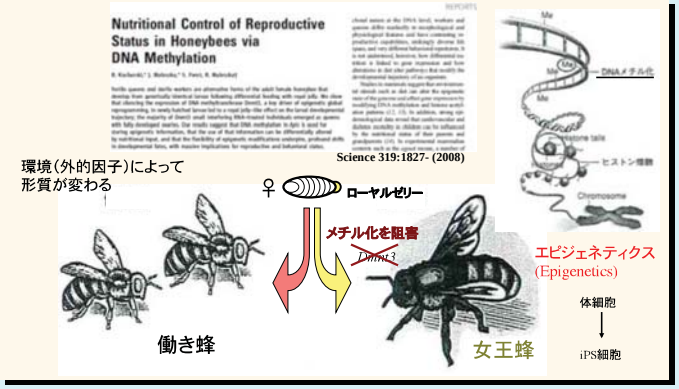


個人全ゲノム情報をどう扱うか
考えておかないといけない

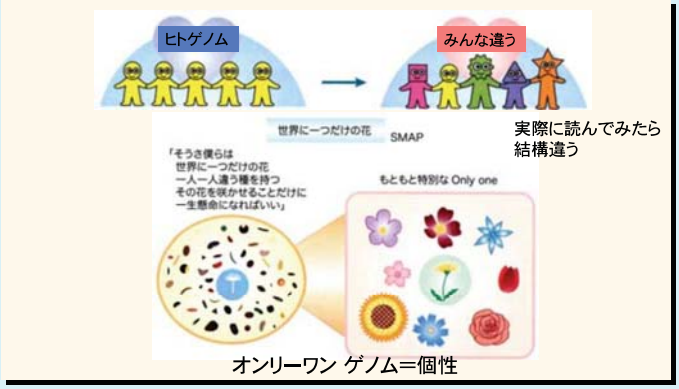
パーソナルゲノムの時代へ向けて（個人的意見）

・個々のゲノムは変わらない、、、でも遺伝子発現・表現型は？
・オンリーワン ゲノム
・ゲノムに関心を持とう！

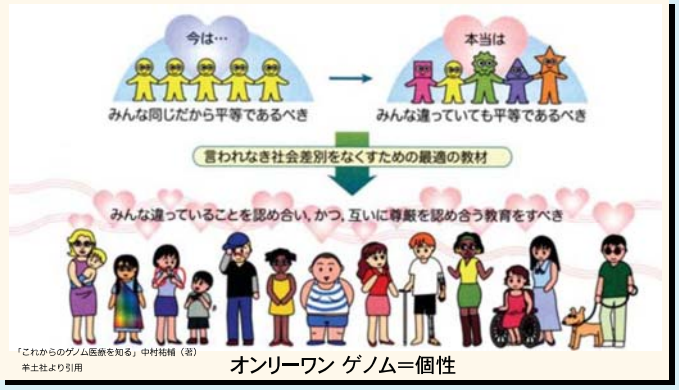
食事で形質が変わる？



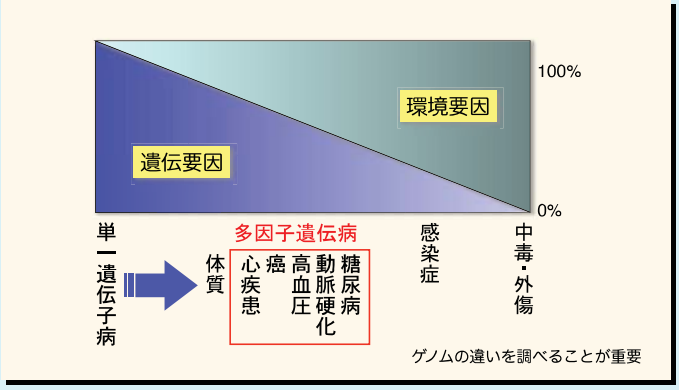
世界にたった一つだけのゲノム



みんな違ってみんないい



遺伝は多くの病気に関わっている



正常形質の遺伝

未解決の正常形質（遺伝）

一重まぶた、二重まぶた
長顎（ハプスブルグ家の顎）
富士額
巻き舌
その他

生トマト嫌い？

2012年9月3日

バイオインフォマティクス　レポート課題

学部・学科：＿＿＿＿＿＿＿＿＿＿　学籍番号：＿＿＿＿＿＿　氏名：＿＿＿＿＿＿＿＿

1．組換えDNA技術を使うことで、例えばオワンクラゲの緑色蛍光タンパク質（GFP）を、種の異なるマウスや大腸菌で作らせ、光らせることが可能となった。なぜこのようなことが可能なのか簡潔に説明せよ。

2．ある両親から生まれる子の血液型が、A型、B型、AB型、O型、いずれの可能性もあるのは、その両親の血液型（遺伝子型）がどのような場合か。また、なぜこのような可能性が生じるのか簡潔に説明せよ。

3.2　生命科学実習①：ゲノムDNA抽出、総RNA抽出

3.2.1概要

講師：池松 真也／沖縄工業高等専門学校生物資源工学科教授

日時：2012年9月4日　10:20～／12:50～／14:40 ～

場所：沖縄工業高等専門学校 SE-271 生物化学工学実験室

内容：生命科学実習①：生命科学研究において利用される基本的な実験技術として、細胞からのDNA抽出および制限酵素処理を加えて電気泳動で分離する処理について講義および実習を交えて学ぶ。生命科学実習②：ポリメラーゼ連鎖反応（PCR）と、DNAシーケンサによる配列解析について、その原理の理解を中心に実習を交えて学ぶ。

ＴＡ：7名

実習体制：4名/班×4班

3.2.2準備

機器

□卓上遠心機	himac CT 13R／HITACHI
□卓上遠心濃縮機	Concentrator 5301／eppendorf
□小型恒温槽	CTU-mini／TAITEC
□ローテーター	Rotator RT50／TAITEC

実験器具・試料・試薬

□マウス肝臓	0.075g×4個（肝臓1/3程度使用）
□エッペンチューブ	12個（3個/人×4人）
□15mLチューブ	4個（試薬分注用）
□p20ピペットマン	1個
□p200ピペットマン	1個
□イエローチップ	1箱
□解剖用ハサミ	1個
□ピンセット	1個
□保護ゴーグル	4個
□手袋	4個
□マスク	4個
□Lysisバッファー	2mL(0.5mL/人×4人)
□proteinaseK	240μL(60μL×4人)
□Tris/フェノール	2mL（0.5mL×4人）
□冷100%エタノール	4mL(1mL×4人)

□2-プロパノール　　4mL(1mL×4 人)
□70%エタノール　　適量 4mL(1mL×4 人)
□PBS(-)　　適量

3.2.3 配布物
・講義スライド
・実習プロトコル

3.2.4 課題
・プリント

3.2.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
<u>分子生物学データベース</u> 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
<u>配列解析</u> アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
<u>タンパク質立体構造解析</u> 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
<u>遺伝学・進化学</u> 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
<u>トランスクリプトーム解析・プロテオーム解析</u> 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
<u>パスウェイ解析・システム生物学</u> ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.2.6 時間割

時間	内容
10:26～	<u>イントロダクション</u> ・ 前日の復習 ・ セントラルドグマ ・ 生き物の体は細胞からできている ・ プロトコル説明
10:46～	【DNA の抽出】 マウス肝臓　0.075g ↓　PBS(−)で血を洗い流す。 ＊今回は、−80℃で保管してあったものを使用 ↓　組織を解剖用ハサミで細かく刻みチューブに移す ↓　Lysis バッファー0.5mL、Proteinase K 溶液 60μL を加える ↓　フタをして、均一になるまで上下転倒混和
11:05～	↓　55℃、8 時間～O/N でインキュベート ＊　今回は、前日からインキュベートしてあるものと交換
11:12～	↓　Tris/フェノールを 0.5mL 加える ↓　30rpm で 1 時間上下転倒 ＊　空き時間に課題に取り組んだ。
11:50～	インキュベート中に昼食
13:10～	↓　RT、3,000rpm、10 分間 ↓　上清を慎重に回収し、新しいチューブに移す ＊　今回は、100μL 程度を回収 ↓　200μL（2 倍量）の 100%エタノールまたは 2-プロパノールを慎重に重層 ↓　竹ぐしを、水槽とエタノール層の界面に挿入し、析出した DNA を巻き取る ＊　今回は十分に回収できなかったため、上清を回収し 100%エタノールを 2 倍量加え、4℃、10,000rpm、3 分間遠心分離 ↓　沈殿を DNA として回収
13:30～	RNA 抽出について ・ 前半の復習 ・ RNA の種類（rRNA、tRNA、mRNA、） ・ RNA 抽出に関する注意事項 ・ プロトコル概要紹介

13:50～	【映像】 細胞からの RNA 抽出ビデオ 14:00～ 【細胞破砕液の調整】 マウス腎臓 1g 弱／班 ↓　＋TRI-Zol (GIBCO) 5mL ↓　ホモジェナイズ 細胞破砕液とする 14:10～ 【Total RNA の抽出】 細胞破砕液 1mL ↓　＋0.2 mL クロロホルム溶液 ↓　チューブの蓋を閉めて、15 秒間激しく浸透 ↓　15℃～30℃にて 2～3 分間インキュベート 14:25～ ↓　4℃、15 分間、12,000g
14:25～	遠心分離中に休憩
	↓　3 層のうち、一番上の層（透明）を新しいチューブに回収する。 ＊使用した TRI-Zol の約 60%程度 ↓　使用した TRI-Zol 1mL に対してイソプロピルアルコール 0.5mL を加え、よく混和する ↓　15℃～30℃にて 10 分間インキュベート 14:50～ ↓　4℃、12,000g 、15min 15:05～ ＊ 遠心後、チューブの壁面から底にかけて白いペレットが確認できる 15:25～ ↓　上清を捨てる ↓　75% EtOH で RNA ペレットを洗う ＊使用した TRI-Zol 1mL に対して 1mL の 75% EtOH を使用 ↓　ボルテックスミキサーでサンプルを混ぜる ↓　4℃、7500g、5min 15:35～ ↓　5～10 分間、Vacuum-Dry 15:48～ ↓　10～100μL の RNase free 滅菌水に溶解 15:58～ ↓　nanoDrop にて吸光度（260nm/280nm）を測定する 16:18～ 課題作成／提出の上解散

バイオインフォマティクス人材育成講座

遺伝子工学とその応用例

ゲノム DNA、総 RNA の抽出

■　本日の実習内容　■

１．　マウス臓器からのゲノム DNA の抽出

２．　マウス臓器からの総 RNA 抽出

2012 年 9 月 4 日

沖縄工業高等専門学校生物資源工学科

池松真也・渡邊謙太・祖納元りえ・當山瑛子

伊波大志・大城龍之助・小松圭・瑞慶覧安里

食品機能科学特別講義Ⅰ　2回目 A.M.

遺伝子工学とその応用例
ゲノムDNAの抽出

4 September 2012

Shinya Ikematsu
Okinawa National College of Technology
Director of Cooperative Technology Center

ゲノムDNAの調製

～マウスの肝臓と腎臓から～

ゲノムDNAの調製の概略

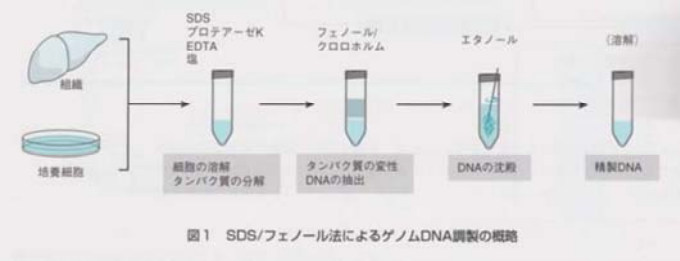


図1　SDS/フェノール法によるゲノムDNA調製の概略

遺伝子工学実験ノート　下　P100

プロトコル

①肝臓を摘出し、PBS(－)で血を洗い流す。今回は－80℃で保管してあったものを使用する。
重さを0.075g量り取る。

②氷上にアルミホイルを敷き、組織を解剖用ハサミで細かく刻んで、エッペンドルフチューブに移す。

遺伝子工学実験ノート　下　P103

プロトコル

③リシスバッファー0.5ml、Proteinase K溶液60μl（組織25mgにつき20μl、脾臓の場合は10mgにつき20μl）を加える。

④フタをして、上下転倒して全体が均一になるまでよく混合し、55℃で振盪しながら8時間～一晩インキュベート。

⑤Tris/フェノールを0.5ml 加え、30rpmで1時間上下転倒。

プロトコル

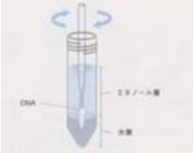
⑥3,000rpm、10分間、室温で遠心する。

⑦広口のピペットで水層を慎重に回収し、新しいチューブに移す。

⑧1ml(2倍量)の100%エタノールまたは2－プロパノールをチューブの壁面に沿わせながら静かに重層する。

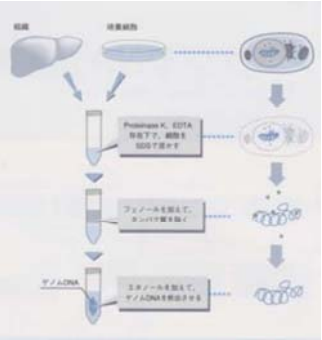
プロトコル

⑨竹ぐしを、水層とエタノール層の界面に挿入して、ゆっくりかき回し、析出したDNAを巻きとる。



⑩エッペンドルフチューブに入った70%エタノールに入れ、DNAをリンスする。

（まとめ）ゲノムDNAの調製



マウスの解剖　肝臓と腎臓を摘出
　-80℃で保存

SDS　プロテアーゼK　EDTA　塩
　→細胞の溶解、タンパク質の分解

フェノール/クロロホルム
　→タンパク質の変性、DNAの抽出

エタノール
　→DNAの沈殿

（溶解）→精製DNA

遺伝子工学実験ノート　下　P102

食品機能科学特別講義 I
＜生命科学実習① A.M.＞　　実習場所：沖縄工業高等専門学校

ゲノム DNA の調製　～マウス組織から～

実験日：＿＿＿＿＿＿＿＿＿＿＿＿＿＿　所属：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

学籍番号：＿＿＿＿＿＿＿＿＿＿＿＿＿＿　氏名：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

■実験機器，器具

1.)エッペンドルフチューブ：□
2.)微量ピペット（P-20）：□
3.)微量ピペット（P-100 または P-200）：□
4.)低速遠心機：□
5.)解剖用ハサミ：□

■試薬

1.）リシスバッファー：□
2.) Proteinase K：□
3.) Tris/フェノール：□
4.）冷 100%エタノール：□
5.) 2-プロパノール：□
6.) 70%エタノール：□
7.) PBS(－)：□

■実験方法

①肝臓を摘出し、PBS(－)で血を洗い流す。今回は－80℃で保管してあったものを使用する。□
　重さを 0.075g（＿＿＿＿g）量り取る。□
②氷上にアルミホイルを敷き、組織を解剖用ハサミで細かく刻んで、エッペンドルフチューブに移す。□
③リシスバッファー0.5ml（＿＿＿＿ml）、Proteinase K 溶液 60μl（＿＿＿＿μl）（組織 25mg につき 20μl、脾臓の場合は 10mg につき 20μl）を加える。□
④フタをして、上下転倒して全体が均一になるまでよく混合し、55℃で振盪しながら 8 時間～一晩インキュベート。（＿＿＿＿：＿＿＿＿～＿＿＿＿：＿＿＿＿）□
⑤Tris/フェノールを 0.5ml（＿＿＿＿ml)加え、30rpm で 1 時間上下転倒。□
⑥3,000rpm、10 分間、室温（＿＿＿＿＿＿rpm、＿＿＿＿分、＿＿＿＿℃）で遠心する。□
⑦ピペットで水層を慎重に回収し、新しいエッペンドルフチューブに移す。□
⑧1ml（＿＿＿＿ml）(2 倍量)の 100%エタノールまたは２－プロパノールをチューブの壁面に沿わせながら静かに重層する。□
⑨竹ぐしを、水層とエタノール層の界面に挿入して、ゆっくりかき回し、析出した DNA を巻きとる。□
⑩70%エタノールの入ったエッペンドルフチューブに移し、DNA をリンスする。□

リシスバッファーの組成

①リシスバッファー

		最終濃度
・1M Tris-HCl (pH8.0)	5mL(__25__mL)	50mM
・5M NaCl	2mL(__8__mL)	100mM
・0.5M EDTA	20mL(__100__mL)	100mM
・10%(w/v)SDS	10mL(__50__mL)	1%
DW up to	100mL(__500__mL)	

食品機能科学特別講義Ⅰ　2回目 P.M.

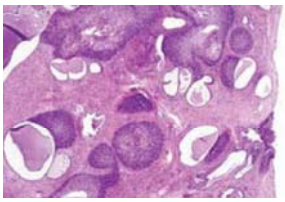
遺伝子工学とその応用例
総RNAの抽出

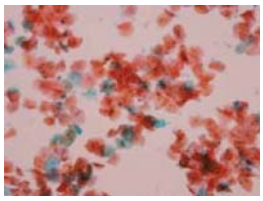
4　September 2012

Shinya Ikematsu
Okinawa National College of Technology
Director of Cooperative Technology Center

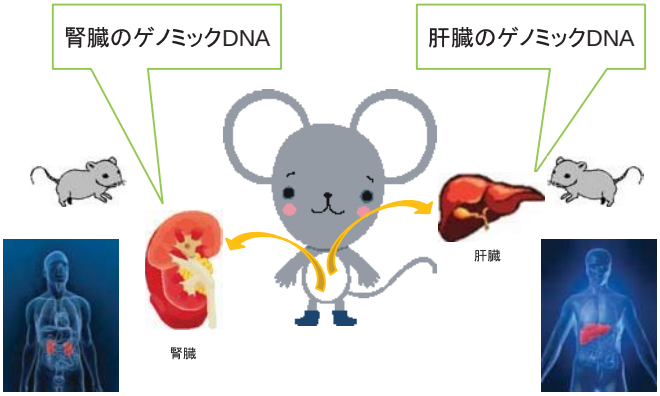
2012年9月4日（火）P.M.

1．マウス臓器から全RNAの取得

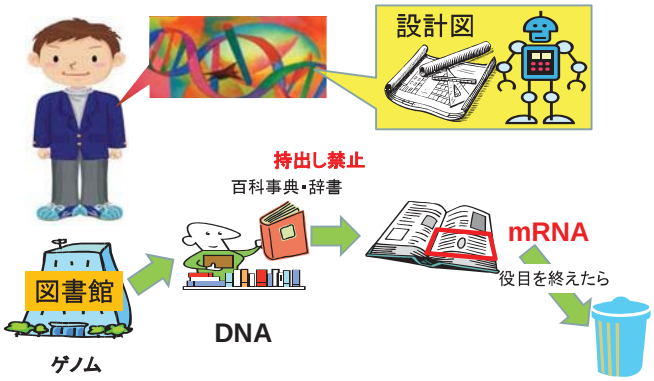




午前中の復習



午後の実験は



RNAの種類

リボソームRNA
（ribosomal RNA, r RNA）：約80%
転移RNA
（transfer RNA, t RNA）
伝令RNA
（messenger RNA, m RNA）：　5%
マイクロRNA（miRNA）
など

実験を行う上での注意点　1

☆安全に充分留意する

非常に危険な薬品を使います。

- TRIZOL Reagent　⇒　フェノール・グアニジンイソチオシアネート（タンパク変性剤）
- クロロホルム　⇒　人体に有害な劇物

対策

・必ず白衣・保護メガネ・ラボグローブを着用する。

・試薬チューブのふたはしっかりと閉める。

・廃液の管理を徹底する。

実験を行う上での注意点　2

☆RNAの取り扱いに注意する

RNAは（DNAと比べても）非常に不安定

⇒壊れやすく、分解酵素も多い

・RNaseの存在

一般的対策

・必ず白衣・ラボグローブ・（マスク）を着用する

・作業環境を正常に保つ

・作業を迅速に行う

実際の実験の流れをビデオで確認しましょう

7分

☆実験をうまく行うコツ

分離した3層のうち、
RNAを含む水層を的確に回収する

☆結果の解釈

A_{260} / A_{280} > 1.8

260nm　：　核酸の吸光度　極大

280nm　：　タンパク質の吸光度　極大

！注意！

- 手袋やゴーグルを使用し、皮膚や衣類にTRI-Zolがつかないようにすること！
- TRI-Zolを吸い込まないこと！

手順1. RNA含有試料溶液の作製

- 試料：マウス臓器（肝臓または腎臓）
  - ＊数mLのTRI-Zolで1個のマウス臓器の処理可能
  - ＊電動ホモジナイザーで処理

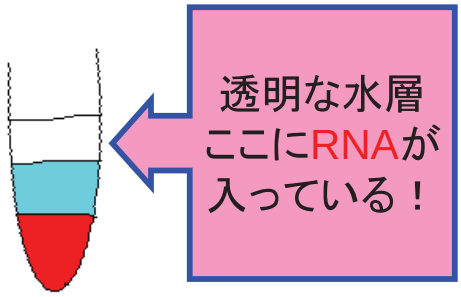
手順2

- 充分にホモジナイズした試料液を、15-30℃（室温）で、5分間インキュベートする。
- 試料液 1mLに対して0.2mLのクロロホルムを加えしっかりふたを閉める。

手順3

- 手で激しく15秒間、チューブを振る。その後、15-30℃で2-3分間インキュベートする。
- 2-8℃で、15分間、12000xg以上で遠心する。

手順3終了後



（使用したTRI-Zolの約60%）

手順4

- 新しいチューブに水層を移す。
- 試料液1mLに対して、（ホモジナイズしたときの）イソプロピルアルコール0.5mLを加え、よく混ぜる。

手順5

- 15-30℃で10分間インキュベートした後、12000xg以上で10分間（2-8℃で）遠心する。
- チューブの底にゲルのようなペレットが確認できる。

手順6

- 上澄みを捨てる。
- 75%アルコールでRNAペレットを洗う。試料液1mLについて、少なくとも1mLの75%アルコールを加える。

手順7

- ボルテックスミキサーでサンプルを混ぜる。
- 7500xgで5分間（2-8℃で）遠心する。

手順8（最後の操作）

- 上澄みを捨てる。
- 5-10分間、vacuum-dryする。
- RNase freeの滅菌水に再溶解し、吸光度を測定。
- A260/A280が1.8から2.0の間であることを確認する。

2012 年 9 月 4 日（火）　　於：沖縄工業高等専門学校

食品機能科学特別講義 I　第 2 回目

学部：＿＿＿＿＿学科：＿＿＿＿＿＿＿氏名：＿＿＿＿＿＿＿＿＿＿＿＿＿＿

＜講義課題＞

以下の質問に端的に解答せよ。

１．界面活性剤が細胞を破壊しやすくするためには、どのような工夫をすれば良いか。

２．細胞を破壊すれば DNA が見えるようになると思うか？どうしてそう思うのか、その理由も記載せよ。

３．細胞膜が溶解すると DNA が溶液中に放出されるが、細胞に含まれるたくさんのその他の物質も放出される。DNA 以外に細胞内にあると考えられる物質を列挙せよ。

４．肝臓にあるタンパク質をコードしている遺伝子を単離したい場合、その遺伝子を腎臓の細胞から見つけることができるか。また、その mRNA を見つけることができるか。理由を付して説明せよ。

2012 年 9 月 4 日（火）　　於：沖縄工業高等専門学校

食品機能科学特別講義 I　第 2 回目

学部：＿＿＿＿＿学科：＿＿＿＿＿＿＿氏名：＿模範解答＿＿＿＿＿

＜講義課題＞

以下の質問に端的に解答せよ。

１．界面活性剤が細胞を破壊しやすくするためには、どのような工夫をすれば良いか。

工夫可能な点を挙げてみると、界面活性剤が効果的に作用できるように温度を上げてみることである。温度を上げるとは、例えば、食器を洗浄する場合、冷水で洗うのと温水で洗うので、どちらの場合がきれいに洗浄できるかを考えてみると良い。

２．細胞を破壊すれば DNA が見えるようになると思うか？どうしてそう思うのか、その理由も記載せよ。

細胞を破壊しただけでは DNA が目に見えるようにはならない。まず、セルリシスバッファー（細胞溶解液）に含まれる界面活性剤に充分に溶かすことで「細胞膜」や「核膜」を壊し、プロテアーゼで DNA 結合タンパク質、DNA 分解酵素などを破壊したり、不活性化する。次に、塩化ナトリウム溶液を加え、DNA の電荷を中和し、良く冷却したアルコールを加えることで DNA が不溶化し、析出した DNA を目で見ることができるようになる。

３．細胞膜が溶解すると DNA が溶液中に放出されるが、細胞に含まれるたくさんのその他の物質も放出される。DNA 以外に細胞内にあると考えられる物質を列挙せよ。

一般的な細胞成分は、タンパク質、脂質、糖類および無機質（塩類）などである。

４．肝臓にあるタンパク質をコードしている遺伝子を単離したい場合、その遺伝子を腎臓の細胞から見つけることができるか。また、その mRNA を見つけることができるか。理由を付して説明せよ。

肝臓にあるタンパク質をコードしている遺伝子は、腎臓の細胞にあるゲノム DNA にも含まれている。しかし、腎臓では、肝臓特有のタンパク質をコードするメッセンジャー RNA は作られていない。なぜなら、肝臓特有のタンパク質を作る遺伝子のメッセンジャー RNA は、肝臓でしか発現していないのである。

3.3　生命科学実習②：PCR、DNA シークエンス

3.3.1 概要

講師：池松 真也／沖縄工業高等専門学校生物資源工学科教授

　　磯村 尚子／沖縄工業高等専門学校生物資源工学科助教

日時：2012 年 9 月 5 日　10:20～／12:50～／14:40 ～

場所：沖縄工業高等専門学校 SE-271 生物化学工学実験室

内容：生命科学実習①：生命科学研究において利用される基本的な実験技術として、細胞からの DNA 抽出および制限酵素処理を加えて電気泳動で分離する処理について講義および実習を交えて学ぶ。生命科学実習②：ポリメラーゼ連鎖反応（PCR）と、DNA シーケンサによる配列解析について、その原理の理解を中心に実習を交えて学ぶ。

ＴＡ：7 名

実習体制：4 名/班×4 班（※受講者 14 名）

3.3.2 準備

機器

□卓上遠心機	himac CT 13R／HITACHI
□卓上遠心濃縮機	Concentrator 5301／eppendorf
□小型恒温槽	CTU-mini／TAITEC
□ローテーター	Rotator RT50／TAITEC
□サーマルサイクラー	
□トランスイルミネーター	
□ボルテックスミキサー	
□吸光度計	
□Mupid-2plus	

実験器具・試料・試薬

□乳酸菌プレート	事前に単離したものを使用
□エッペンチューブ	12 個（3 個/人×4 人）
□15mL チューブ	4 個
□PCR チューブ	4 個
□PCR チューブたて	1 個
□p20 ピペットマン	1 個
□p200 ピペットマン	1 個
□イエローチップ	1 箱
□ブルーチップ	1 箱
□滅菌つまようじ	4 本

□手袋	4 個
□クラッシュアイス	1 個
□発泡スチロール箱	1 個
□パラフィルム	1 枚
□タイマー	1 個
□1.3%アガロースゲル	1 枚
□PrepMan Ultra Reagent	800μL（200μL×4 人）
□10μM primer Forward	2.0μL（0.5μL×4 人）
□10μM primer Reverse	2.0μL（0.5μL×4 人）
□2×buffer	50μL（12.5μL×4 人）
□2.5mM dNTP’s	8μL（2μL×4 人）
□KOD polymerase	2.0μL（0.5μL×4 人）
□dH_2O（滅菌水）	14μL（7μL×4 人）
□1×TAE　適量	350-450mL
□分子量マーカー	5μL
□loading バッファー	5μL
□染色液（GelRed）	適量
□DNA クリーナー	適量
□70%エタノール	適量

3.3.3 配布物

・講義スライド

・実習プロトコル

3.3.4 課題

・プリント

3.3.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
<u>分子生物学データベース</u> 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
<u>配列解析</u> アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
<u>タンパク質立体構造解析</u> 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
<u>遺伝学・進化学</u> 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
<u>トランスクリプトーム解析・プロテオーム解析</u> 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
<u>パスウェイ解析・システム生物学</u> ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.3.6 時間割

時間	内容
10:26～	<u>イントロ</u> ・ 本日の研究意義 亜熱帯での微生物スクリーニングについて 新規性、科学的根拠の必要性 ＊ 以降は磯村先生に引き継ぎ ・ 本日の目的 DNA 抽出から配列決定の流れを学ぶ ゆくゆくは自分一人でできることを目指す 操作に習熟し、トラブルシューティングができるまで ・ プロトコル説明 【乳酸菌の培養】 0.5% 炭酸カルシウム入り MRS Agar 培地を用いて、美ら海水族館熱帯ドリームランド内にてサンプリングを実施 事前にプレスクリーニング、単離を行ったプレートを使用した。 ＊ 事前にスクリーニングおよび単離まで実施しておいた 【乳酸菌からの DNA 抽出】
10:48～	PrepMan Ultra Reagent 200μL をチューブに分注 ↓ MRS Agar 培地のコロニーを滅菌幼児でつつき、懸濁する。 ＊ 必要に応じてボルテックスにて激しく懸濁
10:55～	↓ ヒートブロックにて 100℃、10 分間 ↓ 室温で 2 分間静置 ↓ 室温、13,000rpm、3 分間 上清を DNA 抽出液として PCR に使用 【PCR による 16srRNA（約 1500bp）の増幅】
11:20～	PCR チューブに以下の溶液を混合 ↓ DNA template（上清） 2μL ↓ 4μM Primer F 0.5μL ↓ Primer R 0.5μL ↓ 2×buffer 12.5μL ↓ 2mM dNTP’s 2μL ↓ KOD FX Pol. 0.5μL

	D.W. 7μL Total 25μL ↓ スピンダウン
10:52～	↓ 以下の条件にて PCR 反応 95℃ 3min 95℃ 30sec 62℃ 30sec ｝32cycle 68℃ 1min 68℃ 3min 4℃ ∞
11:52～	インキュベート中に昼食
12:55～	PCR について ・ DNA と RNA の違い ・ Tm 値について ・ DNA ポリメラーゼ、プライマーの働き 各温度で起こる現象を紹介 ・ 電気泳動原理紹介 Agarose は高分子、DNA の移動について ・ 機器の設置
13:18～	泳動槽用意、ゲルセット、バッファー添加 【電気泳動による PCR 産物の確認】
13:38～	PCR 産物 5μL をパラフィルムにとる ↓ 6×Loading buffer 1μL を加える ＊ 100bp マーカーも同様に用意する ↓ サンプルおよびマーカーをウェルに流しこむ
14:00～	↓ 泳動槽のフタをしめ、100V にて電気泳動開始（20 分程度） ＊ 絵移動時間中に並行して PCR 産物の精製を行う ↓ サンプルおよびマーカーをウェルに流しこむ ↓ Loading buffer のフロントが 2/3 程度の位置で泳動を終了
14:35～	↓ ゲルを染色液にひたす
14:52～	↓ UV を照射し、バンドを確認する
	＊電気泳動と並行して実施
14:15～	【PCR 産物の精製】

14:37 ～	残った PCR 産物を新しいチューブに移す ↓　DNA クリーナーを 23μL 加え、ピペッティングにより混和 ↓　室温で 10 分間以上静置する ↓　室温、12,000rpm、10 分間 ↓　ピペッティングにより静かに上清を除く ↓　チューブのフタを開けて室温で 20 分間静置、風乾させる ↓　D.W. 15μL で溶解する
15:20～ 15:30～	【サイクルシークエンス（ダイターミネータ法)】 ・　シークエンス波形とあわせて、ダイターミネータ法の原理を説明 ＊　時間の都合上、サイクルシークエンス、その後のエタ沈作業をスキップ。 ＊　SLS との混合からスタートする 事前にエタ沈精製した PCR 産物（プロトコルの 10） ↓　30μL SLS を加え、ペレットを溶解する。 ↓　白色プレートの各ウェルにサンプルを全量（30μL）入れる ↓　余ったウェルには D.W. を入れる ↓　サンプルの上にミネラルオイルを 1 滴ずつのせる ↓　透明プレートの対応するウェルに、セパレーションバッファーを入れる。 ＊　時間の都合上、スタッフでシークエンスを行い、後日結果を配布する。
16:00～	<u>配列決定法について</u> ・　DNA シークエンスの仕組み サンガー法、キャピラリーシークエンサー、次世代シークエンサーなど ・　次世代シークエンサーとバイオインフォマティクス ・　ヒト 1 人の全ゲノムを解読するのに要する日数 サンガー法：8,000 年、キャピラリーシークエンサー：9 日 次世代シークエンサー：5 日

バイオインフォマティクス人材育成講座

遺伝子工学とその応用例

遺伝子の増幅と配列分析法

■　本日の実習内容　■

乳酸菌のシークエンス

１．　DNA の抽出と増幅

２．　増幅した産物の電気泳動

３．　増幅産物の精製

４．　サイクルシークエンス

５．　シークエンシング

６．　遺伝子情報の解析

2012 年 9 月 5 日

沖縄工業高等専門学校生物資源工学科

池松真也・磯村尚子・渡邊謙太・祖納元りえ・當山瑛子

伊波大志・大城龍之助・小松圭・瑞慶覧安里

遺伝子の解読（DNA シーケンサーを利用して）

乳酸菌のシークエンス

乳酸菌の培養

0.5% 炭酸カルシウム入り MRS Agar 培地による乳酸菌の培養（実施済）

＜ステップ１（操作１）＞

PrepMan Ultra Reagent を用いたコロニーからの DNA 抽出

1. PrepMan Ultra Reagent 200μl を 1.5ml チューブにとる。
2. MRS Agar 培地にできた乳酸菌のコロニーを滅菌つまようじでつつき，1 の溶液に混濁し，ボルテックスにて混合する。
3. ヒートブロックにて 100℃，10 分反応を行う。
4. 室温で 2 分静置する。
5. 遠心（13,000 rpm，3 分）を行う。
6. 上清適量を次の PCR 反応に用いる。

PCR（16S rRNA 領域（約 1500bp）の増幅）

1. 以下の通り試薬を混合する。操作は氷上で行うこと。

（反応系）

template（DNA 抽出後の上清）	1μl
4μM primer Forward	2μl
Reverse	2μl
2×buffer	10μl
2mM dNTP's	2μl
KOD polymerase	0.8μl
dH_2O（滅菌水）	2.2μl
total	20.0μl

2. スピンダウンし，以下の条件で PCR を行う。

（thermo-cycle）

95℃ - 3 min
95℃ - 30 sec
62℃ - 30 sec
68℃ - 1 min
} 32 cycles
68℃ - 3 min
4℃ - ∞

＜ステップ２（操作２）＞

PCR 産物の電気泳動、染色

1. 1.3% Agarose gel を作成する。
2. 100bp マーカー（4μl）およびサンプル（2μl）と 2×Loading buffer（2μl）を混合したものをそれぞれアプライし，100V，25 分泳動を行う。
   ※使用するマーカーには青色素が含まれているため，Loading buffer と混ぜる必要はない。
3. 泳動後のゲルを 1×Gel red にて 20 分染色を行う。
4. UV を照射し，DNA 増幅（バンド）の確認を行う。

＜ステップ３（操作３）＞

DNA クリーナーによる PCR 産物の精製　※未反応のプライマーやプライマーダイマー等を除く。

1. 核酸溶液を含む遠心チューブに，DNA クリーナーを核酸溶液と等量（18μl）添加し，ピペッティングで緩やかに混合する。
2. 室温で 10 分間以上静置する。
3. 室温，12,000rpm で 10 分間遠心分離する。
4. ピペッティングにより静かに上清を除く。
5. 70%エタノール溶液を核酸溶液（実験開始時）の 2 倍量（36μl）添加し，10 秒間ボルテックスミキサーで混合する。
6. 室温，12,000rpm で 10 分間遠心分離する。
7. ピペッティングにより静かに上清を除く。
8. チューブのフタを開けて室温で 20 分間静置し，ペレットを風乾する。
   ※5 分程度のバキュームドライでもよい。
9. 滅菌水（15μl）で溶解する。
10. Nano drop にて DNA 濃度測定を行う。

＜ステップ４（操作４）＞

サイクルシークエンス（ダイターミネータ法）

1. 以下の通りサイクルシークエンス反応に用いるテンプレート濃度計算を行う。
   ※テンプレートサイズにより必要なテンプレート量は異なる。今回増幅した 16S rRNA 領域は約 1500bp，必要なテンプレート量は 37ng である。

   例）DNA 濃度が 60ng/μl のとき
   60÷37≒1.6　　精製後の PCR 産物を滅菌水により 1.6 倍希釈する
   →PCR 産物 5μl と滅菌水 3μl を混合する（全量 8μl 程度になるように）

2. 以下の通り試薬を混合する。
6 つのプライマーを用いて，rRNA 領域約 1500bp を分割して読みきる。
（反応系）

希釈 template	1μl
master mix	2μl
5×bufer	2μl
primer	2μl
(A,B,C,D,E,F それぞれのパターン調製する)	
dH_2O	3μl
total	10.0μl / 1sample

＊サーマルサイクラーによるサイクルシーケンス

サイクルシークエンス後のエタ沈精製

1. チューブに以下の通り stop solution をサンプル分調整する。

stop solution

3M NaOAc（酢酸ナトリウム）	1μl
100mM EDTA	1μl
Glycogen	0.5μl
total	2.5μl / 1sample

2. サイクルシークエンス後のサンプルに 1 の stop solution を 2.5μl ずつ加え，よくピペッティングする。
3. 95%エタノール 30μl を加え，よくピペッティングし，全量を 1.5ml チューブに移す。
4. 13,000rpm，4℃，15min で遠心を行う。
5. 遠心後，上清を残さないよう注意深く吸い取る。　※ペレットを吸い込まないようにする。
6. 70%エタノール 100μl を加え，直ちに 13,000rpm，4℃，2min で遠心を行う。
7. 遠心後，上清を注意深く吸い取る。
8. 6-7 の操作（リンス）をもう 1 度繰り返す（計 2 回）。
9. 約 15 分間 Vacuum dry を行い，ペレットを乾かす。
10. SLS 30μl を加え，ペレットを溶解する。　※SLS に溶解した状態で，-20℃で 1 ヵ月保存可能

＜ステップ５（操作５）＞

シークエンシング

1. サンプルを準備する。
   1）白いプレートの各ウェルにサンプルを全量（30μl）入れる。
   2）1 レーンのうち，あまったウェルには dH_2O を入れる。
   3）サンプルの上にミネラルオイルを 1 滴ずつのせる。
   4）透明プレートのサンプルを入れたウェルと対応するウェルに，セパレーションバッファーを入れる。1 ウェル 10 滴くらい，ウェルの 9 割程度。
2. シークエンサーについては，使用する際に説明する。

＜ステップ６（操作６）＞

解析

1. Chromas Lite でシークエンス結果の波形を確認し，FASTA 形式で配列をコピーする。
2. Forward primer でサイクルシークエンス反応を行った場合はそのままでよいが，Reverse primer の場合，配列を相補鎖にして，さらに逆向きにする必要がある。
   SeqConv v081 を用いて，Complement 機能で変換が可能である。
3. DDBJ，BLAST の CLUSTALW にて Forward と変換後 Reverse の配列をコピーペーストし，送信する。
4. 読めた範囲をつなぎあわせ，その配列を BLAST にて 16S rRNA（Prokaryotes）をチェックし，相同性検索を行う。

（バイオインフォマティクス人材育成講座）
食品機能科学特別講義Ｉ（第３回）　池松真也・磯村尚子、他

2012/09/05(水)

２．増幅した産物の電気泳動

＜cDNA の電気泳動＞

＜試薬＞
・1%アガロースゲル
・1×TAE
・分子量マーカー（100bp DNA ladder）
・loading バッファー（gel loading Dye (6×)）
・染色液（SYBR Safe DNA gel stain(10000×)）

＜操作手順＞
≪前準備≫あらかじめ，それぞれのサンプルを流すレーンを決める.
①電気泳動装置に 1%アガロースゲルをセット.
②パラフィルムに loading バッファーを 1μl ずつ，使用するサンプル数と分子量マーカーの分をとる.
③分子量マーカーを 5μl とり, loading バッファーに加えピペッティングし, それをウェルに流し込む.
④それぞれのサンプルも同様に，ウェルに流し込む.
⑤泳動装置のカバーを閉め，電気泳動（100V，25min）を開始.
⑥loading バッファーがゲルの 2／3 付近にきたら，泳動を終了.
⑦ゲルを染色液に浸し，30 分間振盪.
⑧染色後 UV を照射し，バンドを確認し，写真撮影.
（SYBR Green は，光に当たるとすぐに分解が始まるため，すぐに UV を照射し，写真で記録する）

【　泳動写真　】

食品機能科学特別講義Ⅰ　第3回目-2

遺伝子の増幅と配列分析法
PCRによるDNA増幅

5 September 2012

Shinya Ikematsu
Okinawa National College of Technology
Director of Cooperative Technology Center

2011年8月21日（日）

3. PCR反応でcDNAを増やす

遺伝子の配列を読むために

バイオインフォマティクス人材育成講座　第2回

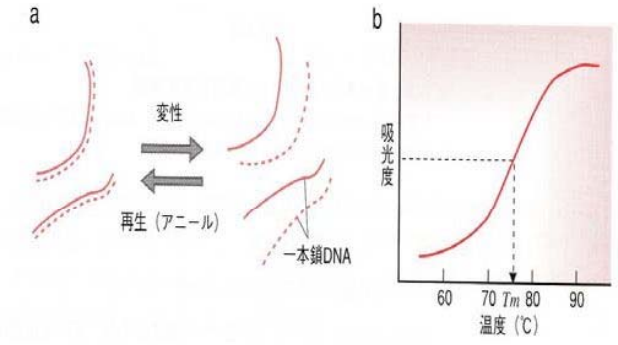


図2-6　DNAの変性
二本鎖DNAは熱などにより水素結合が壊れて，一本鎖DNAへと変化する（a）．このとき紫外部での吸光度が増加する．50%のDNAが変性する温度をTmという（b）

PCR（ポリメラーゼ連鎖反応）法

Polymerase Chain Reaction

はじめに

- DNAの塩基配列の決定や遺伝子組換えなどの遺伝子操作などを行うときには、目的とするDNA断片が多量に必要である。
- PCR（ポリメラーゼ連鎖反応）法を用いることで、目的とするDNA断片を短時間で、大量に増幅させることができる。

発明者

- マリス（Kary Babks Mullis、アメリカ）は1983年にPCR法を開発し、1993年に、この業績に対してノーベル化学賞が贈られた。
- PCR法は、その後の遺伝子工学の進歩に大きく貢献した。

PCR法に必要な酵素とDNA断片

（1） DNA合成酵素
（DNAポリメラーゼ）

（2） DNA断片
「プライマー」

（1）DNA合成酵素（DNAポリメラーゼ）

- 鋳型（いがた）となるDNAのヌクレオチド鎖を基にして、ヌクレオチドを重合する酵素。
- PCR法では高温（72℃）で反応する耐熱性のDNA合成酵素が用いられる。この酵素は、海底火山などの熱水噴出孔に生息する好熱菌 *Thermus aquaticus* から分離精製されたのでTaqポリラーゼとも呼ばれている。

（2） DNA断片 「プライマー」

- DNAの合成を開始するのに必要な短いDNA断片（20～30塩基程度）

PCR法の原理

（1） ステップ1
－ ＜変性＞
（2） ステップ2
－ ＜プライマーの結合＞
（3） ステップ3
－ ＜プライマーの伸長＞

ステップ1

＜変性＞

- 94℃で2本鎖のDNAは水素結合が切れ、1本鎖DNAに解離する。

ステップ2

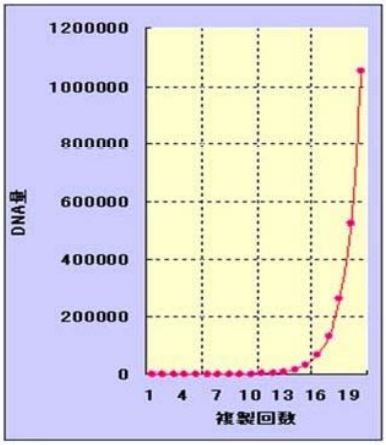
＜プライマーの結合＞

- 55℃に冷却するとプライマーが、1本鎖DNAの相補的な塩基配列の部分に結合する。

ステップ3

＜プライマーの伸長
（cDNAの伸長）＞

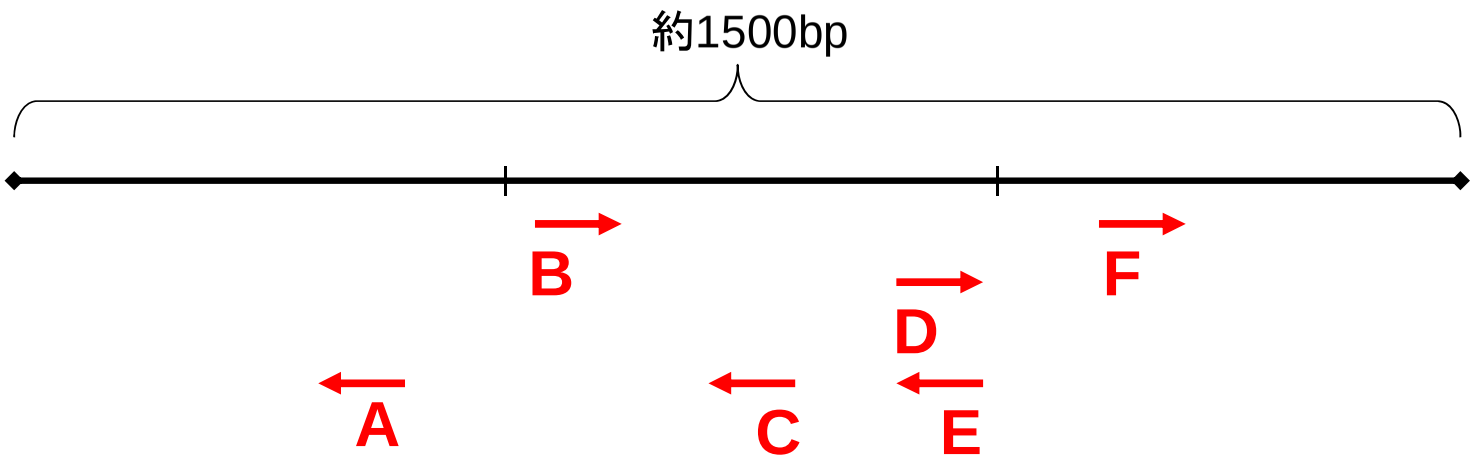
- 72℃にすると、
  DNAポリメラーゼが働いて
  プライマーを伸長させる。
- これで1サイクルが終了し、再び
  ステップ1に戻る。



20回のサイクル（数時間）で100万倍になる。（グラフ）

PCR法で使用する試薬・器具・機器

16S rRNA領域での各プライマー位置
（120905食品機能科学特別講義Ｉ）



2012年9月5日(水)　課題　25点満点

学部：＿＿＿＿＿　学科：＿＿＿＿＿　氏名：＿＿＿＿＿＿＿＿

＜問1＞　【2012食品機能科学特別講義Ⅰ　第3日目】(農学部)

池松教授は、いつもニコニコしているヒトから"ニコチャン"遺伝子をクローニングすることに成功しました。"ニコチャン"遺伝子は、A･･･B･･･C･･･D･･･E のような構造をしていることがわかりました。

(1) A,B,C,･･は塩基配列の並びからできています。DNAやmRNA分子のヌクレオチド3個からなる配列を何というか。　(3点)

(2) A,B,C,D,Eは真核生物の遺伝子の中で発現してタンパク質になる部分です。この部分を何と呼ぶか。　(3点)

(3) A･･･B･･･C･･･D･･･E の「･･･」の部分は非翻訳領域です。この部分を何と呼ぶか。　(3点)

(4) 通常、タンパク質が合成される際には(3)の非翻訳領域は取り除かれる。この作業を何というか　(3点)

(5) いつも不機嫌なヒトの"ニコチャン"遺伝子を調べたところ、いつもニコニコしているヒトとは異なっていた。どのような遺伝子構成になっていると推定できるだろうか。考えられるものを5つ以上書き出せ。正常な"ニコチャン"遺伝子の遺伝子構成は　ABCDE　である。全体的なサイズは5分の2以下にはならないようにして考えること。　(3点)

＜問2＞　【2012食品機能科学特別講義Ⅰ　第3日目　課題】(農学部)

ある遺伝子の塩基配列をショットガン法(大きなDNA鎖を無差別に切断して、たくさんの短いDNA鎖をつくり、それぞれの塩基配列を決定した後、全体の塩基配列を決める方法)で決定しようとしています。

①から⑦までの塩基配列が読み取れました。

メチオニンのコドンはATG、終止コドンは(TAA、TGA、TAG)です。

今回は、重なりが3塩基以上になるよう、問題を作成しました。

6つのアミノ酸をコードしています。①から⑦を正しく並べ直して下さい。

① GCAGTG　② GTTAAG　③ ATGGCA

④ GTGTTA　⑤ GGCAGG　⑥ AAGTGA　⑦ CAGGCA

(10点)

＜ヒント＞終止コドンはアミノ酸をコードしていません。

3.4　生命科学実習③：トランスクリプトーム解析

3.4.1 概要

講師：門田 幸二／東京大学大学院農学生命科学研究科特任准教授

日時：2012 年 9 月 6 日　10:20～／12:50～／14:40 ～

場所：琉球大学総合情報処理センター

内容：細胞における遺伝子発現の全体像を調べるトランスクリプトーム解析について、その実験技術であるマイクロアレイや次世代型シーケンサの活用について学び、サンプル間の発現変動遺伝子の抽出等、そのデータ解析の実例を実習する。

3.4.2 準備

Ｒインストール／Biostrings パッケージインストール

hoge フォルダの配布　▼hoge 内のファイル

名前	種類	サイズ
data_reads.txt	テキスト ドキュメント	1 KB
data_rma_2.txt	テキスト ドキュメント	9,005 KB
hoge4.fa	FA ファイル	1 KB
r.html	Chrome HTML Document	681 KB
r_seq.html	Chrome HTML Document	1,076 KB
sample15.txt	テキスト ドキュメント	1 KB
sample15_cl.txt	テキスト ドキュメント	1 KB
sample16.txt	テキスト ドキュメント	1 KB
sample16_cl.txt	テキスト ドキュメント	1 KB
sample19.txt	テキスト ドキュメント	1 KB

リンク集

1.「門田幸二」のホームページへのリンク

http://www.iu.a.u-tokyo.ac.jp/~kadota/

2.「(R で)塩基配列解析」ウェブページへのリンク

http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

3.「(R で)マイクロアレイデータ解析」ウェブページへのリンク

http://www.iu.a.u-tokyo.ac.jp/~kadota/r.html

4.「講義資料（公開版）」へのリンク

http://www.iu.a.u-tokyo.ac.jp/~kadota/20120906_kadota.pdf

5.「課題」へのリンク

http://www.iu.a.u-tokyo.ac.jp/~kadota/20120906_kadai.pdf

3.4.3 配布物

・講義スライド

3.4.4 課題

・プリント（課題 1～5）

3.4.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.4.6 時間割

時間	内容
10:20～	本日の講義概要紹介 参考ページ等の紹介 ・ (R で)塩基配列解析 ・ (R で)マイクロアレイデータ解析 各種ソフトの起動方法（word、excel、text などを含め）
10:27～	講師自己紹介 ・ 略歴 ・ 研究紹介（遺伝子発現解析）
10:36～	本日の講義ねらい ウェットな分野は限られたプロフェッショナルな人材しか生き残れなくなる。その中で、ドライの分野の人材が、実験そのものをデザインする時代になっていく →バイオインフォマティクス人材が不足しており、スキルのある人材は引く手あまた →次世代シーケンサによるトランスクリプトーム解析について学ぶ。それを通して基礎的なスキルを身につける。
10:38～	トランスクリプトームとは ・ オーム研究について ・ これまでの研究の進め方 vs これからの研究の進め方 ・ トランスクリプトーム取得方法 RNA-Seq→遺伝子発現行列→マッピング
11:17～	【実習】 ・ 目的配列に対して、NGS 由来塩基配列データ（今回は”AGG”）のマッピングを行い、任意のファイル名で出力 ・ ありがちなミスについて確認
11:50～	昼食
12:50～	R を用いた解析 ・ 「(R で)塩基配列解析」「(R で)マイクロアレイデータ解析」の基本を解説 ・ RNA-Seq の長所と短所

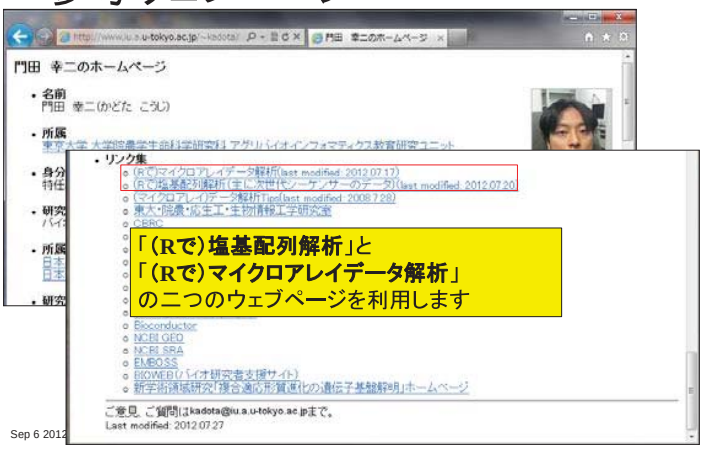
13:06～ 13:18～ 13:30～ 13:37～ 13:46～ 13:57～ 14:05～	・ マイクロアレイの長所と短所 ・ マイクロアレイの臨床応用例 ・ データの正規化 ＊ RPKM については省略 【実習】 ・ 配列長の補正を体験 <u>クラスタリング</u> ・ クラスタリングの実例紹介 ・ 距離の定義 マンハッタン距離、キャンベラ距離、最大距離など ・ 相関係数 【実習】 ・ 2 種類のサンプルについて、様々な距離を求める ・ 「？関数名」で詳細な使用方法を確認 <u>クラスターの併合</u> 最長距離法、群平均法、重心法など 【実習】 ・ 実データのクラスタリング ・ 論文データを読み込み、クラスタリングを行った後に系統樹を作成する ＊ 終わり次第休憩（14:20～）
	休憩
 14:40～ 14:47～ 15:12～	<u>二群間の比較</u> ・ パターンマッチング法 【実習】 ・ パターンマッチング法にて、読み込んだ二群間での発現遺伝子の同定を行う ・ Apply 関数についても解説 <u>R の詳細な使い方</u>

	・ R-Tips ・ cbind 関数（rbind 関数）について
15:23～	【実習】 ・ テンプレートパターンとの比較により、発現が大きく変化している遺伝子を探す
15:28～	<u>様々な組織（条件）での発現遺伝子解析</u> ・ ランキングに基づく方法、外れ値検出、その他 ・ エントロピー解析による検出
15:45～	＊ 全体についてまとめた後、講義終了時刻まで課題に取り組んだ

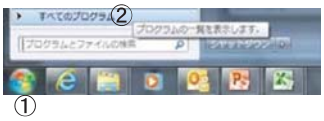
バイオインフォマティクス・スタンダードコース
（食品機能科学特別講義I 第4部 生命科学実習③ ）
「トランスクリプトーム解析」

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
門田 幸二（かどた こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

参考ウェブページ



各種ソフトの場所



スタート － すべてのプログラム － …

自己紹介

- 1995年3月
  - 高知工業高等専門学校・工業化学科　卒業
- 1997年3月
  - 東京農工大学・工学部・物質生物工学科　卒業
- 1999年3月
  - 東京農工大学・大学院工学研究科・物質生物工学専攻　修士課程修了
- 2002年3月
  - 東京大学・大学院農学生命科学研究科・応用生命工学専攻　博士課程修了
  - 学位論文：「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」
    （指導教官：清水謙多郎教授）
- 2002/4/1~
  - 産総研・生命情報科学研究センター　産総研特別研究員
- 2003/11/1~
  - 放医研・先端遺伝子発現研究センター　研究員
- 2005/2/16~
  - 東京大学・大学院農学生命科学研究科
    特任助手→特任助教→特任准教授

バイオインフォマティクス人材育成…

- 現状
  - NGSデータなどの大量実験データを自在に解析できるバイオインフォマティクス人材が不足
  - スキルのある人は引く手あまた
- 私の状況
  - 東大大学院講義（90分×13回）
  - その他、セミナーや講習会の講師
  - 自分の研究（と共同研究の解析）を進める
  - メールでの質問対応（これも頻繁にくるので大変）
  - 初心者でもコピペでデータ解析可能なウェブページの更新
    - （Rで）マイクロアレイデータ解析
    - （Rで）塩基配列解析

ねらい

- マイクロアレイおよび次世代型シーケンサ（NGS）によるトランスクリプトーム解析について学ぶ
  - Rを利用することで、様々なトランスクリプトーム解析が可能
  - プログラミング能力がなくても使いこなし術があれば…
- バイオインフォマティクスの基本的なスキルを身につけることが重要
  - バイオインフォマティクス技術者認定試験合格を目指せ（11/25）
  - 相関係数やエントロピーなどの要素技術を駆使すれば様々なデータ解析が可能であることを紹介

オーム（Ome）研究

ome ：総体

DNA　遺伝子 (Gene) + ome　→ ゲノム（命の設計図）

転写（遺伝子DNAの情報をRNAに写すこと）

RNA　転写 (Transcription) + ome　→ トランスクリプトーム

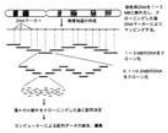
翻訳（転写されたRNA情報をもとにタンパク質を作ること）

タンパク質　タンパク質 (Protein) + ome　→ プロテオーム

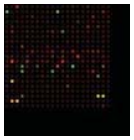
これまでの実験技術

DNA → 転写 → RNA → 翻訳 → タンパク質

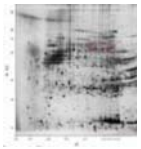
全ゲノムショットガン法

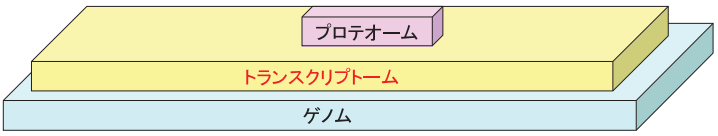


マイクロアレイ



二次元電気泳動法





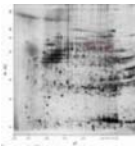
今後の実験技術

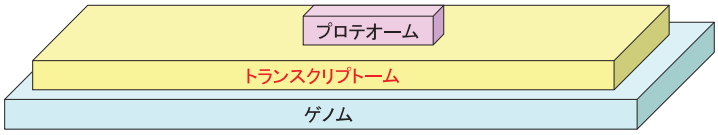
DNA → 転写 → RNA → 翻訳 → タンパク質

次世代シーケンサー（NGS）

GS FLX+　5500xl SOLiD　HiSeq 2000

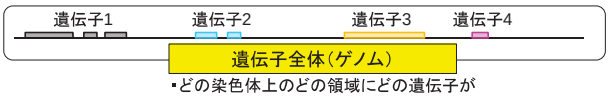
二次元電気泳動法





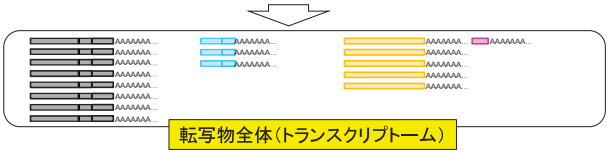
トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域



遺伝子全体（ゲノム）

- どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



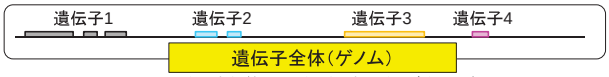
転写物全体（トランスクリプトーム）

- 遺伝子1は沢山転写されている（発現している）
- 遺伝子4はごくわずかしか転写されてない
- …

トランスクリプトームとは

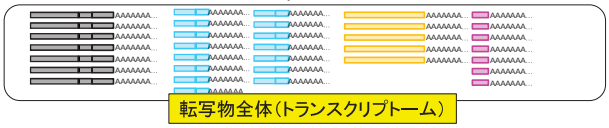
- ある状態のあるサンプル（例：目）のあるゲノムの領域

光刺激



遺伝子全体（ゲノム）

- どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



転写物全体（トランスクリプトーム）

- 遺伝子2は光刺激に応答して発現亢進
- 遺伝子4も光刺激に応答して発現亢進

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム

これがいわゆる「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

- 光刺激後（T2）の目のトランスクリプトーム

- マイクロアレイ
- RNA-Seq（NGS）
- …

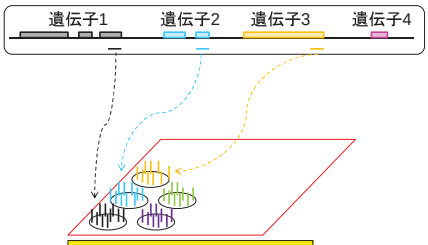
トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全RNA（転写物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - □マイクロアレイ
    - cDNAマイクロアレイ、Affymetrix GeneChip、タイリングアレイなど
  - □配列決定に基づく方法
    - EST、SAGEなど、次世代シーケンサー (NGS)
  - □電気泳動に基づく方法
    - Differential Display、AFLPなど

調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

トランスクリプトーム取得（マイクロアレイ）

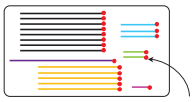
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した"チップ"

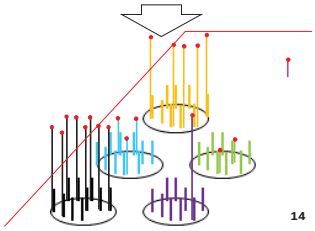
- メーカーによって搭載されている遺伝子の種類が異なる
  →搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



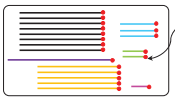
蛍光標識

ハイブリダイゼーション（二本鎖形成）



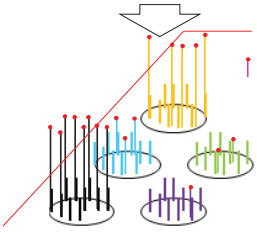
マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム



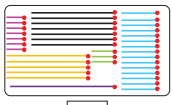
蛍光標識

ハイブリダイゼーション（二本鎖形成）



専用の検出器で各遺伝子に対応する領域の蛍光シグナル強度を測定

光刺激後（T2）の目のトランスクリプトーム



ハイブリダイゼーション と シグナル検出

	T1
遺伝子1	8
遺伝子2	3
遺伝子3	5
遺伝子4	?
遺伝子5	…
…	…

T2
7
15
5
?
…
…

正規化

ハイブリダイゼーション

- 核酸（DNA or RNA）分子が相補的に複合体を形成すること
  - □核酸分子に含まれる塩基はAとT（or U）またはGとCというふうに相補的に結合する性質があるので、この性質を利用

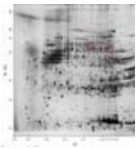
トランスクリプトーム取得（RNA-Seq）

DNA → 転写 → RNA → 翻訳 → タンパク質

次世代シーケンサー（NGS）

GS FLX+　5500xl SOLiD　HiSeq 2000

二次元電気泳動法



ゲノムではなく転写されているRNAの配列決定（Sequencing）をするので、RNA-Seqと呼ばれる

トランスクリプトーム

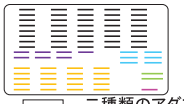
ゲノム

RNA-Seqデータ → 遺伝子発現行列

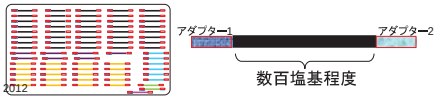
- 次世代シーケンサー（Illumina社の場合）

光刺激前（T1）の目のトランスクリプトーム

数百塩基程度に断片化

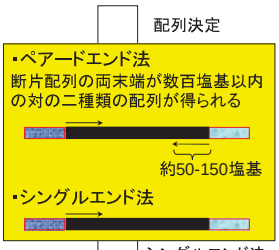


二種類のアダプター配列を両末端に付加

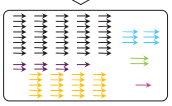


配列決定

- ペアードエンド法
  断片配列の両末端が数百塩基以内の対の二種類の配列が得られる
- シングルエンド法



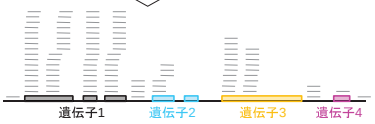
シングルエンド法の場合



RNA-Seqデータ → 遺伝子発現行列

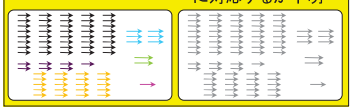
光刺激前（T1）の目のトランスクリプトーム

ゲノム配列にマッピング



ーイメージー
50-125塩基程度からなる配列が沢山ある

ー実際ー
数百万個の配列があり、どの遺伝子に対応するか不明



（短い）配列を読んだものという意味で（ショート）リードなどと呼ばれる

定量化(例：生のリード数をカウント)

	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

正規化

T1
8
3
5
1
…
…

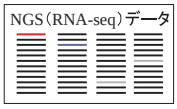
マッピング？

- NGSデータ中の数千万リード（一が数千万個あるということ）の各々がゲノム中のどこにマップされるか、マップされないのはどれか、などを調べるイメージ

「ヒトゲノム配列」

1番染色体　2番染色体　3番染色体　…

NGS（RNA-seq）データ



Linux上で動くNGSデータ専用のマッピングプログラムを用いて実行できます

トランスクリプトーム配列にマップ

- マップされる側の配列はゲノム配列に限らない

遺伝子1　遺伝子2　遺伝子3　遺伝子4

ゲノム配列

>1番染色体
>2番染色体
…

トランスクリプトーム配列

>遺伝子1
>遺伝子2
>遺伝子3
…

マッピング　＝　（大量高速）文字列検索

- マップされる側の配列：4コンティグ（or 4遺伝子 or 4染色体）
- マップする側のNGS由来塩基配列データ："AGG"

hoge4.fa - メモ帳

```
>contig_1
CGGACAGCTCCTCGGCATCCGGAT
>contig_2
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCCTAACATCGCAAGTCG
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCCT
GTC
>contig_3
TGTAGGAGAAGGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT
GTATGAGGTCGGGCA
>contig_4
CGTGCTGATTCCACACAGCAGTAAACGCGGACCTCTACCTATGAACATG
```

出力ファイル：hoge2.txt

	start	end
contig_2	31	33
contig_2	77	79
contig_3	4	6
contig_3	10	12
contig_3	56	58

Rでやってみよう

実習（Rでやってみよう）

- 目的：hoge4.faファイルに対してNGS由来塩基配列データ（例："AGG"）のマッピング（or 文字列検索）を行い、一致領域情報を任意のファイル名（例："hoge2.txt"）で出力

hoge4.fa - メモ帳

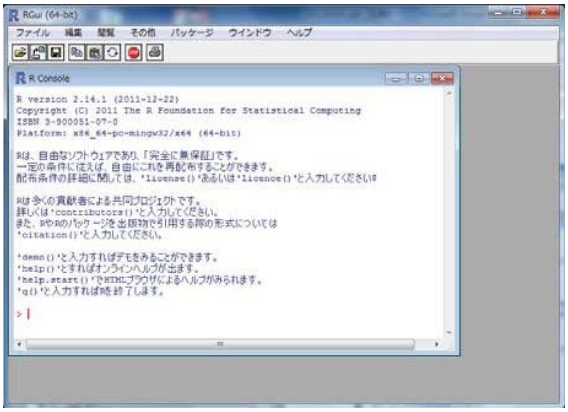
```
>contig_1
CGGACAGCTCCTCGGCATCCGGAT
>contig_2
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCCTAACATCGCAAGTCG
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCCT
GTC
>contig_3
TGTAGGAGAAGGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT
GTATGAGGTCGGGCA
>contig_4
CGTGCTGATTCCACACAGCAGTAAACGCGGACCTCTACCTATGAACATG
```

出力ファイル：hoge2.txt

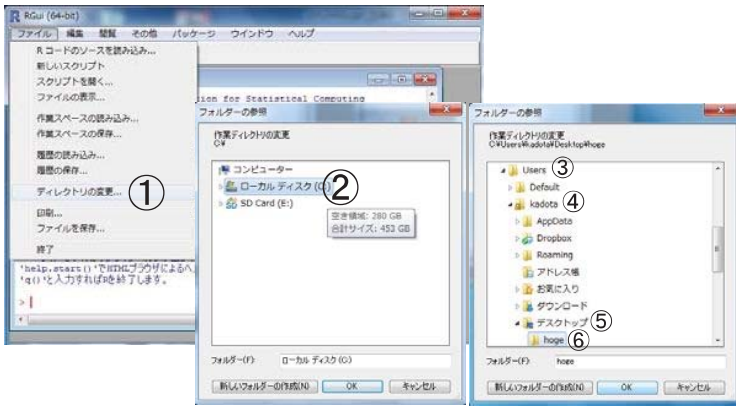
	start	end
contig_2	31	33
contig_2	77	79
contig_3	4	6
contig_3	10	12
contig_3	56	58

デスクトップ上に「hoge」という名前のフォルダがあり、フォルダ中に入力ファイル（**hoge4.fa**）が存在する、という前提

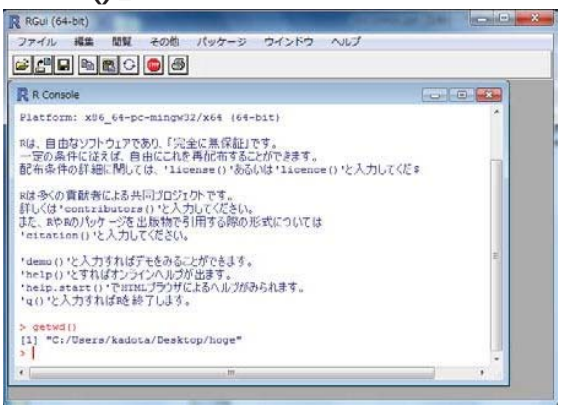
Rの起動



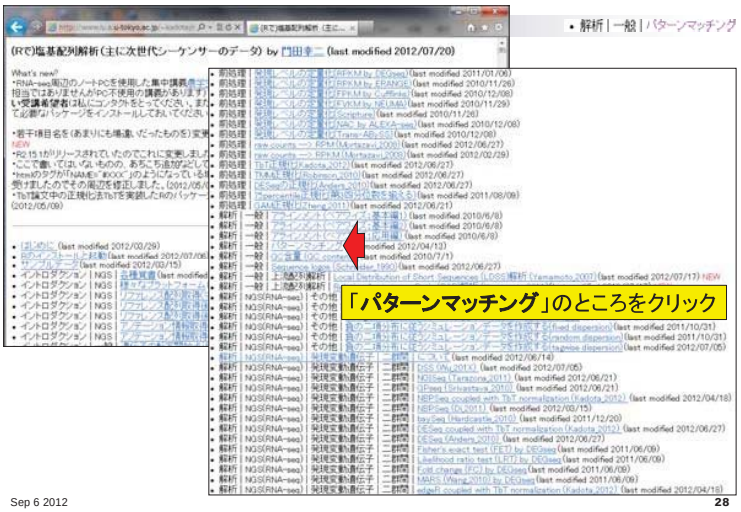
作業ディレクトリの変更



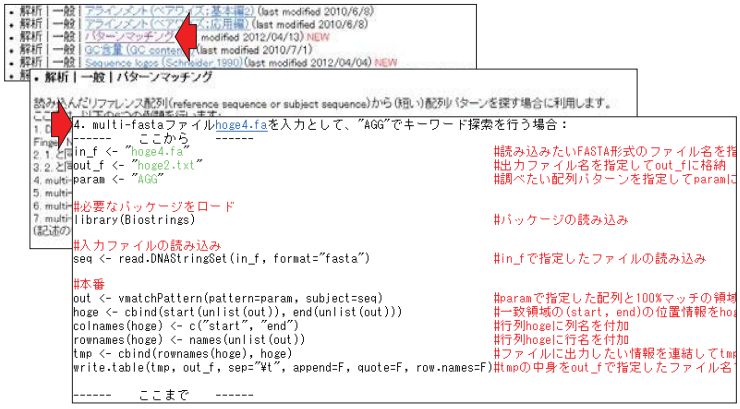
「getwd()」と打ち込んで確認



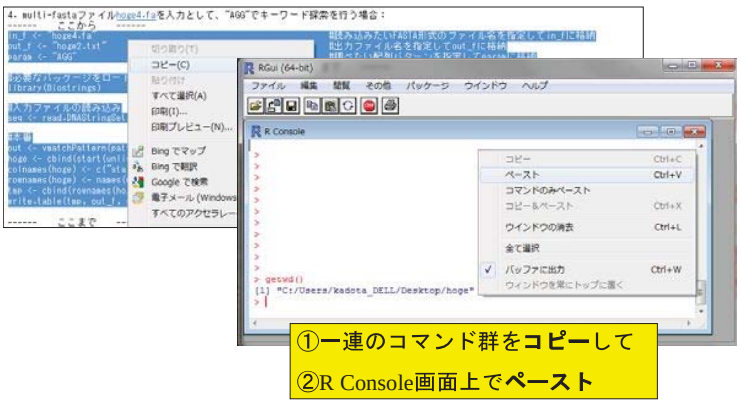
参考ウェブページ



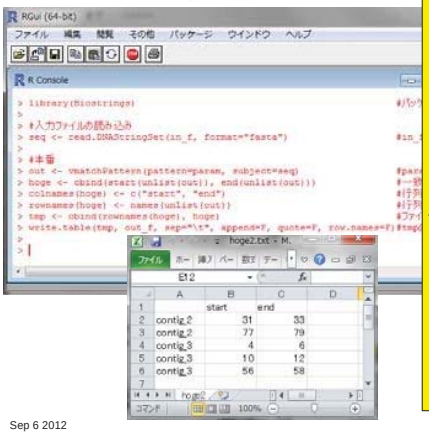
パターンマッチング



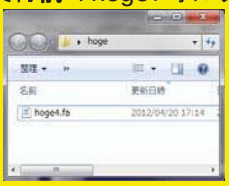
基本はコピペ



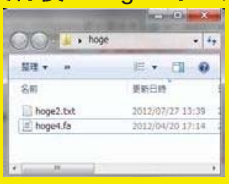
実行結果



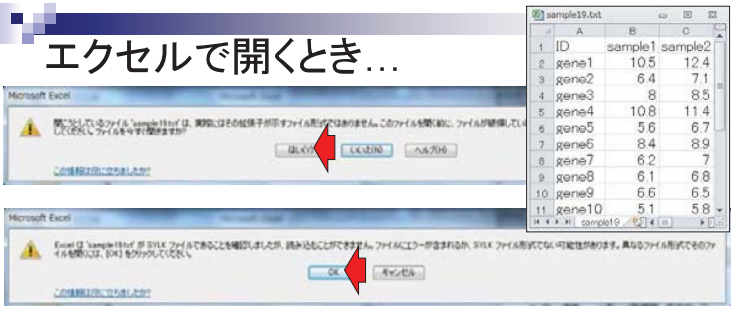
実行**前**のhogeフォルダ



実行**後**のhogeフォルダ

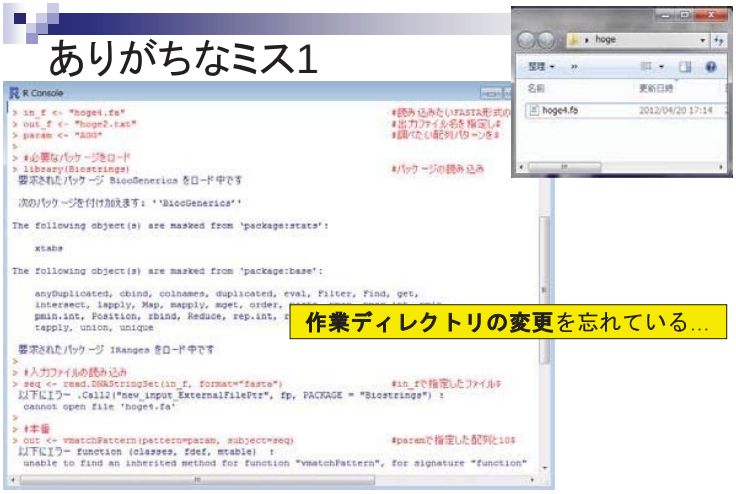


エクセルで開くとき…



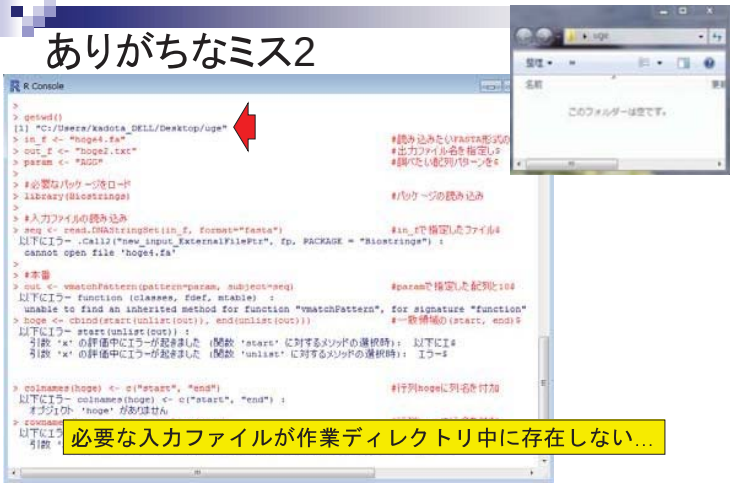
（ドラッグ＆ドロップで開こうとすると）エラーが出て一回目は開けないことがあるが、その後もう一度同じ作業を繰り返すと開けます…

ありがちなミス1



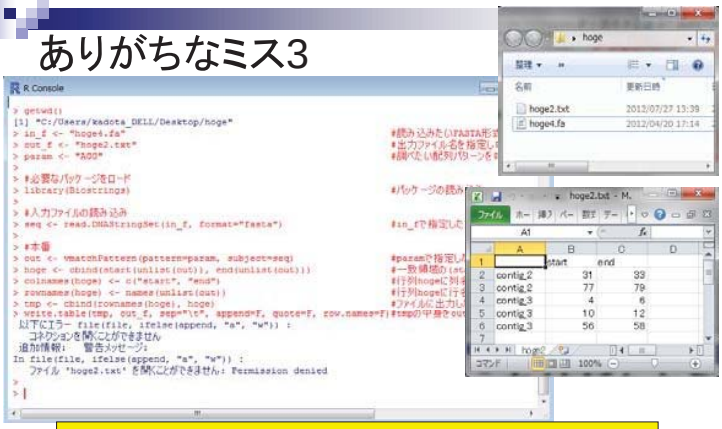
作業ディレクトリの変更を忘れている…

ありがちなミス2



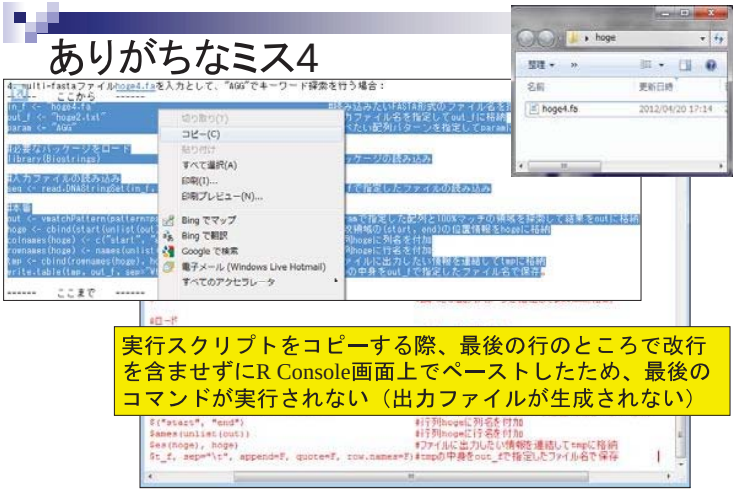
必要な入力ファイルが作業ディレクトリ中に存在しない…

ありがちなミス3

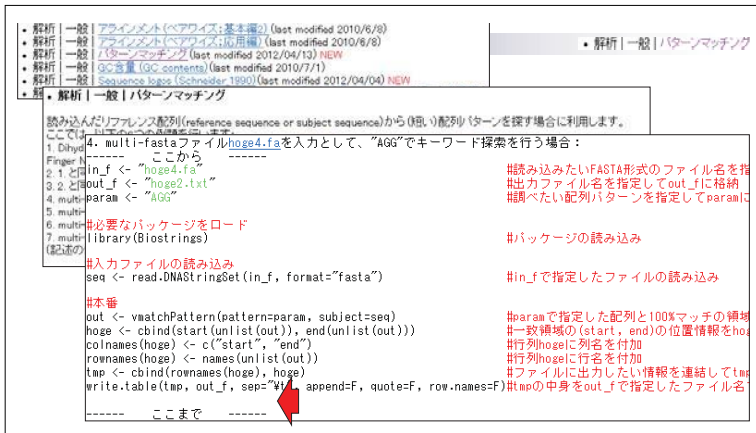


出力予定のファイル名と同じものを別のプログラムで開いているため最後のwrite.table関数のところでエラーが出る

ありがちなミス4

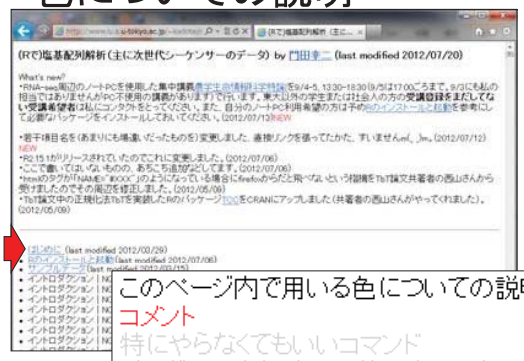


実行スクリプトをコピーする際、最後の行のところで改行を含ませずにR Console画面上でペーストしたため、最後のコマンドが実行されない（出力ファイルが生成されない）

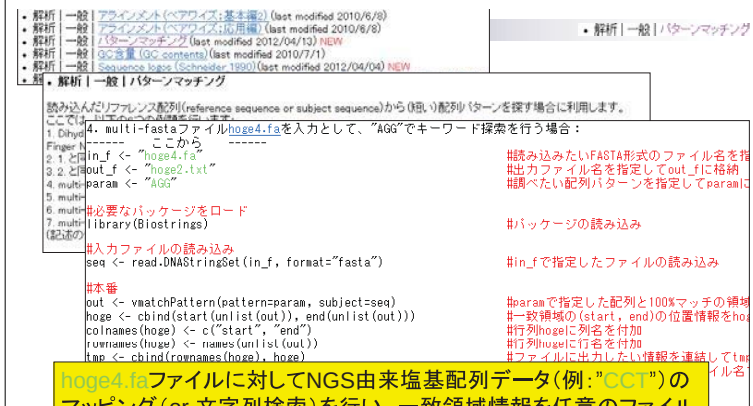


「----- ここまで -----」の一つ上の空行には「スクリプト最終行のコマンドを確実に実行するため」という深い意味があります

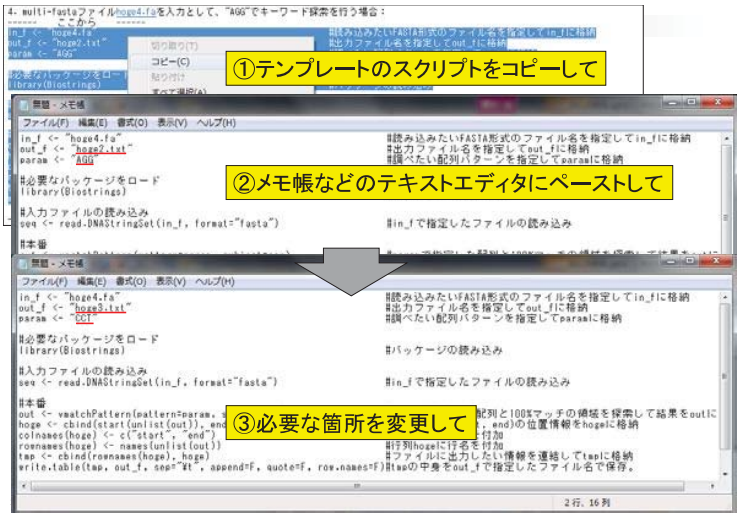
色についての説明

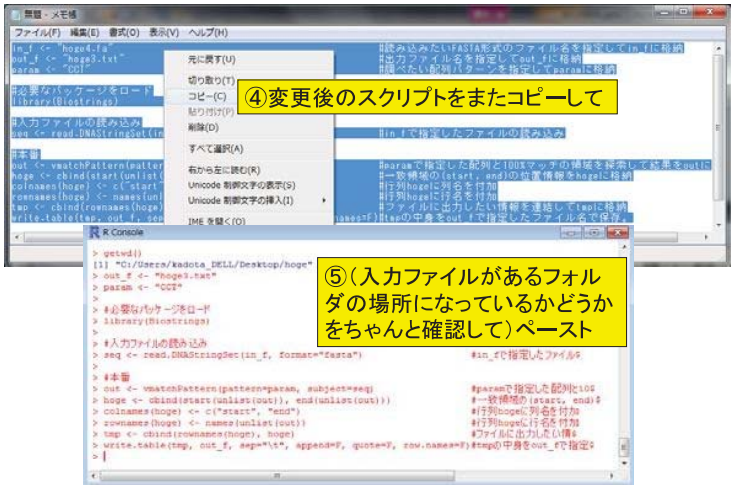


このページ内で用いる色についての説明:
コメント
特にやらなくてもいいコマンド
プログラム実行時に目的に応じて変更すべき箇所

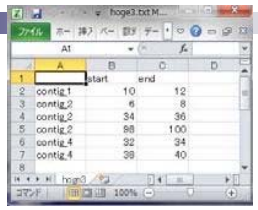


hoge4.faファイルに対してNGS由来塩基配列データ(例:"CCT")のマッピング(or 文字列検索)を行い、一致領域情報を任意のファイル名(例:"hoge3.txt")で出力したいときは？

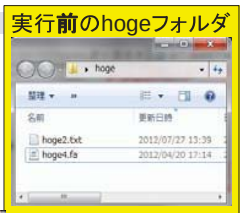


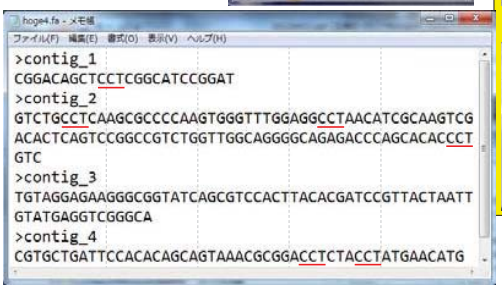


実行結果

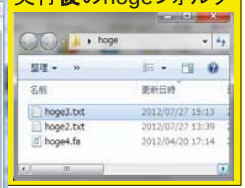


実行前のhogeフォルダ





実行後のhogeフォルダ



より現実に近い解析

data_reads.txt
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA

複数個のリードからなるファイルを読み込んで一度にマッピング結果を返すことも可能です

```
------  ここから  ------
in_f1 <- "hoge4.fa"
in_f2 <- "data_reads.txt"
out_f <- "hoge4.txt"

#必要なパッケージをロード
library(Biostrings)

#入力ファイルの読み込み
seq <- read.DNAStringSet(in_f1, format="fasta")          #in_f1で指定したファイルの読み込み
reads <- read.DNAStringSet(in_f2, format="fasta")        #in_f2で指定したファイルの読み込み

#本番
out <- c("in_f2", "in_f1", "start", "end")               #最終的に得る出力ファイルのヘッダー情報
for(i in 1:length(reads)){                               #リード数分だけループを回す
    tmp <- vmatchPattern(pattern=as.character(reads[i]), subject=seq)#オブジェクトreads中の各塩基配列と
    hoge1 <- cbind(start(unlist(tmp)), end(unlist(tmp)))  #一致領域の(start, end)の位置情報をhoge
    hoge2 <- names(unlist(tmp))                          #ヒットしたリファレンス配列中のIDをhoge
    hoge3 <- rep(as.character(reads[i]), length(hoge2))  #hoge2の要素数分だけ、マッチする側の配列
    out <- rbind(out, cbind(hoge3, hoge2, hoge1))        #cbind(hoge3, hoge2, hoge1)で表される欲
}
write.table(out, out_f, sep="¥t", append=F, quote=F, row.names=F, col.names=F)#resultの中身をout_fで指定
------  ここまで  ------
```


data_reads.txt
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA

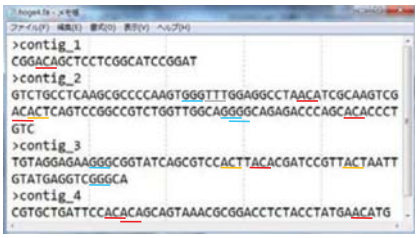
出力ファイル:hoge4.txt

in_f2	in_f1	start	end
TTT	contig_2	26	28
GGG	contig_2	23	25
GGG	contig_2	78	80
GGG	contig_2	79	81
GGG	contig_3	11	13
GGG	contig_3	61	63
ACT	contig_2	53	55
ACT	contig_3	28	30
ACT	contig_3	44	46
ACA	contig_1	4	6
ACA	contig_2	38	40
ACA	contig_2	51	53
ACA	contig_2	94	96
ACA	contig_3	32	34
ACA	contig_4	13	15
ACA	contig_4	15	17
ACA	contig_4	45	47

hoge4.fa - メモ帳

```
>contig_1
CGGACAGCTCCTCGGCATCCGGAT
>contig_2
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCCTAACATCGCAAGTCG
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCCT
GTC
>contig_3
TGTAGGAGAAGGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT
GTATGAGGTCGGGCA
>contig_4
CGTGCTGATTCCACACAGCAGTAAACGCGGACCTCTACCTATGAACATG
```

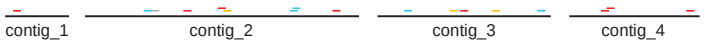

出力結果ファイルと発現量の関係



data_reads.txt
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA

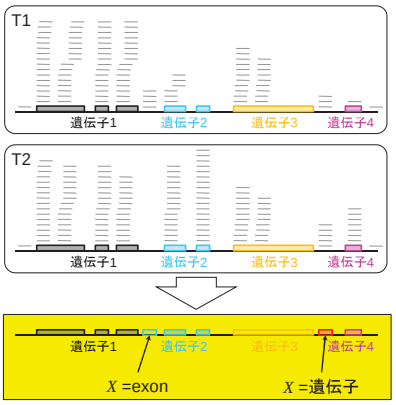
出力ファイル:hoge4.txt

in_f2	in_f1	start	end
TTT	contig_2	26	28
GGG	contig_2	23	25
GGG	contig_2	78	80
GGG	contig_2	79	81
GGG	contig_3	11	13
GGG	contig_3	61	63
ACT	contig_2	53	55
ACT	contig_3	28	30
ACT	contig_3	44	46
ACA	contig_1	4	6
ACA	contig_2	38	40
ACA	contig_2	51	53
ACA	contig_2	94	96
ACA	contig_3	32	34
ACA	contig_4	13	15
ACA	contig_4	15	17
ACA	contig_4	45	47



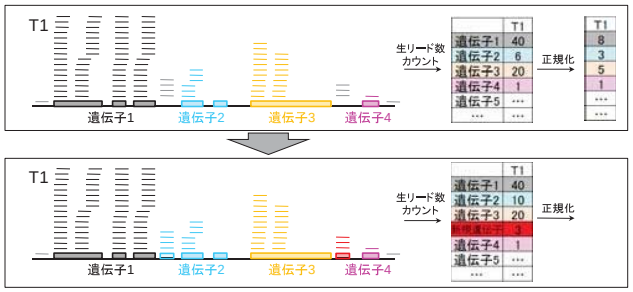
RNA-Seqの長所

- 新規Xの同定
  - X =exon, 遺伝子, …



RNA-Seqの長所

- 新規Xの同定



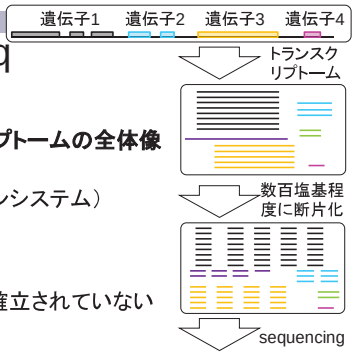
- "トランスクリプトーム(転写物の全体像)"の理解への一番の近道
- よりよい**遺伝子発現行列**を得るための基礎情報充実に貢献

長所・短所:マイクロアレイ

- 長所
  - すでに診断用マイクロアレイが市販されているなど**長年の実績**
  - お手軽、各種データ解析ツールが豊富
- 短所
  - (プローブ搭載のために)解析対象の塩基配列情報を予め知っておく必要がある。(クローズドシステム)
  - プローブが搭載されていない遺伝子の発現レベルは測定不可能(未知遺伝子も当然対象外)

- 主なユーザー
  - 主な解析対象が(アノテーション情報が豊富な)モデル生物で、既知遺伝子のみでいい、という研究者

長所・短所：RNA-Seq



- 長所
  - （未知遺伝子を含む）**トランスクリプトームの全体像を理解することが原理的に可能**
  - 事前情報を必要としない（オープンシステム）
  - ダイナミックレンジが広い
- 短所
  - データ解析が大変、解析手法が確立されていない
- 主なユーザー
  - 無制限（モデル生物・非モデル生物を問わない）
  - （お金持ち…）

マイクロアレイの臨床応用例

- 「MammaPrint」：乳癌予後予測検査サービス
  - 2008年3月
  - 乳癌手術を受けた患者の転移・再発の可能性に関する情報提供
  - 70遺伝子の活性を測定
  - 不必要な補助化学療法などを避けることが可能（ローリスク群）
- 「oncotype DX」：早期浸潤性乳癌の術後再発予測サービス
  - 2007年2月
  - 再発リスクの数値化および化学療法の効果予測
  - 21遺伝子を解析
  - 必要以上の化学療法を回避

既に臨床診断に利用されている

- 「GeneSearch」：乳癌の術中リンパ節転移迅速診断
  - 2007年7月

ここまでのまとめ

- トランスクリプトーム解析技術を紹介
  - マイクロアレイ
  - RNA sequencing (RNA-Seq)

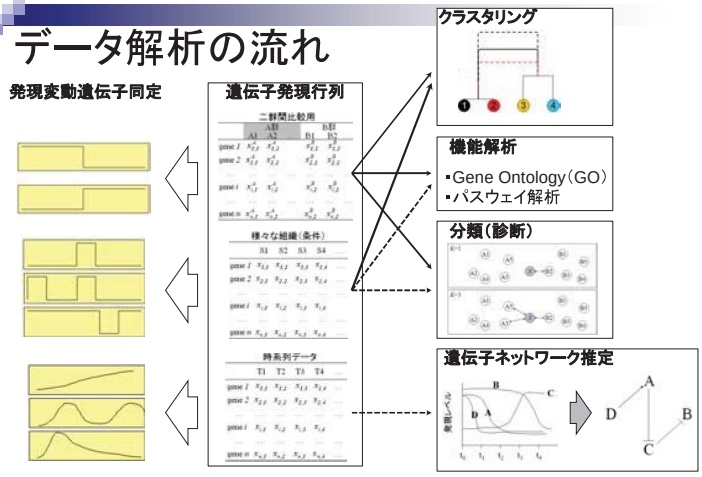
遺伝子発現行列

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

- RNA-Seqデータのマッピング実習
  - 「R」と「**参考ウェブページ**」の基本的な使い方を紹介
- マイクロアレイとRNA-Seqの長所・短所

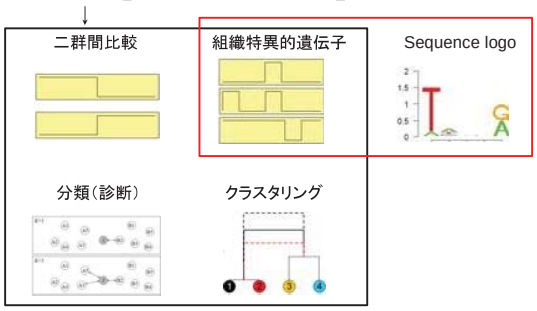
どの実験技術由来データも「遺伝子発現行列」の形式に変換可能

データ解析の流れ



バイオインフォマティクス要素技術

- 「相関係数」や「エントロピー」などの応用例を紹介



マイクロアレイデータの正規化

- 「各サンプルから測定されたシグナル強度の和は一定」と仮定
  - チップ上の遺伝子数が少ない場合は非現実的だが、数千～数万種類の遺伝子が搭載されているので妥当（だろう）

	sample1	sample2
gene1	10.5	12.4
gene2	6.4	7.1
gene3	8.0	8.5
gene4	10.8	11.4
gene5	5.6	6.7
gene6	8.4	8.9
gene7	6.2	7.0
gene8	6.1	6.8
gene9	6.6	6.5
gene10	5.1	5.8
総和	73.7	81.1

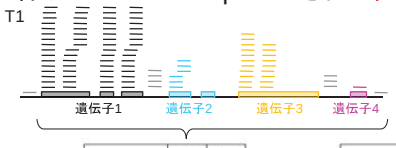
グローバル正規化 →

	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
総和	100.0	100.0

背景：サンプル（or chip）ごとにシグナル強度の総和は異なる
対策：総和が任意の値（例では100）になるような正規化係数を掛ける
例：sample1の正規化係数= 100 / 73.7

RNA-Seqデータの正規化（の一部）

- 「各サンプルからsequenceされた総リード数は一定」と仮定



	T1	T2
遺伝子1	40	7
遺伝子2	6	15
遺伝子3	20	5
遺伝子4	1	1
総リード数	67	28

RPM正規化

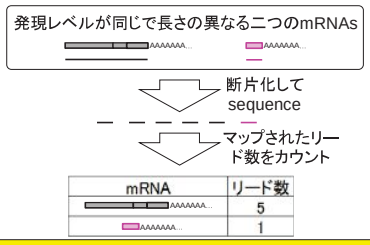
	T1	T2
遺伝子1	597014.9	250000.0
遺伝子2	89552.2	535714.3
遺伝子3	298507.5	178571.4
遺伝子4	14925.4	35714.3
総リード数	1000000	1000000

Reads Per Million mapped reads（RPM）
正規化後の総リード数が100万（one million）になるように補正
例：T1の正規化係数 = 1000000 / 67

Mortazavi *et al.*, *Nat. Methods*, **5**: 621-628, 2008

配列長の補正

- 配列長が長い遺伝子ほど沢山sequenceされる
  - それらの遺伝子上にマップされる生のリード数が増加傾向
  - 配列長が長い遺伝子ほど発現レベルが高い傾向になる



一つのサンプル内での異なる遺伝子間の発現レベルの高低を（配列長を考慮せずに）比較することはできない

Mortazavi *et al.*, *Nat. Methods*, **5**: 621-628, 2008

配列長の補正

mRNA	リード数	配列長（in bp）
AAAAAAA...	5	1500
AAAAAAA...	1	300

- 前提条件：配列長が既知
- 補正の基本戦略：配列長で割る
  - 「1 / 配列長」を掛ける場合
    → 「塩基あたりの平均のリード数」を計算しているのと等価
  - 「1000 / 配列長」を掛ける場合
    → 「その遺伝子の配列長が1000bpだったときのリード数」と等価
    Reads Per Kilobase of exon

Mortazavi *et al.*, *Nat. Methods*, **5**: 621-628, 2008

RPKM

- RPM正規化（マイクロアレイなどと同じところ）
  - Reads per million mapped reads
  - サンプルごとにマップされた総リード（塩基配列）数が異なる。
    →各遺伝子のマップされたリード数を「総read数が100万（one million）だった場合」に補正

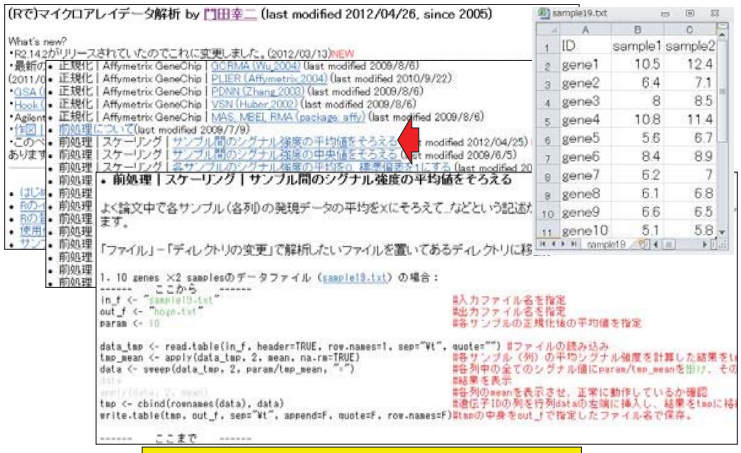
「raw counts : all reads= RPM : 1,000,000 」
A1BGの場合は「744 : 5,087,097 = RPM : 1,000,000」

$$\text{RPM} = \text{raw counts} \times \frac{1{,}000{,}000}{\text{all reads}} = 744 \times \frac{1{,}000{,}000}{5{,}087{,}097} = 146.3$$

geneName	raw counts	RPKM	all reads	gene length	RPM
A1BG	744	82.9	5087097	1764	146.3
A1CF	159	13.7	5087097	2278	31.3
A2BP1	1	0.0	5087097	5415	0.2
A2LD1	4	0.6	5087097	1226	0.8

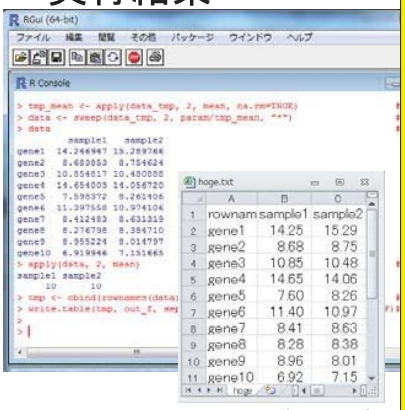
- RPKM正規化（RNA-Seq特有）
  - Reads per kilobase of exon per million mapped reads
  - 遺伝子の配列長が長いほど配列決定(sequence)される確率が上昇
    →各遺伝子の配列長を「1000塩基（one kilobase）の長さだった場合」に補正

$$\text{RPKM} = \underbrace{\text{raw counts} \times \frac{1{,}000{,}000}{\text{all reads}}}_{\text{RPM}} \times \frac{1{,}000}{\text{gene length}} = \text{raw counts} \times \frac{1{,}000{,}000{,}000}{\text{gene length} \times \text{all reads}}$$



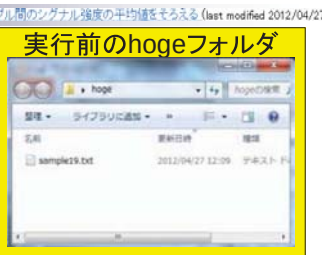
「（Rで）マイクロアレイデータ解析」のほうです

実行結果

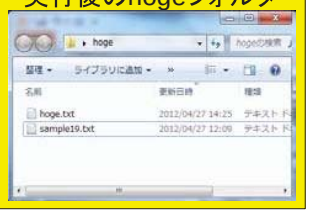


スライド54と対応

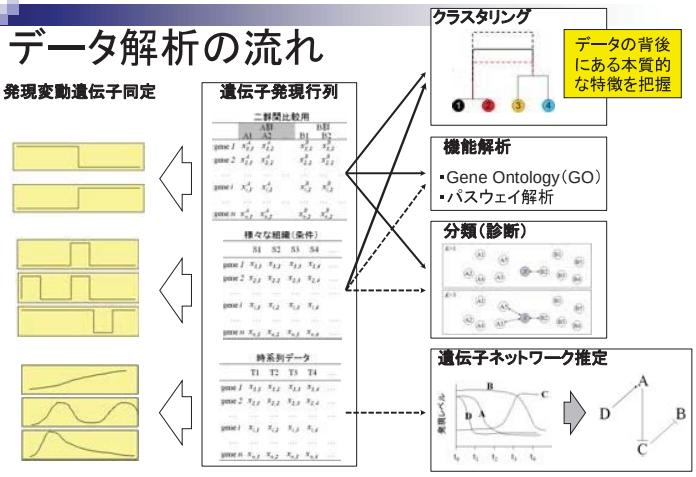
実行前のhogeフォルダ



実行後のhogeフォルダ



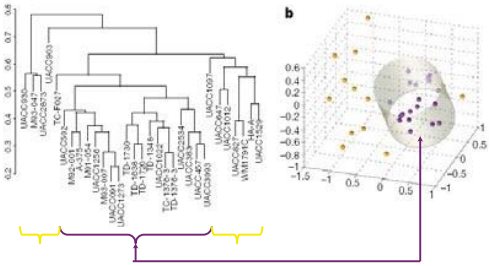
データ解析の流れ



Bittner *et al.*, *Nature*, **406**, 536-540, 2000

（サンプル間）クラスタリングの実例

□悪性黒色腫（メラノーマ）31サンプルのデータ



悪性度の高い癌の**サブタイプ**を発見

クラスタリング（教師なし学習）

■ 決めておくべき二つの基準（事柄）

□距離（類似度）の定義

■ ユークリッド距離、マンハッタン距離など

□クラスターをまとめる（併合する）方法

■ クラスター間の距離を定義する方法、とほぼ同じ

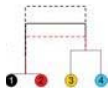
■ 最短距離法、平均連結法、ウォード法など

入力例

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

クラスタリング

出力例



距離（類似度）の定義

■ 遺伝子（or サンプル）***x***と***y***の発現パターンの距離$D(x,y)$

$$\text{相関係数}\ r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (-1 \le r_{xy} \le 1)$$

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

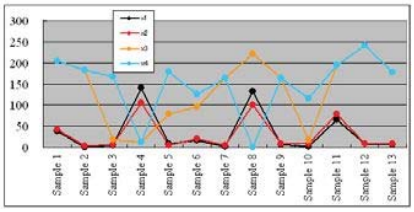
$$\begin{cases} \boldsymbol{x}\text{と}\boldsymbol{y}\text{の発現パターンが酷似} \ \rightarrow\ r \approx 1 \\ \boldsymbol{x}\text{と}\boldsymbol{y}\text{の発現パターンがばらばら} \ \rightarrow\ r \approx 0 \\ \boldsymbol{x}\text{と}\boldsymbol{y}\text{の発現パターンがほぼ正反対} \ \rightarrow\ r \approx -1 \end{cases}$$

$$\text{距離}\ D(x,y) = 1 - r \quad (0 \le D \le 2) \quad \begin{cases} r = 1 \ \rightarrow\ D = 1-1 = 0 \\ r = 0 \ \rightarrow\ D = 1-0 = 1 \\ r = -1 \ \rightarrow\ D = 1-(-1) = 2 \end{cases}$$

「1－相関係数」を**距離**として定義することができます

相関係数 → 距離 （計算例）

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177



相関係数 $r_{x^1x^2} = 0.98 \ \rightarrow$ 距離$D_{x^1x^2} = 1 - 0.98 = 0.02$

相関係数 $r_{x^1x^3} = -0.01 \ \rightarrow$ 距離$D_{x^1x^3} = 1 - (-0.01) = 1.01$

相関係数 $r_{x^1x^4} = -0.78 \ \rightarrow$ 距離$D_{x^1x^4} = 1 - (-0.78) = 1.78$

他の距離（類似度）を定義する手段

■ 遺伝子（or サンプル）***x***と***y***の発現パターンの距離$D(x,y)$

□ユークリッド距離 $D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

□マンハッタン距離 $D = \sum_{i=1}^{n}|x_i - y_i|$

□最大距離 $D = \max(|x_1 - y_1|, \ldots, |x_i - y_i|, \ldots, |x_n - y_n|)$

□キャンベラ距離 $D = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i + y_i|}$

□…

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

計算例（サンプル*x*と*y*間の距離*D*）

	x	y	|xi − yi|	|xi − yi|/|xi + yi|
gene1	10.5	12.4	1.9	0.0830
gene2	6.4	7.1	0.7	0.0519
gene3	8	8.5	0.5	0.0303
gene4	10.8	11.4	0.6	0.0270
gene5	5.6	6.7	1.1	0.0894
gene6	8.4	8.9	0.5	0.0289
gene7	6.2	7	0.8	0.0606
gene8	6.1	6.8	0.7	0.0543
gene9	6.6	6.5	0.1	0.0076
gene10	5.1	5.8	0.7	0.0642

$$D = \sum_{i=1}^{n} |x_i - y_i|$$

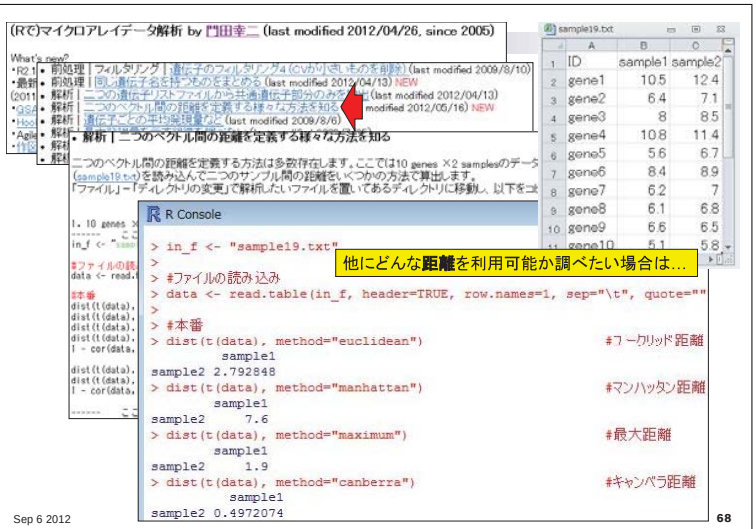
マンハッタン距離 = 1.9+0.7+0.5+0.6+1.1+0.5+0.8+0.7+0.1+0.7 = 7.6

$$D = \max(|x_i - y_i|)$$

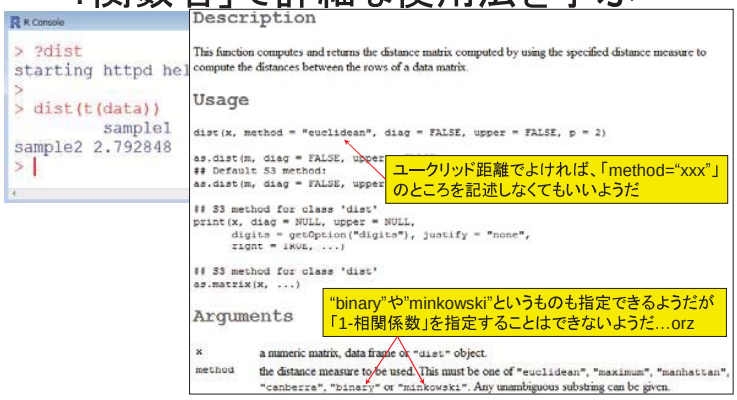
最大距離 = max(1.9, 0.7, 0.5, 0.6, 1.1, 0.5, 0.8, 0.7, 0.1, 0.7) = 1.9

$$D = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|}$$

キャンベラ距離 = 0.0830+0.0519+0.0303+…+0.0642 = 0.4972

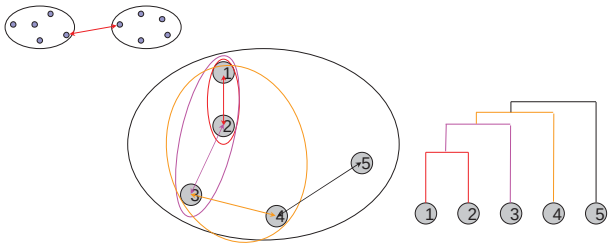


「?関数名」で詳細な使用法を学ぶ



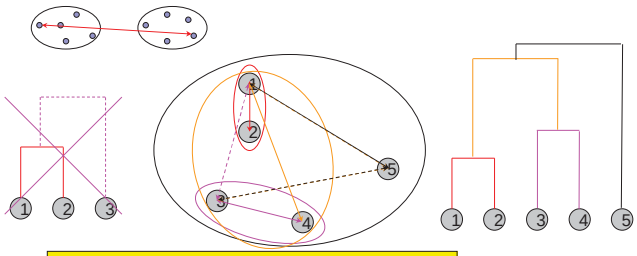
クラスターを併合する方法

- 最短距離法（単連結法；single-linkage）
  - 二つのクラスター間の類似度を、それぞれに含まれる要素対の中で、最も類似性が高い対の間の類似度で定義



クラスターを併合する方法

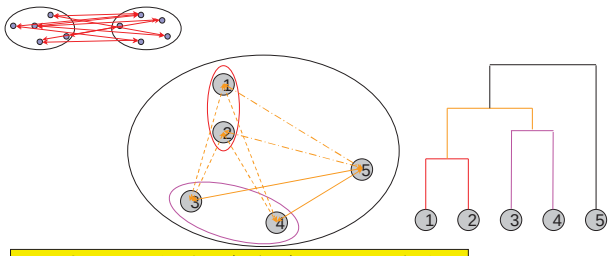
- 最長距離法（完全連結法；complete-linkage ）
  - クラスターをmerge（併合、合併）するときの基準として、最遠距離を用いる



電気泳動波形解析（ピークアラインメント）にも応用可能

クラスターを併合する方法

- 群平均法（平均連結法；average-linkage ）
  - クラスターをmergeするときの基準として、群間平均距離を用いる



用いる方法によって得られるデンドログラム（樹形図）が異なる

http://ua.t.u-tokyo.ac.jp/okabelab/asami/clustering.pdf
http://mjin.doshisha.ac.jp/R/28.pdf

他のクラスター併合手段

- 重心法（Centroid）：重心間距離を利用
- ウォード法：群内平方和の増加量が最小となるクラスターと併合
- メディアン（Median）法：群間距離の中央値を利用
- McQuitty法…
- 可変（flexible）法…

Nakai *et al.*, *Biosci. Biotechnol. Biochem.*, **72**, 139-148, 2008

実データをクラスタリングしてみよう

褐色脂肪_満腹：4サンプル

褐色脂肪_空腹：4サンプル

白色脂肪_満腹：4サンプル

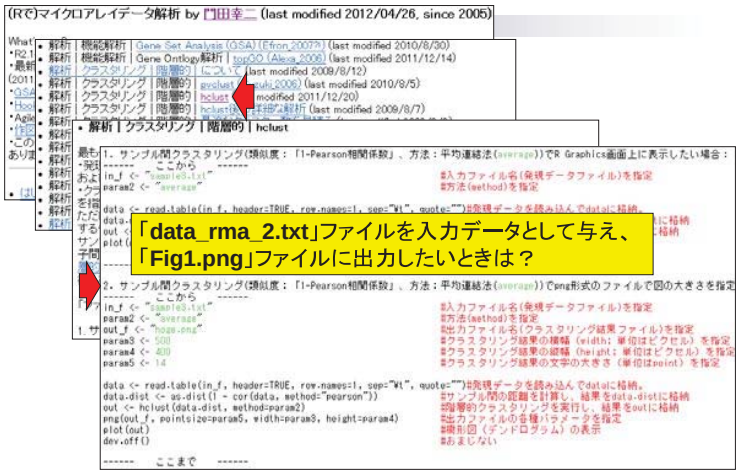
白色脂肪_空腹：4サンプル

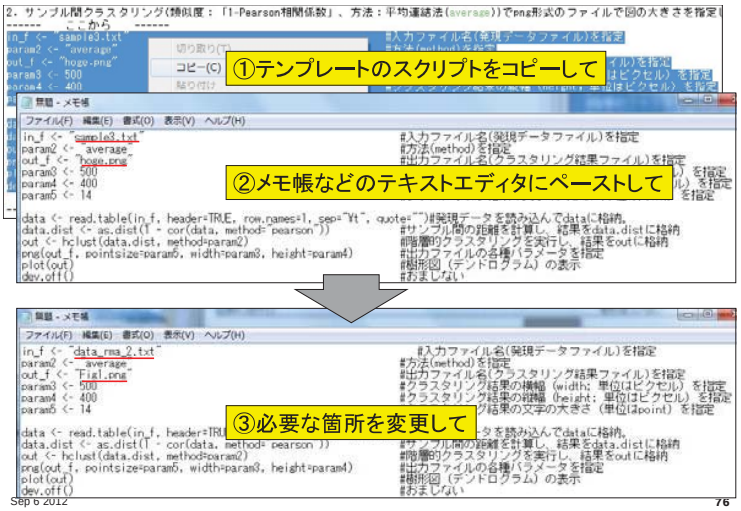
肝臓_満腹：4サンプル

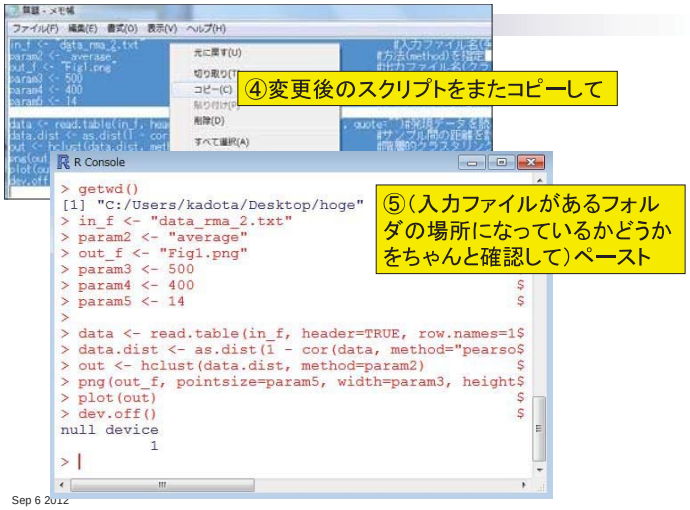
肝臓_空腹：4サンプル

「hoge」–「**data_rma_2.txt**」

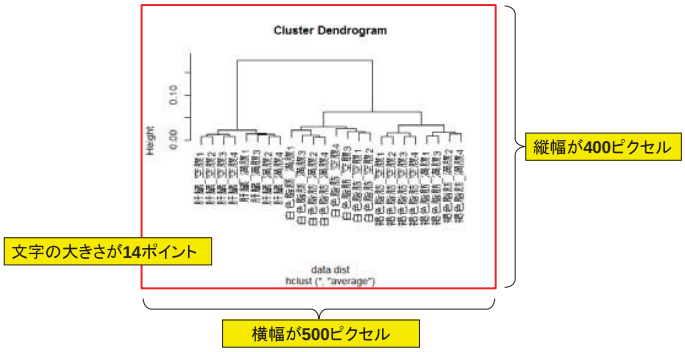
	褐色脂肪_満腹1	褐色脂肪_満腹2	褐色脂肪_満腹3	…
1367452_at	10.52	10.23	10.35	
1367453_at	9.66	10.05	9.90	
1367454_at	9.65	9.40	9.44	
1367455_at	10.76	11.10	10.82	
1367456_at	11.71	11.60	11.59	
…				







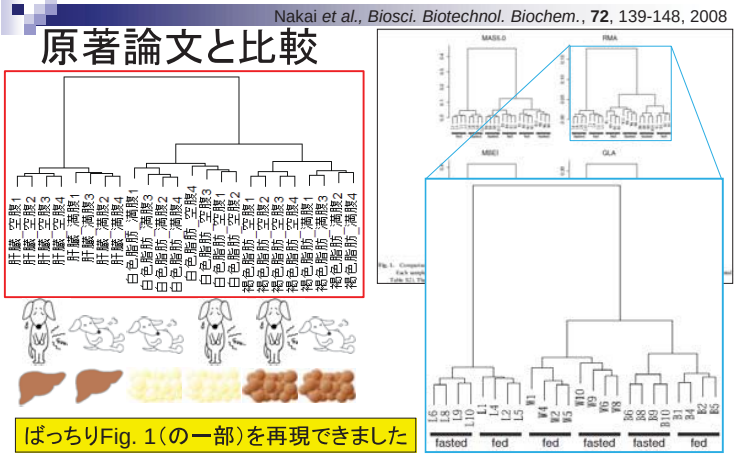
出力ファイル（Fig1.png）の解説




```
param3 <- 500          #クラスタリング結果の横幅（width; 単位はピクセル）を指定
param4 <- 400          #クラスタリング結果の縦幅（height; 単位はピクセル）を指定
param5 <- 14           #クラスタリング結果の文字の大きさ（単位はpoint）を指定
```

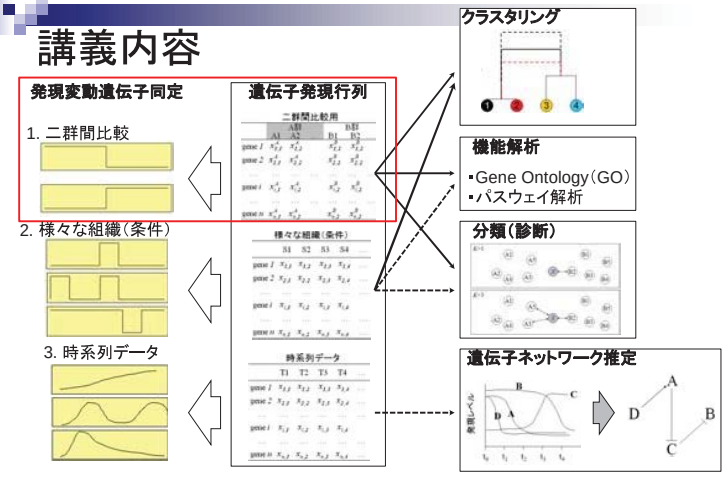
Nakai *et al.*, *Biosci. Biotechnol. Biochem.*, **72**, 139-148, 2008

原著論文と比較

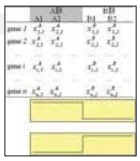


ばっちりFig. 1（の一部）を再現できました

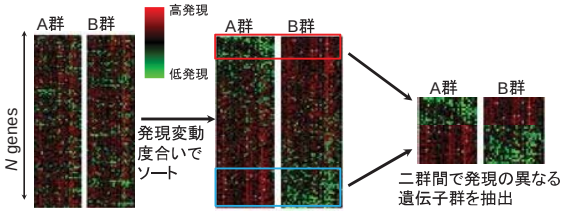
講義内容



1. 二群間比較



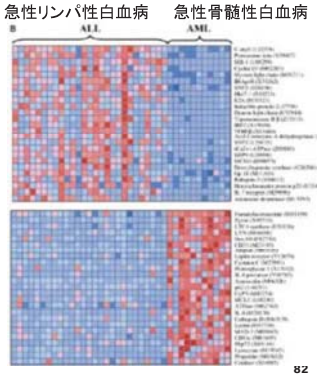
- 農産物の栽培条件の違い（通常 vs. 低温、通常 vs. 乾燥）
- 味の違い（おいしい vs. まずい）
- サンプルの状態の違い（癌 vs. 正常）



参考

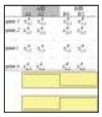
解析例（二群間比較）

- Golub *et al.*, *Science*, **286**: 531-537, 1999.
  - □A：ALL（27サンプル）
  - □B：AML（11サンプル）



発現の異なる遺伝子群を同定するとともに、分類（診断）に適用

1. 二群間比較



- パターンマッチング法
  - □理想的なパターンyとの類似度が高い順にランキング

$$\text{相関係数 } r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (-1 \le r \le 1)$$

y	1	1	1	1	1	1	0	0	0	0	0

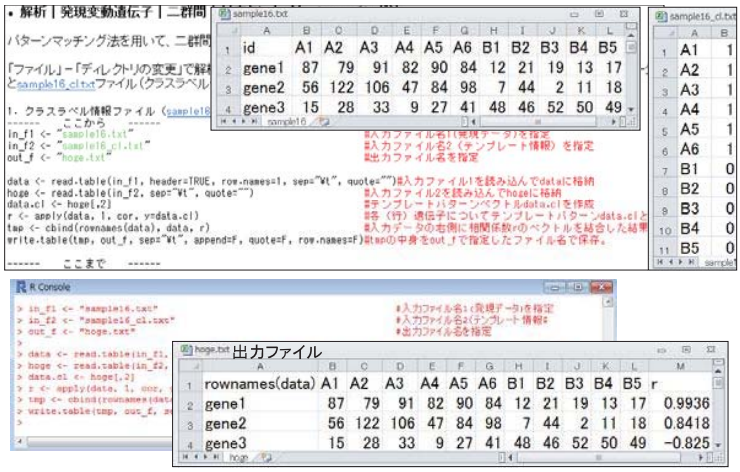
	A群						B群				
	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5
x_{gene1}	87	79	91	82	90	84	12	21	19	13	17
x_{gene2}	56	122	106	47	84	98	7	44	2	11	18
x_{gene3}	15	28	33	9	27	41	48	46	52	50	49

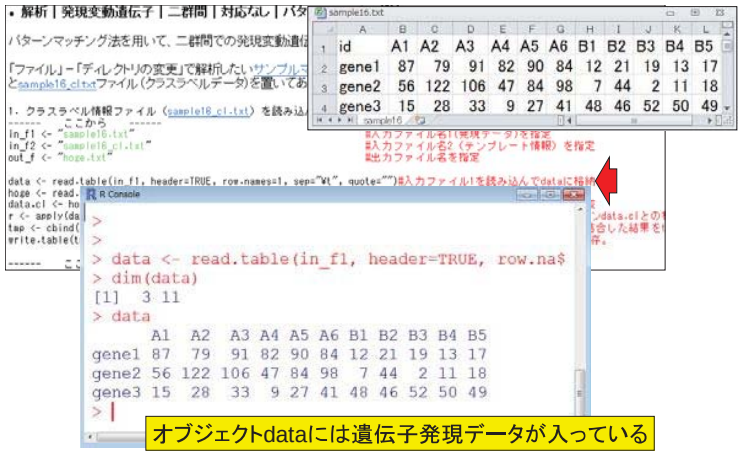
$$r_{gene1} = \frac{18.85}{36.32 \times 0.52} = 0.994$$

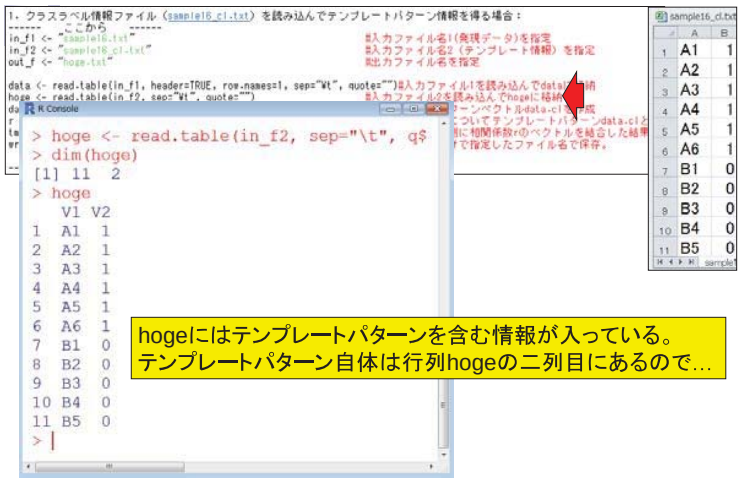
$$r_{gene2} = \frac{18.85}{42.87 \times 0.52} = 0.842$$

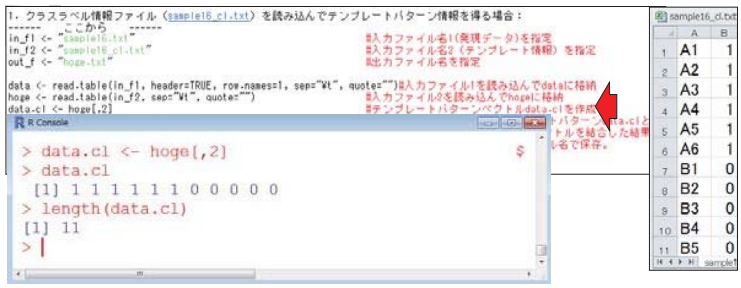
$$r_{gene3} = \frac{-6.41}{14.88 \times 0.52} = -0.825$$

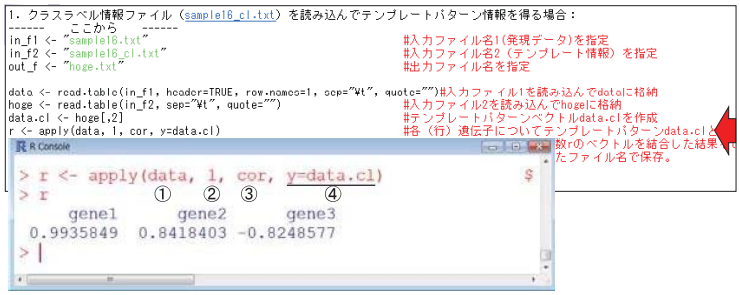
相関係数rの絶対値が大きいほど発現変動の度合いが大きい、と解釈


出力ファイル

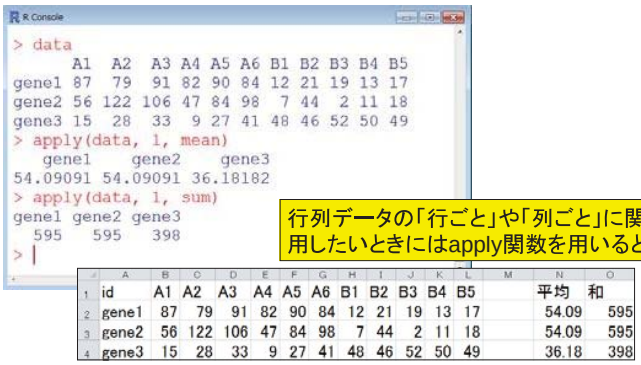

オブジェクトdataには遺伝子発現データが入っている


hogeにはテンプレートパターンを含む情報が入っている。
テンプレートパターン自体は行列hogeの二列目にあるので…


hogeから二列目の情報のみ抽出して、その結果をdata.clオブジェクトに格納している。
当然、data.clの要素数はA群B群のサンプル総数と同じ11個


①dataオブジェクトの、②各行に対して、③cor関数を適用せよ。その際に④引数yはdata.clで与える

apply関数


行列データの「行ごと」や「列ごと」に関数を適用したいときにはapply関数を用いると便利

apply関数

```
> apply(data, 2, mean)
      A1       A2       A3       A4       A5       A6
52.66667 76.33333 76.66667 46.00000 67.00000 74.33333
      B1       B2       B3       B4       B5
22.33333 37.00000 24.33333 24.66667 28.00000
> apply(data, 2, sum)
 A1  A2  A3  A4  A5  A6  B1  B2  B3  B4  B5
158 229 230 138 201 223  67 111  73  74  84
>
```

「列ごと」に何かしたい場合には、apply関数の二番目の引数を2にします

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5
2	gene1	87	79	91	82	90	84	12	21	19	13	17
3	gene2	56	122	106	47	84	98	7	44	2	11	18
4	gene3	15	28	33	9	27	41	48	46	52	50	49
5												
6	平均	52.67	76.33	76.67	46.00	67.00	74.33	22.33	37.00	24.33	24.67	28.00
7	和	158	229	230	138	201	223	67	111	73	74	84


```
> ?apply
>
```

apply {base} R Documentation

Apply Functions Over Array Margins

Description

Returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.

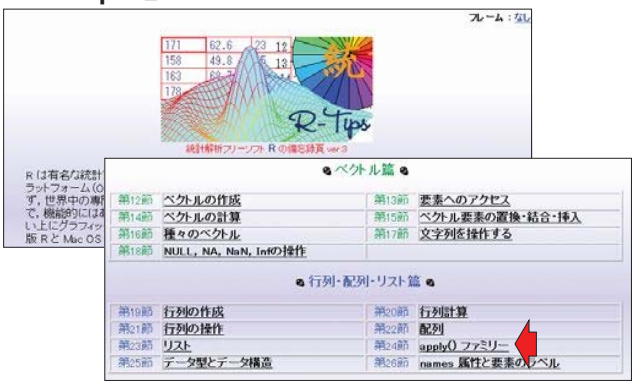
Usage

```
apply(X, MARGIN, FUN, ...)
```

Arguments

X	an array, including a matrix
MARGIN	a vector giving the subscripts which the function will be applied over. E.g., for a matrix 1 indicates rows, 2 indicates columns, c(1, 2) indicates rows and columns. Where X has named dimnames, it can be a character vector selecting dimension names.
FUN	the function to be applied: see 'Details'. In the case of functions like +, %*%, etc., the function name must be backquoted or quoted.
...	optional arguments to FUN.

「R-Tips」にも記述あり




```
> ?cor
>
```

cor {stats} R Documentation

Correlation, Variance and Covariance (Matrices)

Description

var, cov and cor compute the variance of x and the covariance or correlation of x and y if these are vectors. If x and y are matrices then the covariances (or correlations) between the columns of x and the columns of y are computed.

cov2cor scales a covariance matrix into the corresponding correlation matrix *efficiently*.

Usage

```
var(x, y = NULL, na.rm = FALSE, use)

cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor(x, y = NULL, use = "everything",
     method = c("pearson", "kendall", "spearman"))

cov2cor(V)
```

Spearman相関係数なども計算できそうだ

Arguments

x	a numeric vector, matrix or data frame.
y	NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).

apply関数を使いこなす

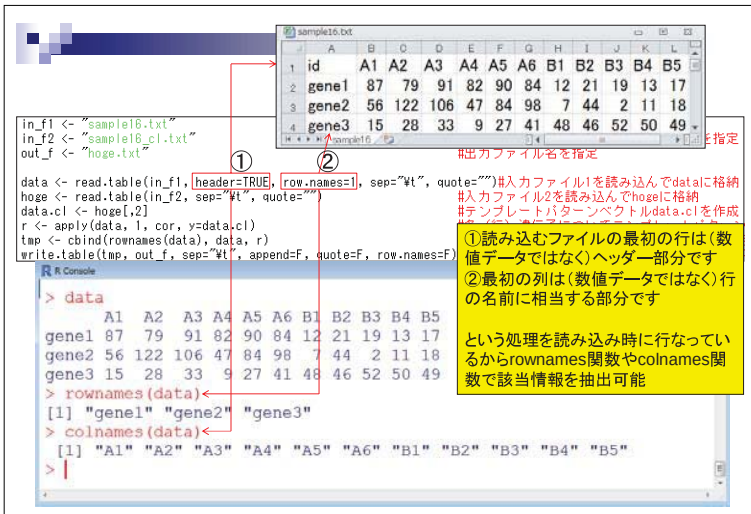
```
> apply(data, 1, cor, y=data.cl, method="spearman")
     gene1      gene2      gene3
 0.8660254  0.8660254 -0.8660254
> apply(data, 1, cor, y=data.cl, method="pearson")
     gene1      gene2      gene3
 0.9935849  0.8418403 -0.8248577
> apply(data, 1, cor, y=data.cl)
     gene1      gene2      gene3
 0.9935849  0.8418403 -0.8248577
>
```

apply関数中で用いる関数（例：cor）の引数もこんな感じで自在に操れます

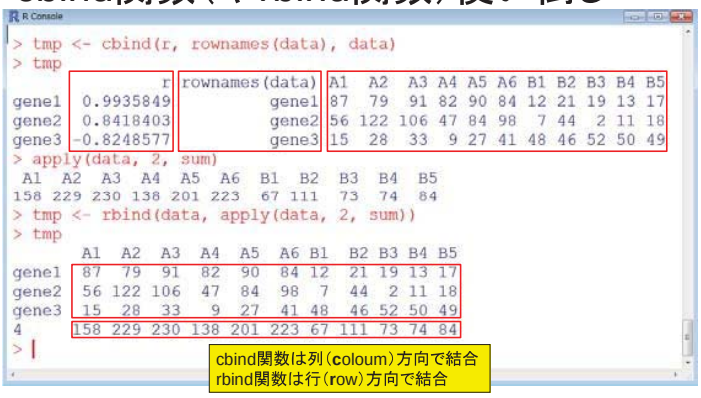

```
in_f1 <- "sample16.txt"                    #入力ファイル名1(発現データ)を指定
in_f2 <- "sample16_cl.txt"                 #入力ファイル名2（テンプレート情報）を指定
out_f <- "hoge.txt"                        #出力ファイル名を指定

data <- read.table(in_f1, header=TRUE, row.names=1, sep="¥t", quote="")#入力ファイル1を読み込んでdataに格納
hoge <- read.table(in_f2, sep="¥t", quote="")   #入力ファイル2を読み込んでhogeに格納
data.cl <- hoge[,2]                        #テンプレートパターンベクトルdata.clを作成
r <- apply(data, 1, cor, y=data.cl)        #各（行）遺伝子についてテンプレートパターンの
tmp <- cbind(rownames(data), data, r)      #入力データの右側に相関係数rのベクトルを結合
```

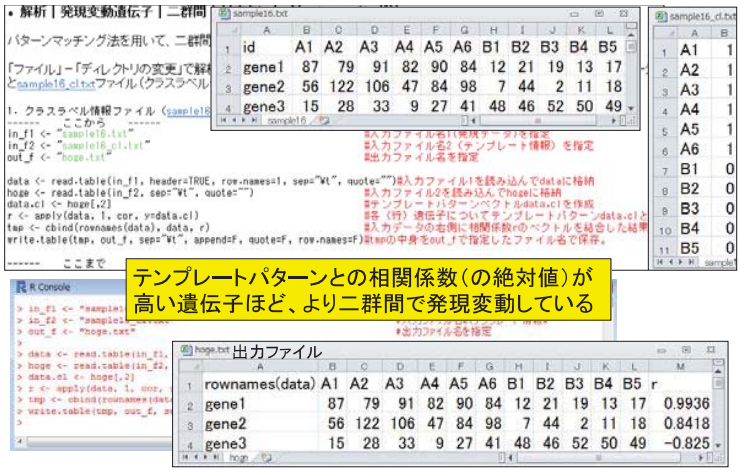
```
> data
      A1  A2  A3 A4 A5 A6 B1 B2 B3 B4 B5
gene1 87  79  91 82 90 84 12 21 19 13 17
gene2 56 122 106 47 84 98  7 44  2 11 18
gene3 15  28  33  9 27 41 48 46 52 50 49
> r
     gene1      gene2      gene3
 0.9935849  0.8418403 -0.8248577
> tmp <- cbind(rownames(data), data, r)
> tmp
      rownames(data) A1  A2  A3 A4 A5 A6 B1 B2 B3 B4 B5          r
gene1          gene1 87  79  91 82 90 84 12 21 19 13 17  0.9935849
gene2          gene2 56 122 106 47 84 98  7 44  2 11 18  0.8418403
gene3          gene3 15  28  33  9 27 41 48 46 52 50 49 -0.8248577
>
```

cbind関数を用いて元の遺伝子発現行列データdataの「左側に遺伝子名情報」を、「右側に相関係数を計算した結果」を結合している



cbind関数(やrbind関数)使い倒し





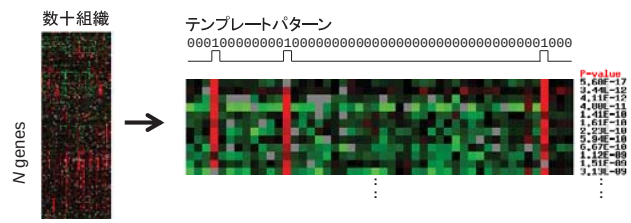
2. 様々な組織(条件)

- 組織(条件)特異的(選択的)発現遺伝子
  - 特定の(複数)組織のみで高(or 低)発現している遺伝子
    - 「脳」特異的遺伝子
    - 「心臓」特異的
    - 「腸」特異的
    - …
    - **栄養条件**特異的

Pavlidis and Noble, *Genome Biol.*, **2**, research0042, 2001

2. 様々な組織(条件)

- パターンマッチング法



• 解析 | 発現変動遺伝子 | 組織特異的(選択的)発現遺伝子 | パターンマッチング法(テンプレートマッチング法)

(基本的には、解析 | 似た発現パターンを持つ遺伝子の同定をご覧ください。)
パターンマッチング法を用いて、指定した理想的なパターンとの類似度が高い遺伝子の同定を行うやり方を紹介します。

「ファイル」-「ディレクトリの変更」で解析したいサンプルマイクロアレイデータ14中のsample15.txtファイル(遺伝子発現データ)とsample15_cl.txtファイル(sample4で特異的高発現パターンを検出するためのテンプレートパターンのデータ)を置いてあるディレクトリに移動し、以下をコピペ

```
1. 基本形:
------  ここから  ------
in_f1 <- "sample15.txt"
in_f2 <- "sample15_cl.txt"
out_f <- "hoge1.txt"

data <- read.table(i
hoge <- read.table(i
data.cl <- hoge[,2]
r <- apply(data, 1,

tmp <- cbind(rownam
write.table(tmp, out

------  ここまで
```

#入力ファイル名1(発現データ)を指定
#入力ファイル名2(テンプレート情報)を指定

sample15.txt

	A	B	C	D	E	F	G	H	I
1	id	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8
2	gene1	1	0	0	9	0	0	0	0
3	gene2	5	2	1	2	4	6	3	5
4	gene3	6	6	6	6	6	6	6	6
5	gene4	4	4	4	4	10	4	4	4
6	gene5	10	10	10	10	4	10	10	10

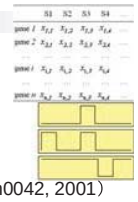
sample15_cl.txt

	A	B
1	tissue1	0
2	tissue2	0
3	tissue3	0
4	tissue4	1
5	tissue5	0
6	tissue6	0
7	tissue7	0
8	tissue8	0

hoge1.txt

	A	B	C	D	E	F	G	H	I	J
1	rownam	tissue1	tissue2	tissue3	tissue4	tissue5	tissue6	tissue7	tissue8	r
2	gene1	1	0	0	9	0	0	0	0	0.99
3	gene2	5	2	1	2	4	6	3	5	−0.34
4	gene3	6	6	6	6	6	6	6	6	NA
5	gene4	4	4	4	4	10	4	4	4	−0.14
6	gene5	10	10	10	10	4	10	10	10	0.14

2. 様々な組織(条件)



- ランキングに基づく方法
  - Dixon test (Greller and Tobin, *Genome Res.*, **9**, 282-296, 1999)
  - Pattern matching(Pavlidis and Noble, *Genome Biol.*, **2**, research0042, 2001)
  - Entropy (Schug *et al.*, *Genome Biol.*, **6**, R33, 2005)
  - Tissue specificity Index (Yanai *et al.*, *Bioinformatics*, **21**, 650-659, 2005)
- 外れ値検出に基づく方法
  - Akaike's Information Criterion (AIC) (Kadota *et al.*, *Physiol. Genomics*, **12**, 251-259, 2003)
  - Sprent's non-parametric method (Ge *et al.*, *Genomics*, **86**, 127-141, 2005)
- その他
  - Tukey-Kramer's Honest Significance Difference (HSD) test (Liang *et al.*, *Physiol. Genomics*, **26**, 158-162, 2006)
  - ROKU (Kadota *et al.*, *BMC Bioinformatics*, **7**, 294, 2006)
  - QDMR (Zhang *et al.*, *Nucleic Acids Res.*, **39**, e58, 2011)

• 解析 | 発現変動遺伝子 | 組織特異的(選択的)パターン | Schug's H(x) statistic (Schug 2005) (last modified 2011/10/13)

エントロピー(組織特異的遺伝子検出)

- 遺伝子iのエントロピー $H(x_i) = -\sum_{j=1}^{N} p_{ij} \log_2(p_{ij})$, where $p_{ij} = x_{ij} / \sum_{j=1}^{N} x_{ij}$

N:組織数(jの数) = 8
Hの取りうる範囲:$0 \leqq H \leqq \log_2 N \rightarrow 0 \leqq H \leqq 3$

x_{ij} (j: 組織, i: 遺伝子)

	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
組織1	1	5	6	4	10	...
組織2	0	2	6	4	10	...
組織3	0	1	6	4	10	...
組織4	9	2	6	4	10	...
組織5	0	4	6	10	4	
組織6	0	6	6	4	10	
組織7	0	3	6	4	10	
組織8	0	5	6	4	10	
$\Sigma_j x_{ij}$	10	28	48	38	74	

p_{ij}

	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.10	0.18	0.13	0.11	0.14	...
2	0.00	0.07	0.13	0.11	0.14	...
3	0.00	0.04	0.13	0.11	0.14	...
4	0.90	0.07	0.13	0.11	0.14	...
5	0.00	0.14	0.13	0.26	0.05	
6	0.00	0.21	0.13	0.11	0.14	
7	0.00	0.11	0.13	0.11	0.14	
8	0.00	0.18	0.13	0.11	0.14	
Σ_j	1.00	1.00	1.00	1.00	1.00	

$p_{ij} \log_2(p_{ij})$

	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.33	0.44	0.38	0.34	0.39	...
2	0.00	0.27	0.38	0.34	0.39	...
3	0.00	0.17	0.38	0.34	0.39	...
4	0.14	0.27	0.38	0.34	0.39	...
5	0.00	0.40	0.38	0.51	0.23	
6	0.00	0.48	0.38	0.34	0.39	
7	0.00	0.35	0.38	0.34	0.39	
8	0.00	0.44	0.38	0.34	0.39	
Σ_j	0.47	2.83	3.00	2.90	2.96	

組織特異的遺伝子は低いエントロピー　そうでないものは高い値

Schneider and Stephens., *Nucleic Acids Res.*, **18**, 6097-6100, 1990

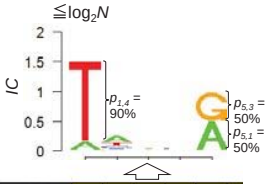
エントロピー(Sequence logos)

- position iの情報量 $IC_i = \frac{\log_2(N)}{n} - H(x_i)$

N:塩基の種類数= 4
Hの取りうる範囲: $0 \leqq H \leqq \log_2 N$

position i

	1	2	3	4	5	...
配列 1	T	A	C	G	G	...
配列 2	T	A	A	C	G	...
配列 3	T	G	T	A	G	...
配列 4	A	C	T	T	A	...
配列 5	T	T	G	G	A	...
配列 6	T	C	A	A	G	...
配列 7	T	A	C	T	A	...
配列 8	T	T	G	C	A	...
配列 9	T	A	A	C	A	...
配列10	T	A	C	T	G	...



	1	2	3	4	5	...
IC	1.53	0.24	0.03	0.03	1.00	...

x_{ij} (j: A, C, G, T, i: 1 2 3 4 5 ...)

	1	2	3	4	5	...
Aの数 (j=1)	1	5	3	2	5	...
Cの数 (j=2)	0	2	3	3	0	...
Gの数 (j=3)	0	1	2	2	5	...
Tの数 (j=4)	9	2	2	3	0	...
$\Sigma_j x_{ij}$	10	10	10	10	10	

p_{ij}

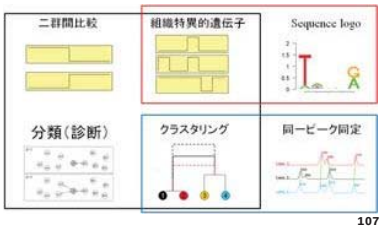
	1	2	3	4	5	...
1	0.1	0.5	0.3	0.2	0.5	...
2	0.0	0.2	0.3	0.3	0.0	...
3	0.0	0.1	0.2	0.2	0.5	...
4	0.9	0.2	0.2	0.3	0.0	...
Σ_j	1.0	1.0	1.0	1.0	1.0	

$-p_{ij} \log_2(p_{ij})$

	1	2	3	4	5	...
1	0.33	0.50	0.52	0.46	0.50	...
2	0.00	0.46	0.52	0.52	0.00	...
3	0.00	0.33	0.46	0.46	0.50	...
4	0.14	0.46	0.46	0.52	0.00	...
$H = \Sigma_j$	0.47	1.76	1.97	1.97	1.00	

まとめ

- マイクロアレイおよび次世代型シーケンサ(NGS)によるトランスクリプトーム解析について紹介
  - Rを利用することで、様々なトランスクリプトーム解析が可能
- バイオインフォマティクスの基本的なスキルを身につけることが重要
  - バイオインフォマティクス技術者認定試験合格を目指せ(11/25)
  - 相関係数やエントロピーなどの要素技術を駆使すれば様々なデータ解析が可能であることを紹介



課題

1. スライド79中の樹形図は大きく二つの群に分かれているが、何と何に大別可能か？**①～③の中から選べ**:
   ①肝臓と脂肪、②空腹と満腹、③白色と褐色
2. (スライド79中の樹形図を眺めて)**次の二つのうち**どちらの発現プロファイルの組合せがより似ているか？:
   ①「白色脂肪_空腹2」と「褐色脂肪_空腹1」、②「褐色脂肪_満腹1」と「褐色脂肪_満腹4」
3. (スライド83, 89, 99を参考にして)sample16.txt中の**三つの遺伝子のうち**、テンプレートパターンyと**Pearson相関係数が最も高かったもの**はどれか？:
   ①gene1, ②gene2, ③gene3
4. (スライド103):sample15.txt中の**五つの遺伝子のうち**、テンプレートパターンとの**Pearson相関係数が最も低かったもの**はどれか？(計算できなかったものは除く):
   ①gene1, ②gene2, ③gene3, ④gene4, ⑤gene5
5. (スライド103):sample15_cl.txt中で指定した数値ベクトルは、「tissue4という組織で特異的高発現パターンを示す遺伝子を抽出するためのもの」という解釈が可能である。では、「**tissue4特異的低発現パターンを示す遺伝子**」を抽出したい場合には、どのようにすればいいだろうか？考えを簡潔に述べよ。

1.「門田幸二」のホームページへのリンク
http://www.iu.a.u-tokyo.ac.jp/~kadota/

2.「(Rで)塩基配列解析」ウェブページへのリンク
http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

3.「(Rで)マイクロアレイデータ解析」ウェブページへのリンク
http://www.iu.a.u-tokyo.ac.jp/~kadota/r.html

4.「講義資料（公開版）」へのリンク
http://www.iu.a.u-tokyo.ac.jp/~kadota/20120906_kadota.pdf

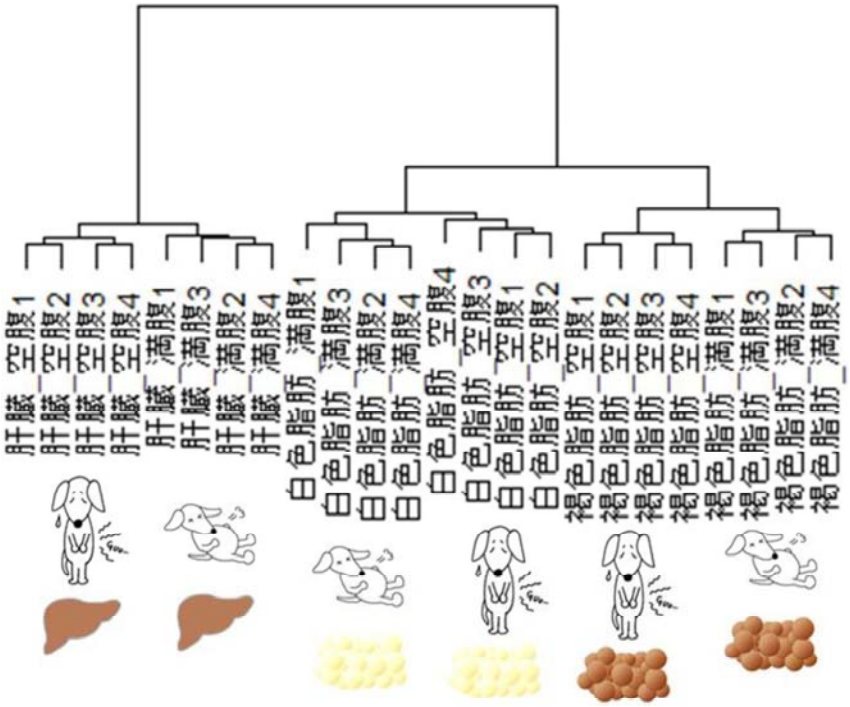
5.「課題」へのリンク
http://www.iu.a.u-tokyo.ac.jp/~kadota/20120906_kadai.pdf

「食品機能科学特別講義Ⅰ」 第四部 生命科学実習③（9/6 分）課題

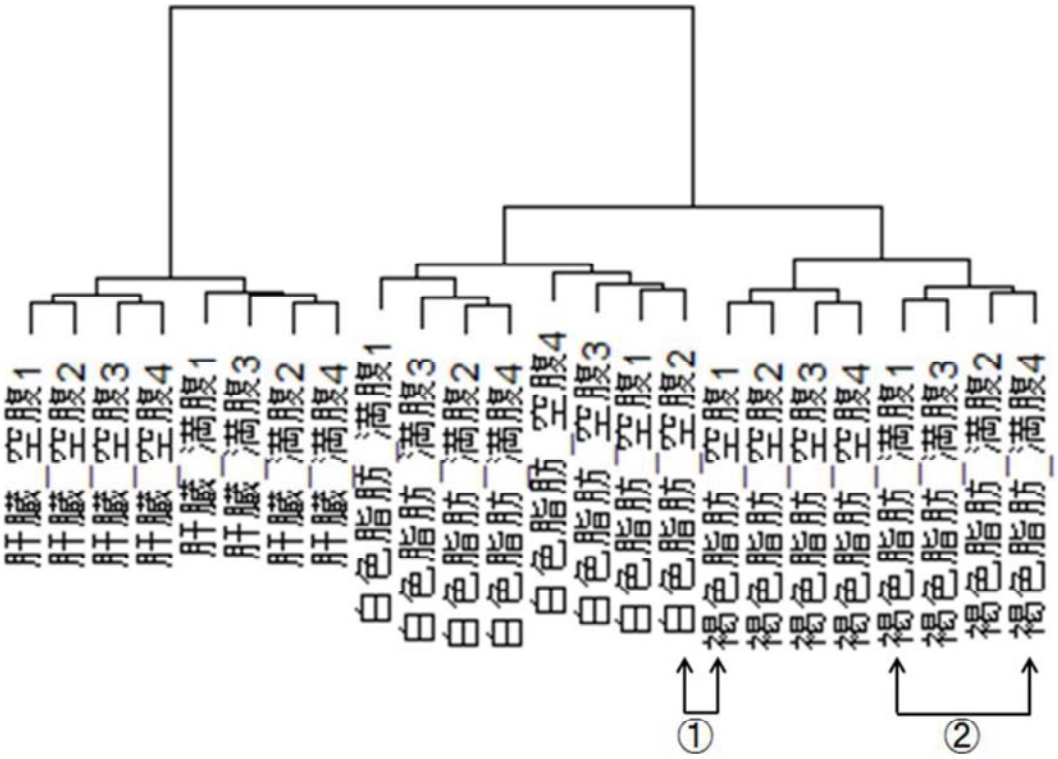
所属：

名前：

課題 1（スライド 79）：以下の樹形図は大きく二つの群に分かれているが、何と何に大別可能か？①～③の中から選べ：①肝臓と脂肪、②空腹と満腹、③白色と褐色



課題 2（スライド 64, 72, 79）：次の二つのうちどちらの発現プロファイルの組合せがより似ているか？：①「白色脂肪_空腹 2」と「褐色脂肪_空腹 1」、②「褐色脂肪_満腹 1」と「褐色脂肪_満腹 4」



課題 3（スライド 83, 89, 99）：sample16.txt 中の三つの遺伝子のうち、テンプレートパターン $\boldsymbol{y}$ との Pearson 相関係数が**最も高かったもの**はどれか？：①gene1, ②gene2, ③gene3

課題 4（スライド 103）：sample15.txt 中の五つの遺伝子のうち、テンプレートパターンとの Pearson 相関係数が**最も低かったもの**はどれか？（計算できなかったものは除く）：①gene1, ②gene2, ③gene3, ④gene4, ⑤gene5

課題 5（スライド 103）：sample15_cl.txt 中で指定した数値ベクトルは、「tissue4 という組織で特異的高発現パターンを示す遺伝子を抽出するためのもの」という解釈が可能である。では、「tissue4 特異的**低発現**パターンを示す遺伝子」を抽出したい場合には、どのようにすればいいだろうか？考えを簡潔に述べよ。

3.5　生命科学実習④

3.5.1 概要

講師：田邊 俊朗／沖縄工業高等専門学校生物資源工学科准教授

日時：2012 年 9 月 7 日　10:20～／12:50～／14:40 ～

場所：沖縄工業高等専門学校 SE-271 生物化学工学実験室

内容：機能を担うタンパク質の総体を取り扱うプロテオーム解析として、目的とするタンパク質の検出を行う活性染色法や、二次元電気泳動および質量分析によるタンパク質発現のサンプル間比較を実習し、タンパク質同定の技術について学ぶ。

ＴＡ：3 名

実習体制：5 名/班×3 班（受講生 14 名）

3.5.2 準備

機器

□卓上遠心機	himac CT 13R／HITACHI
□ローテーター	Rotator RT50／TAITEC
□タンパク質泳動装置	BioRAD
□パワーサプライ	BioRAD
□インキュベーター	
□二次元電気泳動装置	GE Healthcare
□Strip Holder	Ettan IPGphor3
□ZipTip $_{c18}$ ピペットチップ	適量
□ターゲットプレート	1 本
□質量分析装置	島津製作所

実験器具・試料・試薬

□工業用アクレモニウムセルラーゼ（明治製菓製）	
□エッペンチューブ	適量
□15mL チューブ（試薬分注）	4 個
□p20 ピペットマン	1 個
□p200 ピペットマン	1 個
□イエローチップ	1 箱
□ピンセット	1 個
□手袋	4 個
□タッパー	2 個
□サランラップ	適量
□Immobiline DryStrip（7cm、3-11NL）	

□Lysis バッファー　　2mL(0.5mL/人×４人)
□100mM　DTT 入り SDS 化処理液
□25%イソプロピルアルコール入り 0.05M リン酸緩衝液(pH6.8)
□0.1%CBB R-250 染色液　　適量
□0.1%Congo Red　　適量
□1M NaCl　　適量
□0.1%TFA 入り 50%アセトニトリル　　30μL／サンプル
□0.1%TFA　　60μL／サンプル
□マトリックス飽和溶液　　2μL／サンプル

3.5.3 配布物
・実習プロトコル

3.5.4 課題
・プリント（記述３題）

3.5.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.5.6時間割

時間	内容
10:24～	<u>イントロ</u> ・　自己紹介、経歴紹介 ・　本日のテーマ 　　プロテオーム解析 ・　活性染色（酵素とは何か） 　　分類と名称 　　EC1：オキシドレダクターゼ（酸化還元酵素） 　　EC2：トランスフェラーゼ（転移酵素） 　　EC3：ヒドラーゼ（加水分解酵素） 　　EC4：リアーゼ（脱離酵素） 　　EC5：イソメラーゼ（異性化酵素） 　　EC6：リガーゼ（合成酵素）
10:36～	【SDS-PAGE】 工業用アクレモニウムセルラーゼ（50μL/μg）を100mMDTT入りSDS化処理液で100℃で５分加熱処理 　　＊　時間の都合上、事前に用意 ↓　１人１サンプルをアクリルアミドゲルにロード ↓　200Vで電気泳動 　　＊　時間の都合上、ここで終了。 　　＊　以降は事前に用意してある、泳動後、脱SDSをしたゲルに交換
11:08～	<u>SDS-PAGEについて</u> ・　SDS-PAGEの原理紹介 ・　その後の操作を解説 　　イソプロパノールによる脱SDS 　　酵素活性の原理紹介 【活性染色】
11:16～	脱SDS済みのゲル（事前に用意） ↓　基質を含むゲルときれいに重ね合わせる ↓　ラップできれいに包む
11:32～	↓　37℃で30～60分間インキュベート

11:35～	その他の電気泳動法 ・ native-PAGE との比較
11:47～	昼食
11:50～	【ゲルの染色】 インキュベート後のゲル ↓ アクリルアミドゲルを CBB R-250 にて染色 ↓ アガロースゲルは 0.1% Congo Red にて染色 ↓ 振盪機を使い、20～40 分染色
13:07～	等電点電気泳動について ・ アミノ酸の性質 20 種類のアミノ酸からなる（翻訳後修飾によりもっと多い） アミノ酸の構造（不斉炭素、L 型と D 型） 極性アミノ酸、荷電アミノ酸、芳香族アミノ酸など 等電点＝タンパク質の表面電荷が 0 になる pH
	【等電点電気泳動】
13:31～	工業用アクレモニウムセルラーゼをサンプルとする ↓ 膨潤液の作成（時間の都合上、事前に作成しておいた） ↓ サンプルを 125μL（40～240μg 程度）乗せる ↓ Immobiline DryStrip（7cm、3-11NL）を乗せる ＊ ミネラルオイル等の作業は省略 ↓ 泳動装置にセットし、泳動開始 ＊今回は機械にセットするところまでで終了
13:55～	プロテオーム解析の流れ 等電点電気泳動 1 日～ 二次元電気泳動 1 日～ スキャン 半日～ サンプル調整、解析スタート ・ プロテオーム解析の活用例 ・ 写真等を見せながら詳細な作業例を紹介
14:10～	【ゲルの脱色】 CBB 染色したゲル ＊ 今回は脱色を省略

	Congo Red 染色したゲル ↓　廃液入れに染色液を捨てる ↓　1M NaCl をタッパーに入れる ↓　振盪機で脱色 　　＊ 10 分間休憩
	休憩
14:40～	【ゲルの切り出し、ゲル抽出】 　＊写真を見せながら行程を紹介。作業は省略
14:47～	<u>質量分析について</u> ・ MALDI-TOF-MS ・ 原理と結果
14:57～	【MALDI-TOF-MS　結晶形成】 ゲル抽出後のサンプル（事前に用意したものを使用） ↓　ZipTip による精製 ↓　マトリックスとの共結晶を作成するため、ターゲットプレートに乗せる ↓　自然風乾させ、共結晶の形成をまつ
15:17～	【MALDI-TOF-MS　解析】 　　＊ 部屋のスペースと時間の関係で、3 班に分け、順番に解析を実施。 　　＊ 残りの班は課題に取り組む
15:55～	課題に取り組んだ

平成24年度沖縄県産業振興基金事業
「バイオインフォマティクス人材育成推進事業」
公益財団法人沖縄科学技術振興センター

琉球大学夏期集中講義
～バイオインフォマティクス・スタンダードコース～
農学部　食品機能科学特別講義Ⅰ

第5部　生命科学実習④

沖縄工業高等専門学校生物資源工学科
准教授　田邊俊朗

活性染色

酵素とは？

- 生体触媒
- タンパク質でできている
- 生体内の化学反応　ー>　生命維持

Enzyme Nomenclature

- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)
  国際生化学分子生物学連合の命名法委員会が推奨する酵素の分類
- その酵素が触媒する反応に基づいて分類

- Enzyme Nomenclature 1992 [Academic Press, San Diego, California, ISBN 978-0-12-227164-9 (ハードカバー)として出版、沖縄高専の図書館にも所蔵
- その後補足や追加情報：Supplement 1 [Eur. J. Biochem. 1994, 223, 1-5]、Supplement 2 [Eur. J. Biochem. 1995, 232, 1-6]、Supplement 3 [Eur. J. Biochem. 1996, 237, 1-5]、Supplement 4 [Eur. J. Biochem. 1997, 250, 1-6]、Supplement 5 [Eur. J. Biochem. 1999, 264, 610-650]

酵素の系統的分類と名称

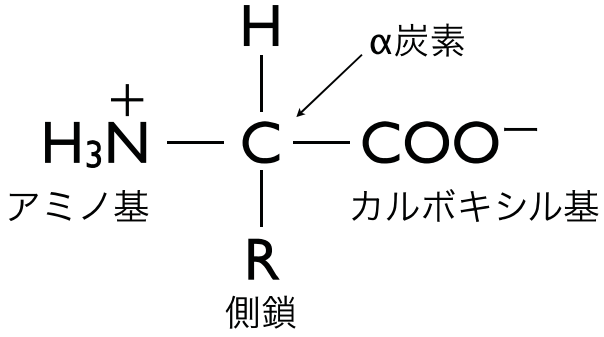
EC 1	オキシドレダクターゼ（酸化還元酵素）
EC 2	トランスフェラーゼ（転移酵素）
EC 3	ヒドロラーゼ（加水分解酵素）
EC 4	リアーゼ（脱離酵素）
EC 5	イソメラーゼ（異性化酵素）
EC 6	リガーゼ（合成酵素）

二次元電気泳動

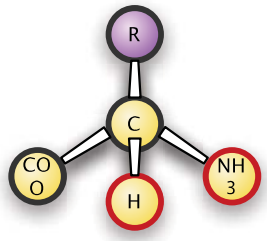
アミノ酸の性質

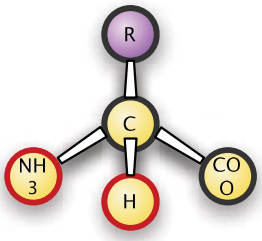
- 天然タンパク質を構成するアミノ酸は通常２０種類
- （翻訳後修飾によって実際は２０種を越える）

アミノ酸の構造



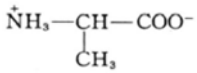
L-アミノ酸とD-アミノ酸



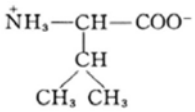


天然タンパク質を構成するのはL-アミノ酸

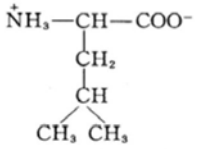
脂肪族アミノ酸



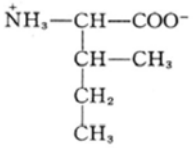
アラニン(Ala, A)



バリン(Val, V)

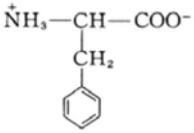


ロイシン(Leu, L)

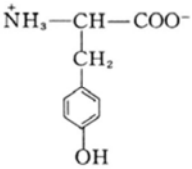


イソロイシン(Ile, I)

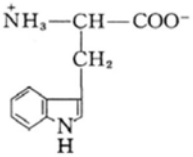
芳香族アミノ酸



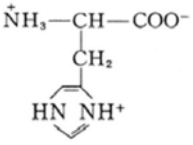
フェニルアラニン
(Phe, F)



チロシン(Tyr, Y)

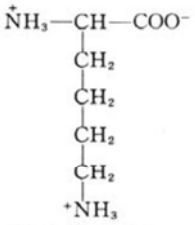


トリプトファン
(Trp, W)

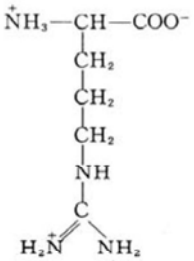


ヒスチジン(His, H)

荷電アミノ酸



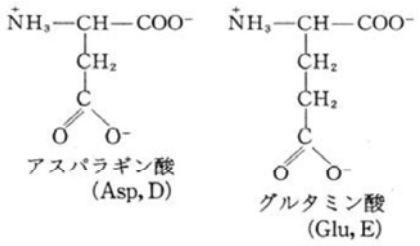
リジン(Lys, K)



アルギニン(Arg, R)

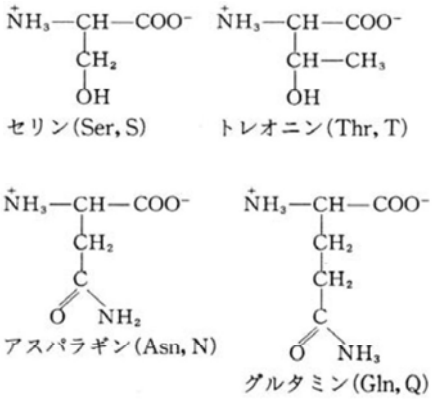
正電荷

荷電アミノ酸２

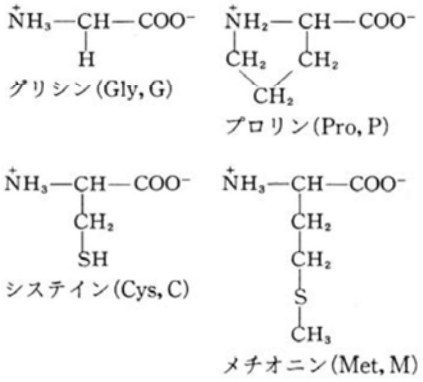


負電荷

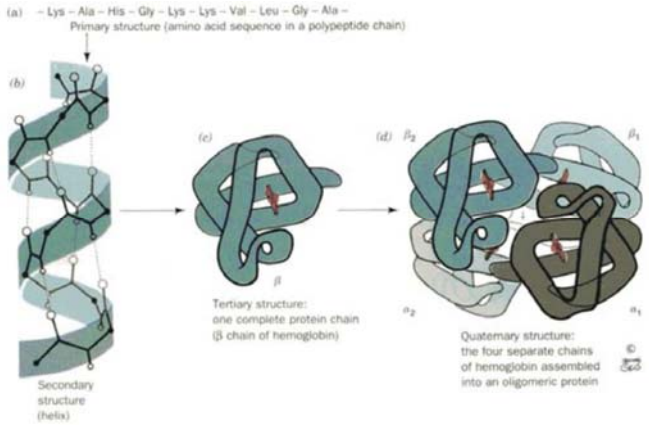
極性アミノ酸



非極性アミノ酸



タンパク質の構造

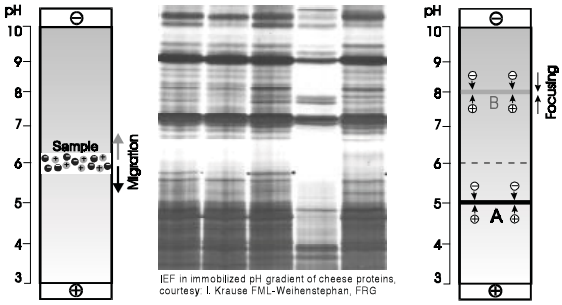


quoted from Donald Voet, Judith Voet, Biochemistry 2nd. ed. , p.106, Figure 6-1.

等電点

- タンパク質の表面電荷が「０」になるpH

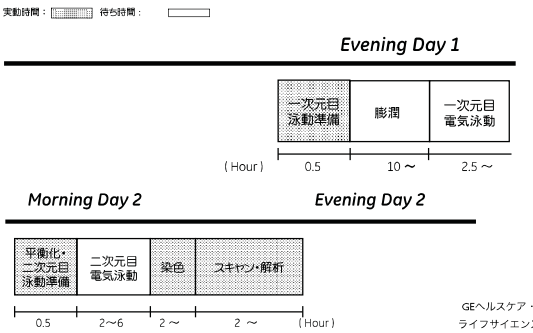
等電点電気泳動



IEF in immobilized pH gradient of cheese proteins, courtesy: I. Krause FML-Weihenstephan, FRG

GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動



GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

① 膨潤液を準備します。
Immobiline DryStrip 1本当たりに必要な膨潤液量は
下表の通りです。(長さにより膨潤液量が変わります)

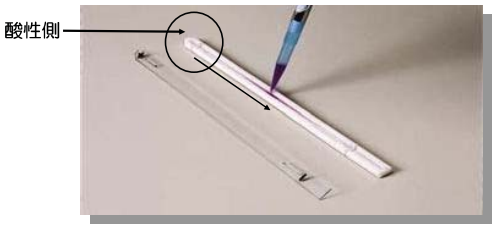
検収で使用

IPG Strip の長さ(cm)	1本あたりの液量(μl)
7	125
11	200
13	250
18	340
24	450

GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

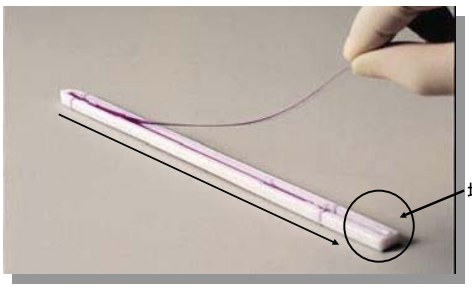
② 各ストリップホルダーに膨潤液を必要量加えます。
膨潤液は中央の溝の酸性側(下記写真参照)から中央にかけて
ゆっくりいれます。



GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

最後にImmobiline DryStripの両端がストリップホルダーの電極部分に
接触するようにフラットエンド側(塩基性側)を溝におとします。

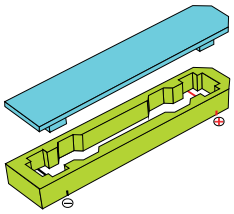


GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

④ IPGカバー液を添加します。
1 mlのスポイトにカバー液をとります。
ストリップホルダーの端からゆっくりと滴下し、Immobiline DryStrip
を半分位覆ったらもう一方の端からストリップ全体が覆われるまで
滴下します。

⑤ ストリップホルダーにカバーをします。



GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

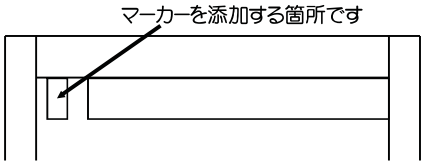
二次元電気泳動

GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

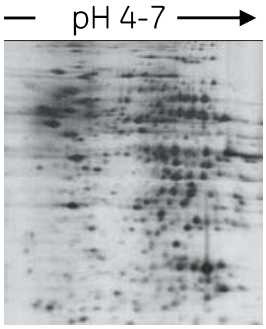
●平衡化終了後のImmobiline DryStripのセット

① SDS- PAGE用の二次元目電気泳動用プレキャストゲルの包装を開けます。



GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

二次元電気泳動

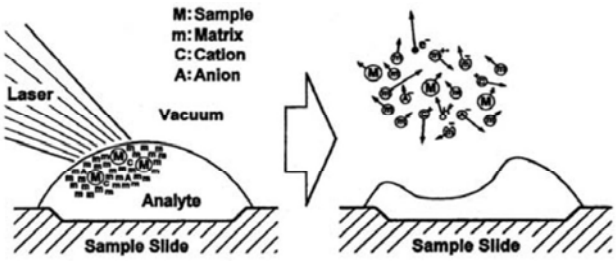


GEヘルスケア・ジャパン
ライフサイエンス統括本部
製品取扱説明書より抜粋

プロテオーム解析

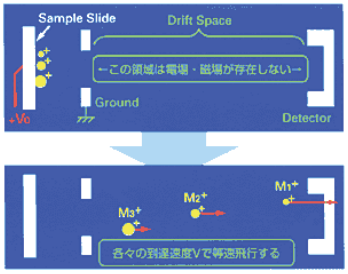
Matrix Assisted Laser Desorption/Ionization
（マトリックス支援レーザー脱離イオン化法）
Time of Flight Mass Spectrometry
(飛行時間型質量分析法)

MALDI



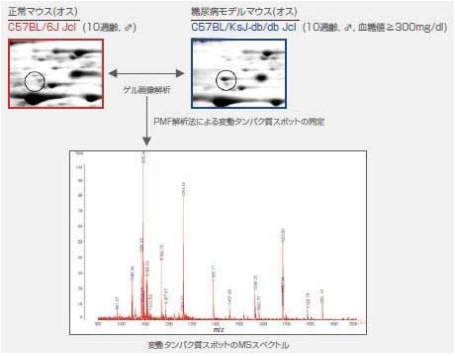
島津製作所HP
製品情報より引用

TOF　MS



島津製作所HP
製品情報より引用

プロテオーム解析



島津製作所HP
製品情報より引用

1.　活性染色

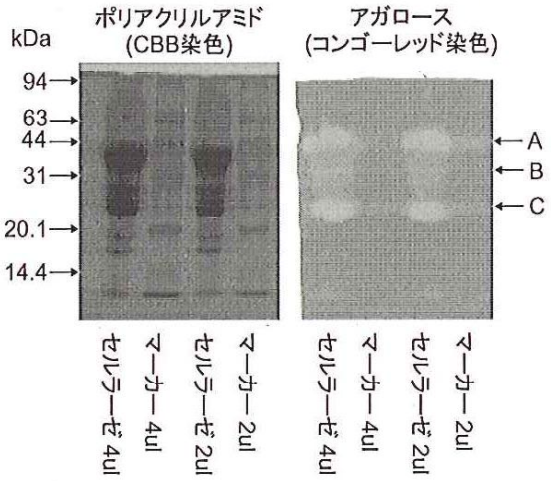
（酵素活性のゲル中復元と可視化）

1.1　目的

酵素の活性染色について実習する。

1.2　背景

タンパク質の総体を取り扱うプロテオーム解析の前段階として、目的とするタンパク質や酵素を特に精製することなく、その存在を確認する方法はいくつかある。既に抗体が用意できる場合にはウェスタンブロットが特異性も高く、良い方法である。しかし、新属新種の微生物を分離し、その分泌するタンパク質の中に自分の欲する活性を有するものがあるかどうかを見るには最適とは言えない。このような場合、培養液全タンパク質を等電点電気泳動とSDS-PAGEで二次元電気泳動に供した後、活性が見えれば好都合である。SDS-PAGEでは、還元剤と界面活性剤の作用でタンパク質分子は、ほぼ直鎖状に引き延ばされると考えられる。しかし幸いなことに、加水分解酵素は電気泳動後に界面活性剤をゆっくり除去していくことで再フォールディングされ、ゲル中で活性が復元できることがある。ここで基質を加え酵素反応を行い、反応が起こった部分だけが見えるように染色なり蛍光発光を行う。すると、活性の有無とともに等電点および分子量も明らかとなり、精製計画が立てやすくなる。もちろん全ての加水分解酵素に適用できるわけではないが、研究初期で全く手探りの段階では、活性染色は有効な手法である。特に糖質加水分解酵素の研究では活性染色がよく用いられているので、ここでは代表的な糖質加水分解酵素であるセルラーゼを用いる。



1.3　実験概要

工業用セルラーゼを試料とし、その中にあるセルラーゼ活性を分子量で分離後、酵素活性を回復し、基質と反応させて可視化する。工業用セルラーゼが単一酵素なのか、複合酵素なのかを明らかにする。

1.4　材料と方法

1.4.1　実験材料

工業用アクレモニウムセルラーゼ（明治製菓製）、100 mM DTT入りSDS化処理液、25％イソプロピルアルコール入り0.05Mリン酸緩衝液（pH 6.8）、アガロース、ラップ、0.1％CBB R-250染色液、0.1％Congo Red、1 M NaCl

1.4.2　操作

(i)　サンプル調製

工業用アクレモニウムセルラーゼ（50μg/μL）を100 mM DTT入りSDS化処理液で100℃で5分の加熱処理し、通常のSDS-PAGEを行う。サンプル添加量2μL/レーン、200V定電圧。

(ii)　脱SDS（再フォールディング）

泳動終了後、ゲルを25％イソプロピルアルコール入り0.05Mリン酸緩衝液（pH 6.8）で振盪しながら10分間洗浄する。この洗浄操作を3回行う。

(iii)　基質アガーシート調製

0.03％カルボキシメチルセルロース入り0.05Mリン酸緩衝液（pH 6.8）で1.5％となるよう寒天を溶解させ、アガーシートを作成する。

(iv)　酵素反応

洗浄済みゲル表面から水分をとり、アガーシートを密着させるように重ね、乾燥を防ぐためラップでしっかり包む。37℃で１～１６時間インキュベートし、酵素反応を行う。

(v)　染色

反応終了後、ゲルはCBB R-250で染色する。アガーシートは、0.1％Congo Redで20～45分染色し、1 M NaClで脱色する。活性部分が見えるまで30～60分である。

(vi)　脱色

脱イオン水で振盪洗浄しNaClを除く。必要に応じてスキャンする。

2.　二次元電気泳動

2.1　目的

等電点電気泳動とSDS-PAGEを組み合わせた二次元電気泳動の原理と方法について学ぶ。

2.2　実験概要

二次元電気泳動は、二段階の電気泳動によりタンパク質を二次元に展開分離する手法である。分離したタンパク質の同定法が進歩するに従い、プロテオーム解析という研究分野を拓くことにつながった。細胞の産生するタンパク質のごく一部を解析しようとする活性染色に対し、プロテオーム解析では細胞の産生する全タンパク質を解析しようとした。今回は実際の泳動は行わないが全体的な操作を行い、実験の流れを実習する。

一次元目：等電点電気泳動

タンパク質の等電点（pI）の違いを利用して分離する。タンパク質を構成しているアミノ酸側鎖やN末端、C末端の電荷はpH条件によって変化する。泳動ゲルにpH勾配を作製し、電場をかけることによって、タンパク質は固有のpIと同じpHに向かってゲル中を移動する。電荷の総和がゼロになるところでタンパク質はとどまり、バンドを形成する。電荷の総和がゼロになるところのpHの値が等電点である。

二次元目：SDS-PAGE

目的タンパク質の高次構造を変性して分子量の違いにより分離する手法。等電点電気泳動は2つの手法がある。

(i) キャリアアンフォライトを用いた方法、両性担体(キャリアアンフォライト)をゲルに添加し
て電場をかけてpH勾配を形成。等電点の差がpH 0.01~0.02で分離可能。

(ii) IPG（Immobilized pH radient）法、様々なpIの側鎖を持つアクリルアミド誘導体を用いて
ゲル作製と同時にpH勾配を形成する。 等電点の差がpH 0.001の違いでも分離可能。

2.3　材料と方法

2.3.1　一次元目の泳動（等電点電気泳動）

2.3.1a　実験材料

Immobiline DryStrip（7cm, 3-11NL）、カバー液（Immobiline DryStrip Cover Fluid）、
ゲル膨潤原液（8M Urea, 2%CHAPS, 0.5%IPG Buffer, 0.002%BPB, 7mg/2.5mL DTT）、
一次元電気泳動装置：Ettan IPGphor3、Strip Holder
泳動試料（工業用アクレモニウムセルラーゼ、先染めの場合は蛍光色素と反応させておく）

2.3.1b　操作

(i)　ゲル膨潤液の調製

ゲル膨潤原液と泳動試料を混合し、ゲル膨潤液を調製する。

サンプル添加量は、DryStripのサイズによって異なり、今回使用する7cm DryStripでCBB染色を行う場合の推奨サンプル量は40～240μgである。

7cmDryStripを完全に膨潤させるのに必要な液量は125μLである。

調製しやすいように泳動試料の調製は、2000μg/1mL膨潤液とした。

今回用いた酵素は未精製であり、タンパク質と糖が7：3で混在している。そこで、実際のタンパク質濃度と添加量は、2000μg × 0.7 ＝ 1,400μgタンパク質 / 1mL膨潤液、175μgタンパク質 / 125μL膨潤液となる。

(ii) Immobiline DryStripへのサンプルアプライ

ピペットでゲル膨潤液125μL をStrip Holderにまんべんなく酸性側から中央にかけてゆっくり入れる。

DryStripの保護カバーを外し、ゲルの面を下に向けてStrip Holderに入れる。この際、ゲル全体に膨潤液が広がるように注意する。

このとき、ゲルの酸性側（＋）をStrip Holderのエンドポイント側（先が尖った側）に向け、エンドポイント側からDryStripをゆっくりなじませながら置いていく。

Holderの底とゲルの間に気泡が入っていないか確認し、ピペットでカバー液（Immobiline DryStrip Cover Fluid）1 mL を、Strip Holderに端から、ストリップの全体が覆われるまで滴下する。Strip Holderの蓋をする。

(iii) 等電点電気泳動の開始

Ettan IPGphor3を用いて、以下の条件で等電点電気泳動を行う。

表1. IPGphor3の泳動条件

ステップ	電圧変化パターン	電圧(V)	時間(Hr)	kVhr
1	Step and Hold	300	0:30	0.2
2	Gradient	1000	0:30	0.3
3	Gradient	5000	1:20	4.0
4	Step and Hold	5000	0.06-0.25	0.5-2.0
Total			2:26-2:45	5.0-6.5

2.3.2　二次元目の泳動（SDS-PAGE）

2.3.2a　実験材料

・平衡化buffer A

（50 mM Tris-HCl buffer(pH 8.8), 6M Urea, 30%グリセロール, 1% SDS, 0.25% DTT）

・平衡化buffer B

（50 mM Tris-HCl buffer(pH8.8), 6M Urea, 30%グリセロール, 1% SDS,
4.5%ヨードアセトアミド）

・0.5%アガロースゲル

・二次元目泳動装置（Ettan Dalt six, Hoefer SE 600 Rubyなど）

2.3.2b　操作

Immobiline DryStripをゲルにセットする前に0.5%アガロース（0.05g/10mL）を調製した。

(i)　一次元目の泳動Immobiline DryStripの平衡化

Immobiline DryStripを二次元目の分離に必要なSDS buffer系に平衡化する。

DryStripのカバー液を、ピペットで除去する。

試験管に500μLの平衡化buffer Aを入れ、1次元目泳動終了後のDryStripのゲル面全体がbufferに浸かるように入れ、15分間振盪する。

別の試験管に500μLの平衡化buffer Bを入れ、DryStripを移し、15分間振盪する。

(ii)　平衡化終了後のImmobiline DryStripをSDS-PAGE用ゲルの上部に乗せる

平衡化終了後のDryStripのプラスチックフィルムの部分をピンセットで持てるぐらい残して、はさみで切る。

SDS 泳動bufferでDryStripを軽くすすぐ。

DryStripのプラスチックフィルムを一方のガラスプレートにつけ、スパーテルを用いて隙間や空気が入らないようにDryStripをゆっくり押し込む。

少量の0.5%アガロースをSDS-PAGE用ゲルに添加し全体を封入する。

（マーカーレーンに分子量マーカーをアプライする）

(iii)　二次元目の泳動（SDS-PAGE）

泳動装置にゲルをセットし、装置内を泳動bufferで満たす。

冷却装置を動かし、電気泳動を開始する。

泳動条件の例：600V, 40mA, 50W, 2hr(中型ゲル2枚分の目安)

(iv)　泳動後ゲルを取り出し、解析する。（後染めの場合は染色・脱色後に解析）

（解析については単にソフトウェアの説明になるので割愛・概要を説明する）

3.　ゲル抽出と質量分析

3.1　目的

二次元電気泳動で得られたタンパク質スポット中の微量タンパク質を、質量分析で解析し同定する手法を実習する。

3.2　実験概要

二次元電気泳動後のゲル中にあるタンパク質スポットを切り出し、ゲル内にタンパク質を留めたまま、プロテアーゼで消化する。生じたペプチド断片を質量分析し、その質量値セットをデータベース検索にかけ、もとのタンパク質を同定する。

3.3　材料と方法

3.3.1　ゲル内消化とペプチド断片の抽出

3.3.1a　実験材料

メス、ピンセット、マイクロチューブ、減圧濃縮遠心エバポレーター、トリプシン溶液、
ペプチド抽出液

3.3.1b　操作

＊人体からのケラチン混入を避けるため、以下の操作は清浄な環境で行うのが望ましい。

(i)　ゲル片の切り出し

カッティングシートなどの上にラップを敷き、メスでタンパク質スポットを切り出す。
なお、切り出した切片はラップ上の別位置でさらに細かく1mm角に刻んでおく。

(ii)　脱色

必要であれば、脱色液でタンパク質を染めている染色剤を取り除く。マイクロチューブに刻んだゲルを入れ、適切な脱色液を加えてよく振盪して脱色する。（2〜3分を2〜3回。先染めの場合はこの操作は不要である。）

(iii)　脱水

マイクロチューブに100μLのアセトニトリルを加え、10分間振盪する。ゲルが小さく収縮したらアセトニトリルを取り除き、 減圧濃縮遠心エバポレーターで10分間ほど乾燥する。

(iv)　トリプシン消化

乾燥後のマイクロチューブにトリプシン溶液を20μL加えて、4℃に冷却しながら30分間静置する。ゲルが再膨潤したら余分なトリプシン溶液を除く。37℃のインキュベーターで一晩置き、反応させる。

(v)　ペプチド断片抽出

マイクロチューブに50μLのペプチド抽出液を加え、室温で30分間振盪する。遠心機でフラッシュし、抽出液を新しいマイクロチューブに移す。一度では全て抽出できないので、抽出操作を2回繰り返す。その都度、抽出液は一本の同じマイクロチューブに回収する。 減圧濃縮遠心エバポレーターで10μL以下に濃縮する。（時間がかかる）

3.3.2　MALDI-TOF/MSによるペプチド断片の質量分析

3.3.2a　実験材料

ZipTip$_{C18}$ピペットチップ、アセトニトリル、トリフルオロ酢酸、
ペプチド用マトリックス（CHCA）

3.3.2b　操作

(i)ZipTipの平衡化

0.1％TFA入り50％アセトニトリルを10μL、ZipTipで吸引し廃液用マイクロチューブに捨てる。この操作を3回繰り返す。続いて、0.1％TFAを10μL吸引し、廃液用マイクロチューブに捨てる。この操作を3回繰り返す。

(ii)ペプチド断片を吸着・脱塩

ペプチド断片抽出液を平衡化したZipTipでゆっくりと吸引し、同じマイクロチューブ内にゆっくりと排出する。この操作を５回繰り返し担体にペプチドを吸着させる。

0.1％TFAを吸引し、 廃液用マイクロチューブに捨てる。この操作を3回繰り返し洗浄する。

(iii)ペプチドの溶出、マトリックスとの共結晶形成

ペプチドを吸着させているZipTipでマトリックス飽和溶液を2μL吸引し、質量分析計のターゲットプレート上で、2〜3回ピペッティングし溶出を行う。そのまま溶出液を吸引し、ターゲットプレート上にドーム上の液滴を形成するように滴下する。自然風乾し、マトリックスとペプチドが共結晶形成するのを待つ。

(iv)質量分析計にターゲットプレートを挿入し、質量分析を行う。

9月7日　（於　沖縄工業高等専門学校）

平成２４年度沖縄県産業振興基金事業「バイオインフォマティクス人材育成推進事業」
公益財団法人沖縄科学技術振興センター

琉球大学　夏期集中講義
～バイオインフォマティクス・スタンダードコース～
食品機能科学特別講義Ⅰ（農学部）
第5部　生命科学実習④　プロテオーム解析

課題レポート

所属＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿氏名＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

1.　活性染色実験の流れをまとめなさい。

2.　二次元電気泳動の手順をまとめなさい。

3.　プロテオーム解析の流れをまとめなさい。

3.6　情報科学基礎

3.6.1 概要

講師：名嘉村 盛和／琉球大学工学部情報工学科教授

　　岡﨑 威生／琉球大学工学部情報工学科講師

日時：2012 年 9 月 10 日　8:30～／10:20～／12:50～／14:40 ～／16:30 ～

場所：琉球大学総合情報処理センター

内容：情報科学分野の基礎として、生命科学分野の研究者／技術者が修得すべきデータ構造とアルゴリズム、最適化計算の基本を解説と演習を通して学ぶ。

3.6.2 準備

特筆事項なし

3.6.3 配布資料

・講義スライド

3.6.4 課題

・プリント（演習問題（1）～（5））

3.6.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
<u>分子生物学データベース</u> 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
<u>配列解析</u> アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
<u>タンパク質立体構造解析</u> 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
<u>遺伝学・進化学</u> 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
<u>トランスクリプトーム解析・プロテオーム解析</u> 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
<u>パスウェイ解析・システム生物学</u> ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.6.6 時間割

時間	内容
8:33～ 8:46～	＊岡﨑先生 イントロダクション ・ バイオインフォマティクス人材育成推進事業について 2 カ年の育成プログラム スタンダード、アドバンスト、プロフェッショナル ・ バイオインフォマティクス技術者認定試験について ・ 琉球大学講義内での位置づけ紹介 情報科学とは ・ 計算機の発展により、急激に加速した分野 収集、発生、分類、操作、蓄積、…etc ・ 計算機の普及について トランジスタの発達、集積回路の進化 ・ 情報科学と生物学との関わり ゲノム、トランスクリプトーム、メタボローム…etc ・ バイオインフォマティクスの果たす役割 ・ バイオインフォマティクス人材の果たす役割
	休憩
10:20～ 10:36～	＊名嘉村先生 コンピューターの基礎 ・ コンピューターと 2 進数 ・ 2 進数と 10 進数 ・ 情報を数字で表現・処理する ・ 論理回路（AND、OR、NOT） ・ 文字コード（ASCII、JIS） ・ プログラム、アセンブラ言語 コンピューターの仕組み ・ コンピューターの構成要素 ・ コンピューターの基本動作 CPU＝制御装置、メモリ＝主記憶装置 ＊ 時間の都合上、クロックパルス、キャッシュ、補助記憶装置については資料配布のみ ・ 計算機の進化について

	コンピューターは 18 ヶ月、ネットワークは 9 ヶ月で倍になる
10:46～	アルゴリズムの基礎 ・ プログラミングとアルゴリズム ・ 擬似言語とフローチャート表現 代入文、ブロック、if 文、for 文、while 文、…etc ・ アルゴリズムと時間計算量 n 次多項式計算、初等探索法（線形探索法)、二分探索法
11:10～	・ データの工夫 2 分探索木、部分和と総当り解法
11:24～	・ 解き難い問題について 部分和問題、総当り解法 ・ アルゴリズムの評価基準、計算複雑度
11:40～	データ構造の基礎 ・ 配列 ・ リスト 基本操作、探索、合併、切り取り、 ・ スタックとキュー ・ グラフ／有向グラフ 隣接リストによるグラフの実現
11:50～	昼食
12:50～	ヒープ（順序付きキュー） ・ 基本操作 insert、findmax、deletemax、decreasekey など
13:20～	・ 基本操作の時間計算量 ・ ヒープの構成方法
13:25～	探索 ・ 2 分探索木と対象順 ・ 2 分探索木における基本操作 挿入、削除、回転
13:37～	・ ハッシュ法とその事例
13:51～	整列（ソーティング） ・ バケットソートと基数ソーティング

	・ 選択法 ・ バブルソート ・ 挿入法 ・ ヒープソート ・ マージソート
14:40～	・ クイックソート
	休憩
14:48～	グラフ・アルゴリズム
	・ グラフ ・ 有向グラフ ・ グラフのコンピューター内での表現方法
14:55～	・ 最小木問題
15:25～	・ 最短経路問題 ダイクストラ法
15:40～	・ オイラー閉路問題 パス、閉路、連結グラフと非連結グラフ
15:44～	・ オイラーパス問題 ・ 巡回セールスマン問題 ・ ハミルトニアンパス問題（時間の都合で省略）
15:52～	・ ゲノムアセンブリング ＊残りの時間は課題作成に取り組んだ
	休憩
	＊岡﨑先生
16:30～	イントロダクション
	・ 似ているってどういうこと？ ・ 確立／統計の役割 ・ スケジュール
16:37～	計算環境
	・ ハードウェア ・ ソフトウェア ・ データ記述 ・ 遺伝子発現データ ・ データの代表値 ・ データのばらつき

16:59～	【R 演習】 ・ 発現解析データ（CSV ファイル）を読み込む ・ データ表示

情報工学科　特別講義V

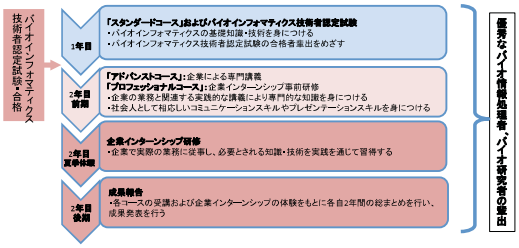
2012年9月10日（月）
琉球大学工学部

バイオインフォマティクス人材育成推進事業

平成22～24年度　沖縄県産業振興基金事業
- 主体：公益財団法人沖縄科学技術振興センター
- 後援：琉球大学、沖縄工業高等専門学校
- 運営：株式会社リバネス

事業目的：バイオインフォマティクス人材育成基盤の構築
事業実績：
- **フォーラム等イベントの開催**
- **バイオインフォマティクス人材育成講座の開講**
- **日本バイオインフォマティクス学会・沖縄地域部会の設立**
- **バイオインフォマティクス技術者認定試験の県内誘致**

2か年の人材育成プログラム



1年目：「スタンダードコース」
バイオインフォマティクスの基礎的な知識・技術を学ぶ座学と実習

2年目：「アドバンストコース」
企業で通用する実用的な専門知識を学ぶ講義とインターンシップ研修

「プロフェッショナルコース」
知識・技術を仕事の中で有効に活用できる社会人（プロ）の育成を目指す

バイオインフォマティクス技術者認定試験

琉球大学　夏季集中講義

- 農学部食品機能科学特別講義Ⅰ
- 工学部情報工学科特別講義Ⅴ

バイオインフォマティクス人材育成講座
「スタンダードコース」をもとに構築

食品機能科学特別講義Ⅰ

第1部
バイオインフォマティクスの歴史、
遺伝医学講座

第2部　生命科学実習①･②
ゲノム
DNA抽出から配列解析

第3部
トランスクリプトーム
NGSやマイクロアレイのデータの解析

第4部
プロテオーム
二次元電気泳動、質量分析、タンパク質同定

情報工学科　特別講義Ⅴ

第1部
情報科学基礎

第2部
類似性発見

第3部
配列と機能

第4部
タンパク質の機能と構造

情報科学基礎

情報科学とは？

情報

収集　発生　分類　操作　蓄積　伝達　復旧　処理　…etc.

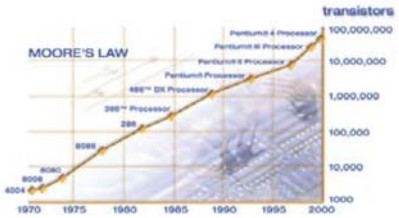
e.g.
収集：サーチエンジン
発生：データマイニング
蓄積：データベース
伝達：コンピュータネットワーク
処理：計算機科学、ソフトウェア

情報の相互作用の研究や
情報の利活用
（問題の解法そのものに興味）

☞ **計算機の発展により、急激に加速した分野**

計算機の普及

電子計算機
真空管方式→トランジスタ
集積回路（IC：Integrated　Circuit）
※集積度：1つのICチップに含まれるトランジスタの数）／ムーアの法則



☞ **計算機の小型化、低価格化、高速化、高機能化、大容量化**

情報科学と生物情報

1953　DNA二重らせん
1970～1980
配列データベース
ホモロジー検索　などの基礎

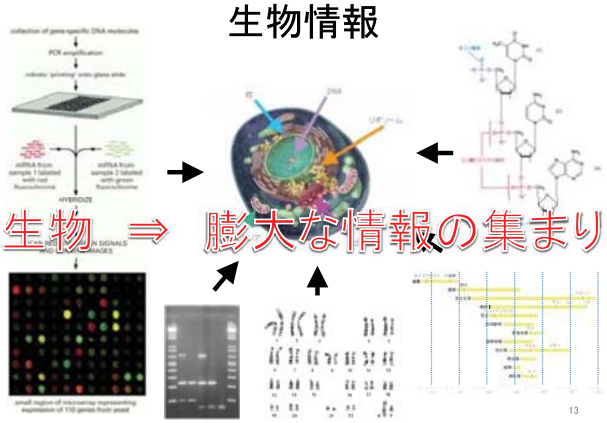
1990年代
「ゲノムインフォマティクス」
↓
1993　文献に3報
「バイオインフォマティクス」
- ゲノム
- 構造ゲノムプロジェクト
- トランスクリプトーム
- プロテオーム
- メタボローム
- システム生物学
- ケモインフォマティクス　…etc.

1990～2003
ヒトゲノムプロジェクト

日本
1991～1995
ゲノム解析に伴う大量知識情報の研究（重点領域研究）
1996～2000
ゲノムの生物知識情報（特定領域研究）
2000～2004
ゲノム情報科学（特定領域研究C）

☞ **塩基配列やアミノ酸配列の情報など大量のデータの管理、解析に**
コンピュータが使われ、バイオインフォマティクスの基盤となった。

生物情報



バイオインフォマティクス

生命科学

- ゲノム情報、配列比較
- 遺伝子発現情報
- 遺伝子多型情報
- タンパク質、RNA構造情報、構造予測
- システム生物学

…etc.

情報の蓄積および情報の活用
バイオインフォマティクスの重要性

必要不可欠 →

← 応用の1分野

情報科学

- 数学
- 社会学
- 心理学
- 経済学
- 生命科学

…etc.

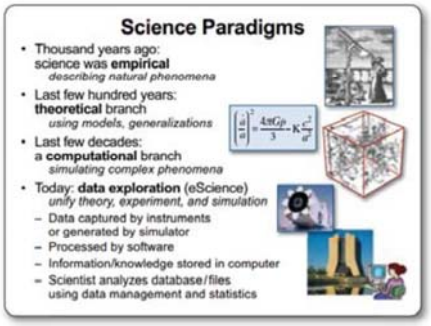
興味対象の1分野
さまざまな情報との融合
データマイニング

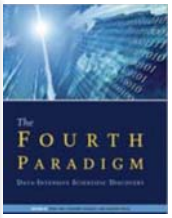
バイオインフォマティクス

⇒ データが大きければ大きいほど、
情報が蓄積されればされるほど、発展する分野

データの蓄積と利活用

http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf





☞ これからの時代、
膨大に蓄積された情報から新たな科学的発見が生まれる

バイオインフォマティクス人材

大量の実験データの解析、
蓄積された膨大な情報を利活用できる人材が不足

☞ **スキルのある人材は引く手あまた**

日本バイオインフォマティクス学会
「バイオインフォマティクス技術者認定試験」

試験日： 11月25日（日） 13:30-15:30

申込： 9月3日（月）～10月26日（金）

※試験対策講座 （無料）
11月23日（金）、24日（土） 於 沖縄高専

アルゴリズムの基礎

プログラミングとアルゴリズム

プログラムの作成と実行のステップ

1. 何を処理させるのか，問題を明確にする.
2. 処理手順（アルゴリズム）を決める.
3. 適当なプログラミング言語によりプログラムを作る.
4. 作成したプログラムを実行する.
5. コンピュータの出力を吟味する.

アルゴリズムの設計と解析の要点

1. 対象とする問題の解きやすさを予め検討する.
2. 種々のアルゴリズムを考え，その中で最も良いものを採用する.
3. 適当なデータ構造を工夫する.

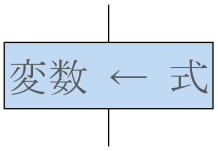
アルゴリズムの記法

- 自然言語表現
  - 日本語などで簡潔に表現したもの
- 図形表現
  - フローチャート（Flow Chart ： 流れ図）
  - UML （シーケンス図）
- 擬似言語表現
  - C, Pascal などの高級言語の制御構造と自然言語を組み合わせた表現方法

擬似言語とフローチャート表現(1)

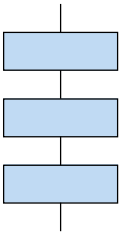
- 代入文

  変数 ← 式 ；



- ブロック

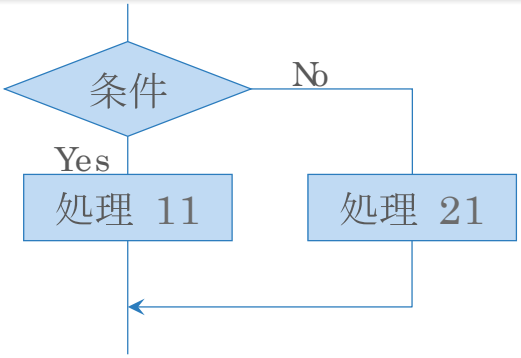
  “{”， “}” で囲まれた部分



擬似言語とフローチャート表現(2)

- if 文

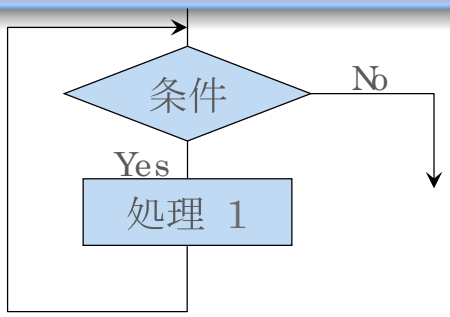
```
if (条件) {
    処理 11 ;
    処理 12 ;
        :
}
else {
    処理 21 ;
    処理 22 ;
        :
}
```



擬似言語とフローチャート表現(3)

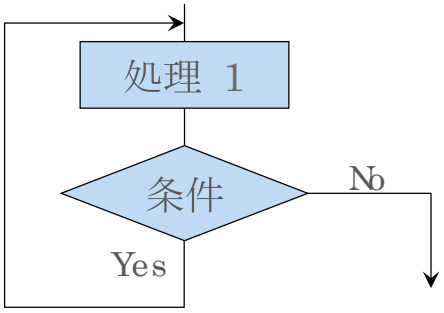
- while 文 (前判定)

```
while (条件) {
    処理 1 ;
    処理 2 ;
        :
}
```



- do - while 文 (後判定)

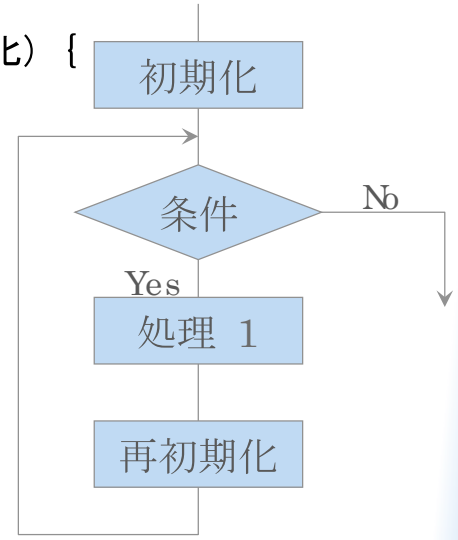
```
do {
    処理 1 ;
    処理 2 ;
        :
} while (条件);
```



擬似言語とフローチャート表現(4)

- for 文

```
for（初期化 ; 条件 ; 再初期化）{
    処理 1 ;
    処理 2 ;
        :
}
```



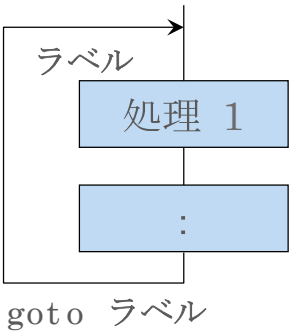
擬似言語とフローチャート表現(5)

- goto 文

```
goto ラベル ;
```

- 関数と return 文

```
関数名（仮引数の並び）{
    処理 1 ;
    処理 2 ;
        :
    処理 3 ;
    return（式）;
}
```



アルゴリズムと時間計算量

n次多項式計算

処理すべき問題の定式化（Step 1）

- n次多項式の計算を次の通り定式化する.

$$f(x)=a_n x^n+a_{n-1}x^{n-1}+\cdots+a_1 x+a_0$$

処理手順（アルゴリズム）の決定（Step2）

(1) 初等的計算法

$$f(x)=a_n*x*\cdots*x+a_{n-1}*x*\cdots*x+\cdots+a_1*x+a_0$$

乗算 n(n+1)/2, 加算 n, 変数 n+1

(2) ホーナー法(Horner's Rule)

$$f(x)=(((\cdots(a_n*x+a_{n-1})*x+a_{n-2})*x+\cdots+a_1)*x+a_0$$

乗算 n, 加算 n, 変数 1

アルゴリズムと時間計算量

探索問題(1)

探索問題

- 10個の整数の集合S={18, 9, 22, 4, 21, 12, 15, 31, 7, 25}

  member(x)： 集合Sの中に要素xがあるか否かを調べ, xがあればYesを, なければNoを出力する.

アルゴリズムと時間計算量

探索問題(2)

初等的探索法（線形探索法）

初期データが一次元配列に蓄えられているものとする.

18	9	22	4	21	12	15	31	7	25

```
int member(int x){
      for(i=0; i<=DATASIZE; i++)
        if(R[i]=x) return ("Yes");
      return("No");
    }
```

- 基本操作の回数
  - member(4)　　4回（左→右）　7回（右→左）
  - member(15)　　7回（左→右）　4回（右→左）

アルゴリズムと時間計算量

探索問題(3)

二分探索法（バイナリーサーチ）

初期データが一次元配列に蓄えられているものとする.

ただし、あらかじめ昇順、あるいは降順に並べてある.

```
int member(int x){
  low = 0; high = n-1;
  while (low <= high){
    mid = (low+high)/2;
    if(R[mid]==x) return ( "Yes" );
    else if (key < R[mid]) high = mid - 1;
    else low = mid + 1;
  }
  return("No");
}
```

4	7	9	12	15	18	21	22	25	31

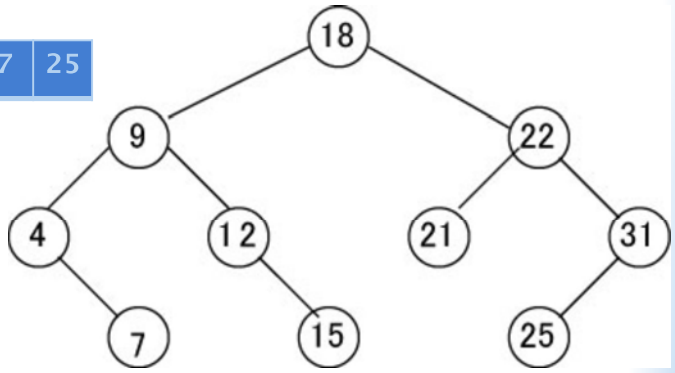
基本操作の回数は？

データ構造の工夫(1)

2分探索木：

- 各点の値は左の子の子孫のどの値より大きく，右の子の子孫のどの値より小さい
- 対称順(symmetric order or inorder)．

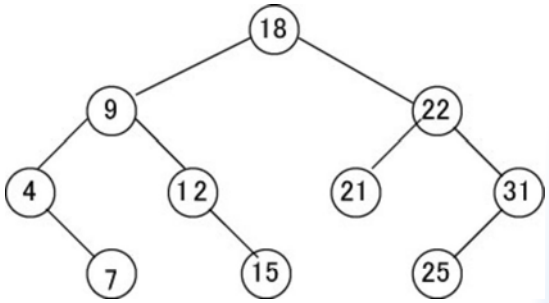
18	9	22	4	21	12	15	31	7	25



データ構造の工夫(3)

- 2分探索木に基づく探索法

```
int member(int x){
      v = root;
      while(1){
        if(vの値 == x)
          return ("Yes");
        else if(vには子がいない)
          return ("No");
        else if(vの値 < x)
          v = vの右の子;
        else
          v = vの左の子;
      }
}
```



- 基本操作の回数
  - member(4)　3回，member(15)　4回
- しかし，２分探索木のバランスが悪いと→初等的探索と同じ計算量 O(n)．

解き難い問題

部分和問題と総当たり解法

- 部分和問題
  - 正整数の集合Sと１個の正整数bが与えられたとき，Sの要素を適当に選んでその和がbに等しくできるか否かを判定せよ．
- 総当たり解法
  - Sのすべての部分集合についてその中の要素の和がbに等しいか調べる．
- Sの部分和の数は，2の|S|乗
  - 1つの部分集合の判定に百万分の1秒必要だと仮定すると
  - n=10のとき：0.001秒，n=30のとき：1000秒
  - n=100のとき：1000000000000000000000000秒（>46億年×540万）
- 入力サイズnがほんの少し大きくなっただけで総当たり解法は現実的ではなくなる．

アルゴリズムの評価基準と計算複雑度

アルゴリズムの計算時間

- O(f(n))
  - 入力サイズnが十分大きくなると最悪計算量はf(n)に比例する時間以下である．例：$O(n^2)$，$O(2^n)$，O(log n)
  - 漸近的上界
- Ω(f(n))
  - nを十分大きくとると，T(n)はf(n)に比例する計算時間以上である．例：$\Omega(n^2)$，$\Omega(2^n)$，Ω(log n)
  - 漸近的下界
- Θ(f(n))
  - T(n)は，最も早く実行できてもf(n)に比例した時間がかかり，逆に最悪計算時間もf(n)に比例する．

アルゴリズムの評価基準と計算複雑度

計算時間による問題の分類

- 解き易い問題
  - 多項式時間アルゴリズムが存在する問題
  - クラスPに属する問題
- 解き難い問題
  - $\Omega(2^n)$, $\Omega(n^n)$, $\Omega(n!)$
- 不明，ただし，解き難いと思われている.
  - NP-完全問題

コンピュータの基礎

２進数

コンピュータの仕組み

２進数

コンピュータと2進数

- コンピュータの内部は電子回路
  - 電子が流れているかいないか
  - 電圧が高いか低いか
  - 電荷がたまっているかいないか
- このような物理現象で情報処理？
  - 情報（データ）をどのように表しているか？
  - どのように処理しているか？
- 秘密は2進数：0と1の世界
  - 電圧が低い->0 高い->1
  - 数字も，画像も，音声も0と1だけで表現
  - CD，デジタルカメラ等々

2進数の単位

- 2進数1桁を1ビット
- 8ビットを1バイト
- 大きな数字と小さな数字の表現
  - 10^3をキロ(K)，10^6をメガ(M)，10^9をギガ(G)
  - 10^{-3}をミリ(m)，10^{-6}をマイクロ，10^{-9}ナノ(n)
- 例：
  - ハードディスクの容量は60Gバイト
  - メモリサイズは512Kバイト
  - CPUの動作周波数は2.0GHz（ギガヘルツ）

2進数と10進数(1)

- 2進数で数字(10進数)を表そう
  - 0と1の二種類の記号を使う
  - 0, 10, 11, 100, 101, …
  - 3桁で2^3 (=8)通りの数字, k桁で2^k通りとなる
  - 10進数より桁数が長くなる
  - 2進数と10進数を当てはめてみると…

2進数と10進数(2)

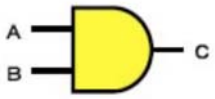
10進数	2進数
0	0
1	1
2	10
3	11
4	100
..	..
9	1001
10	1010
..	..
100	1100100

情報を数字で表現・処理

- シャノンの情報理論
  - 1916- 2001:情報理論の創始者
  - 情報に関する数学を確立
  - 1937年「継電器とスイッチ回路の記号論的解析」
    - スイッチ回路のオン，オフで記号論の真偽を表す
    - スイッチの直列でAND(かつ)，並列でOR(または)
    - ANDとORとNOT(反転)で全ての論理演算を表現できる
    - 全ての論理演算をコンピュータで処理可能となる
- 種々の情報を数字で表現(コード化)
  - あらゆる処理を論理演算として処理可能
- コンピュータは基本的な回路で構成される
  - 基本回路を組み合せる事で様々な処理が可能

論理回路(AND, OR, NOT)

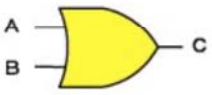
AND



- AND＝かつ
  - Aが正しい
    かつ
  - Bが正しい

A	0	0	1	1
B	0	1	0	1
C	0	0	0	1

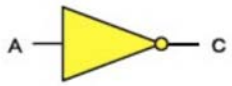
OR



- OR＝または
  - Aが正しい
    もしくは
  - Bが正しい

A	0	0	1	1
B	0	1	0	1
C	0	1	1	1

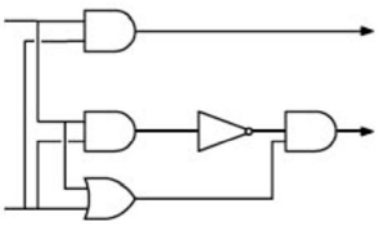
NOT



- 0を1に
- 1を0に

A	1	0
C	0	1

1ビット加算器



文字の表現

- 文字も数字で表現される
- 文字と2進数の対応表があれば良い
- 2進数7桁あれば，128 (=27)種類の文字が表現可能
- 例えば‘A’は，0100001と対応づけられている
  - キーボードから’A’が入力されると実際には0100001を表す信号がコンピュータ入力される
  - ‘B’は，0100010に対応付けられている，.....
- 大切なのは，同じ対応表を使う事
  - 異なる対応表だとコミュニケーションができない
  - 文字化けの原因

文字コード(ASCII, JIS)

= 32	! = 33	" = 34	# = 35	$ = 36	% = 37	& = 38
' = 39	(= 40	) = 41	* = 42	+ = 43	, = 44	- = 45
. = 46	/ = 47	0 = 48	1 = 49	2 = 50	3 = 51	4 = 52
5 = 53	6 = 54	7 = 55	8 = 56	9 = 57	: = 58	; = 59
< = 60	= = 61	> = 62	? = 63	@ = 64	A = 65	B = 66
C = 67	D = 68	E = 69	F = 70	G = 71	H = 72	I = 73
J = 74	K = 75	L = 76	M = 77	N = 78	O = 79	P = 80
Q = 81	R = 82	S = 83	T = 84	U = 85	V = 86	W = 87
X = 88	Y = 89	Z = 90	[= 91	\ = 92	] = 93	^ = 94
_ = 95	` = 96	a = 97	b = 98	c = 99	d = 100	e = 101
f = 102	g = 103	h = 104	i = 105	j = 106	k = 107	l = 108
m = 109	n = 110	o = 111	p = 112	q = 113	r = 114	s = 115
t = 116	u = 117	v = 118	w = 119	x = 120	y = 121	z = 122

実際には2進数で表されている.

他のデータは？

- 日本語の漢字は通常16bitで表現される
  - 2^{16}(=65536)種類の組合せが表現可能
- 画像(静止画)は，色のついた点の集まり
  - 光の色は，光の三原色の重ね合わせ
  - 赤，緑，青のそれぞれの強さを2進数で表す
- 動画，音声も二進数で表現できる
- プログラムも2進数で表される
  - プログラムは命令の集まり
  - 命令は2進数で表現される

プログラム

C言語	アセンブラ言語	機械語
main(){	LEA r1, 0	00110000
x=0;	STR r1, X	01111001
y=5;	LEA r1, 5	00110101
z=x+y;	STR r1, Y	01111010
print(z);	ADD r1, X	10011001
}	STR r1, Z	01111011
	OUT r1	11010000
	HLT	11100000

プログラム

- 機械語によるプログラム開発は人間にとって分かりづらい
- より人間に分かり易いアセンブラ言語が発明された
  - アセンブラ言語の命令を機械語に変換
  - 大規模なプログラムはやはり分かりづらい
- 人間の言葉に近い高級言語が開発される
  - Java, C, Fortran, BASIC等

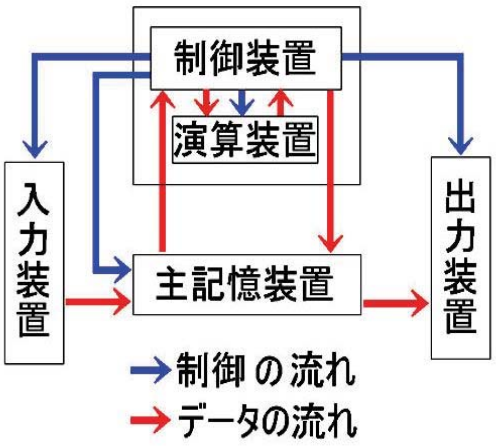
コンピュータの仕組み

コンピュータの構成要素

- CPU(中央処理装置)
  - プログラムを構成する命令を解読し，各構成要素に適切な指令を送る．コンピュータの中枢部
- 演算装置
  - いろんな計算をするための回路
- メモリ
  - CPU，演算装置が使うデータやプログラムを置く場所
- 外部記憶装置(ハードディスク)
  - データやプログラムをファイルとして保存する．電源を切っても記録は消えない
- 入出力装置
  - キーボード，マウス，ディスプレイ，プリンタ

コンピュータの基本動作

1. ハードディスクから実行したいプログラムをメモリにコピー(ロード)する
2. プログラムの先頭の命令をCPUが受取り，どのような命令か解読する
3. 命令の種類に応じて，周辺の構成要素に適切な指令信号を送る
4. 周辺要素(演算装置，メモリ，入出力装置等)は指令に従って動作する
5. CPUはプログラムの次の命令を取り出し解読する．3に進む



クロックパルス

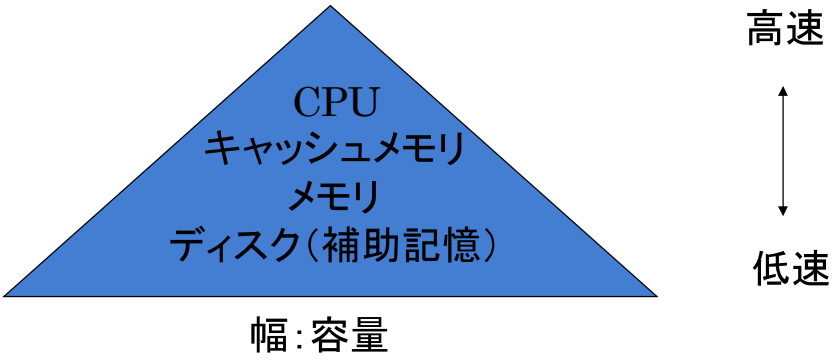
- コンピュータは，複数の回路・装置が協力しながら処理を進めている
  - クロックパルス信号が全ての回路・装置に同じタイミングで入力される.
  - 複数の装置のタイミングはクロックパルスに合わせる.
- クロックパルスの間隔が短いと高速処理が可能である.
  - クロックパルスの間隔は，CPUの動作周波数として表され，コンピュータの処理速度を決める最も重要な尺度の一つである.
  - 1GHz →1秒間に10億回のクロックパルスがでる.

なぜ，メモリとディスクが必要か？

- メモリはデータの取り出し・書込みが高速. しかし容量（蓄えられる情報の量）は小さい
  - 512kバイト，数十ナノ秒
- ディスクはデータの取り出し・書込みが低速. しかし容量は大きい.
  - 200Gバイト，数ミリ秒
- メモリは電源が切れると消される. ディスクは電源がきれても磁気なので残る.
- CPUは高速に動作する必要があるので，データをやりとりする相手も高速な方が良い.
  - 実はメモリでも遅すぎるので，高速なメモリ（キャッシュ）というのが，メモリとCPUの間に存在する.

キャッシュの役割

- CPUとメモリ間の速度差を吸収する.
  - CPUとメモリの速度は，数十倍から百倍
  - キャッシュメモリは，小容量・高速なメモリ



補助記憶装置

- フロッピーディスク
  - 磁気，720KB～1.44MB
- ハードディスク
  - 磁気，10GB～300GB～，パソコン標準内蔵，高速
- CD-ROM (CD-R, CR-RW)
  - 光，640MB～800MB
- DVD-ROM(DVD-RAM, DVD-R, DVD-RW)
  - 4.7GB～9.4GB
- SD, CF, コンパクトフラッシュ
  - フラッシュメモリ，16MB～1GB～

その他の周辺機器

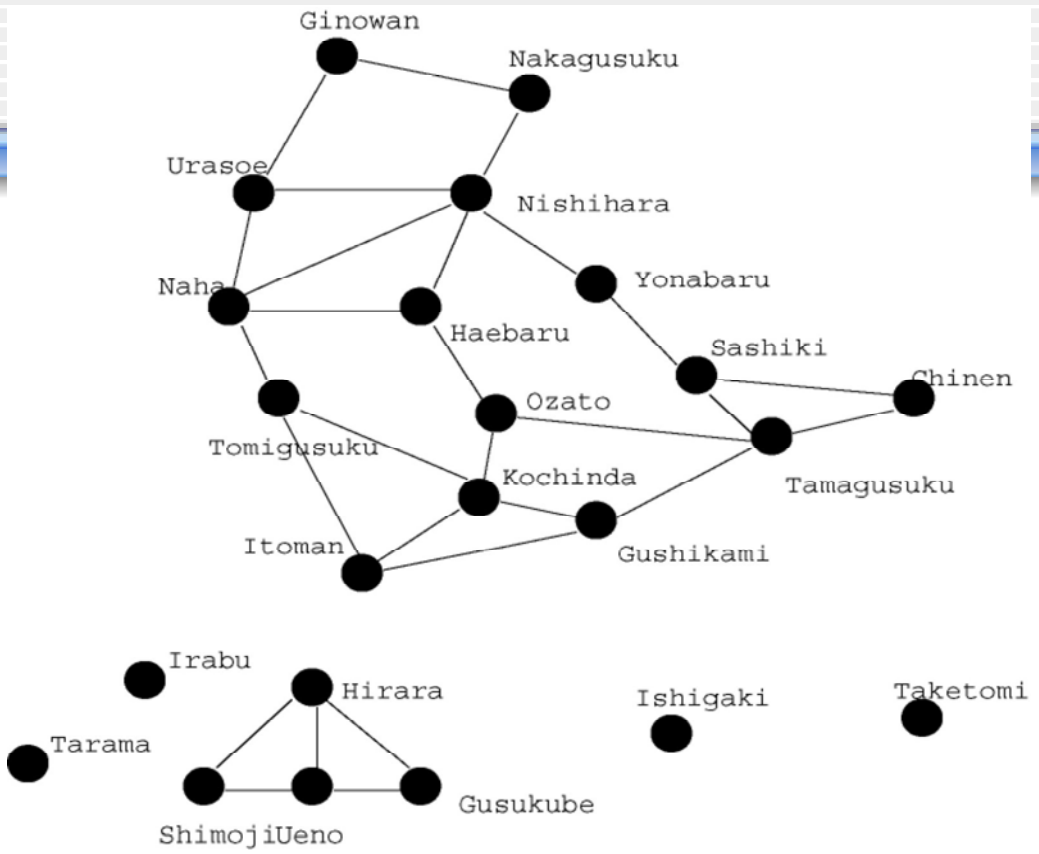
- スキャナ（イメージスキャナ）
  - 紙媒体に印刷された図形，写真，文字等を画像データとしてコンピュータに取り込む．
- デジタルカメラ
  - 光に反応する半導体(CCD)を使って露光し，画像データとして取り込む．

グラフ・アルゴリズム

グラフ

グラフとは

- グラフ(Graph)
  - 点集合Vと辺集合Eの組,G=(V, E).
  - 有向グラフ(directed graph):辺(edge)に向きがある場合
  - 無向グラフ(undirected graph):辺(edge)に向きがない場合

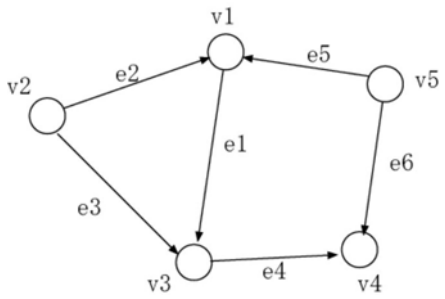


有向グラフの例

n個の頂点とm本の辺を持つグラフG=(V,E)を考える。

$$V=\{v_1,v_2,v_3,\ldots,v_n\},\quad E=\{e_1,e_2,e_3,\ldots,e_m\}$$

例:



$$V=\{v_1,v_2,v_3,v_4,v_5\}$$

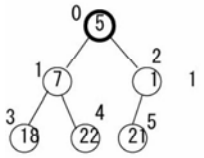
$$E=\{(v_1,v_3),(v_2,v_1),(v_2,v_3),(v_3,v_4),(v_5,v_1),(v_5,v_4)\}$$

グラフのコンピュータ内での表現法(1)

3.5.1 配列による実現

(1) 隣接行列(adjacency matrix)

- 無向グラフ：点iと点jが辺で結ばれている時には，(i, j)要素に1を入れる．
  - 対象行列になる．
- 有向グラフ：点iから点jに有向辺がある時には，(i, j)要素に1を入れる．
- 使用メモリ量：グラフの点を整数として表すと仮定し，各整数がc1バイトで表現されているとすると，単純グラフの隣接行列は$|V|^2$・c1使用する．

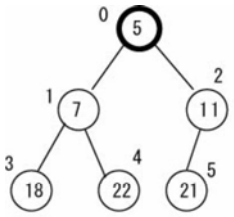


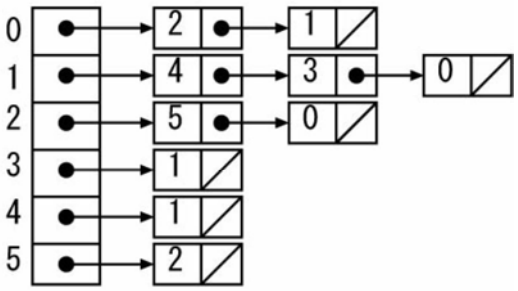
	0	1	2	3	4	5
0	0	1	1	0	0	0
1	1	0	0	1	1	0
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	1	0	0	0	0
5	0	0	1	0	0	0

隣接行列

グラフのコンピュータ内での表現法(2)

- 隣接リスト(adjacency list)
  - グラフの各点に対して，それに隣接する点の集合またはそれに接続する辺の集合を１本の線状リストとして表現したもの．
  - 使用メモリ量：O(|V|+|E|)



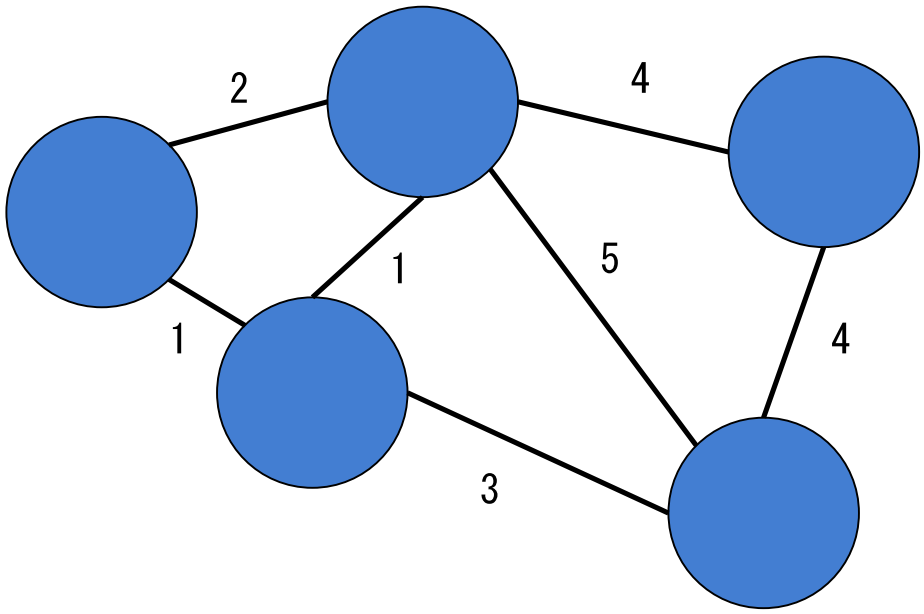


隣接リスト

最小木問題

- 最小木問題
  - グラフG=(V,E)で、全ての頂点を結ぶ辺の集合（Eの部分集合）で辺の重みの総和が最小となるもの
- 最小木
  - 木はVの全ての頂点を含む根付き木となる
  - 一般に一意には決まらない

最小木問題



最短路問題とダイクストラ法

枝の費用が非負のネットワークに対する
最短路問題の代表的なアルゴリズム

だいたいの教科書にはこれが最初に
書いてある.

ダイクストラ法（アイディア）

点の走査順序を「点が一度走査されたら、それ以降
走査する必要がない」ことが肝要

枝の費用が非負ならば，そのようにできる

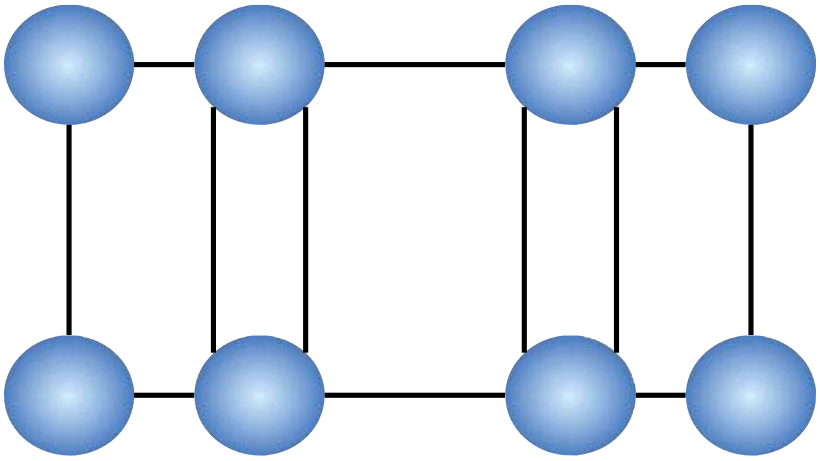
ポテンシャルが最小の点を走査すれば，
その点はその後走査する必要はない（なぜか？）

やってみよう

ダイクストラ法

Dijkstra法
　　ポテンシャルyの初期設定
　　S:=V
　　while Sが空でないならば do
　　y_vが最小の点 $v \in S$ を選択
　　S:=S＼{v}
　　点vの走査

オイラー閉路問題



オイラー（Euler）閉路

＝ すべての枝をちょうど１回ずつ経由する閉路

Euler閉路の求め方

グラフがEuler閉路を持つための必要十分条件

グラフの点の次数がすべて偶数かつグラフが連結

↓

Euler閉路の求め方（アルゴリズム）

適当な点から出発し、出発点と通過していない枝からなるグラフが連結であるように枝を適当な順番でたどる。

探索

member(x,S):

Sの中にxがあるか否かを調べ，あればYesを，なければNoを出力する.

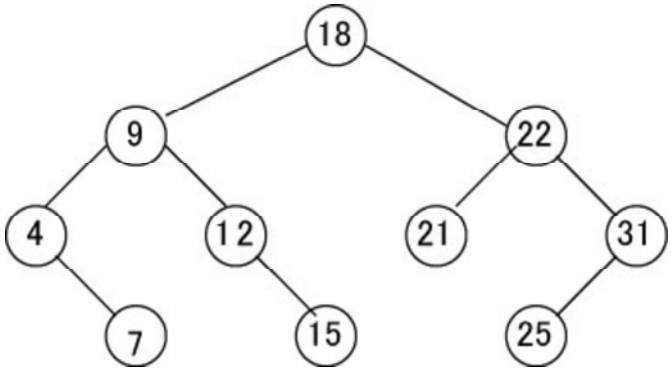
2分探索木と対称順

2分探索木

- 2分木の各点に全順序集合の要素を1つずつ対称順に配置したもの

対称順(symmetric order/inorder)

- (vの左子の子孫にある要素)<(vにある要素)<(vの右子の子孫にある要素)



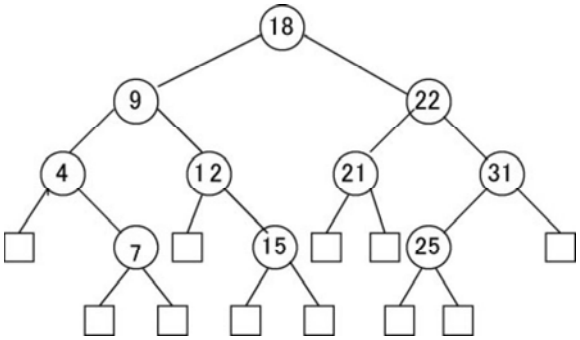
2分探索木と対称順(2)

正則2分木(regular binary tree)

- 各点が左子と右子の両方を持つかまたは全く子供を持たないか，いずれかの根付き木

2分探索木

- 正則2分木の各内点にSの要素を一つずつ対称順に配置したもの．但し，外点には配置しない
- 全ての葉が外点になるように正則2分木をつくる



2分探索木と対称順(3)

記号等

- 点vの左子，右子をそれぞれ，left(v), right(v)
- 内点vに配置されたSの要素をkey(v), vの親をparent(v)又はp(v)
- 点vが外点ならばleft(v)=NULLかつright(v)=NULL

実現

- ポインタを利用

2分探索木における基本操作：探索

2分探索木における基本操作

- 要素xの探索：member(x,S)
- 要素xの2分探索木への挿入：insert(x,S)
- 要素xの2分探索木からの削除：delete(x,S)

member(x,S)

- 根から葉に向って，各点で左子または右子を辿ることを反復してxを探す操作

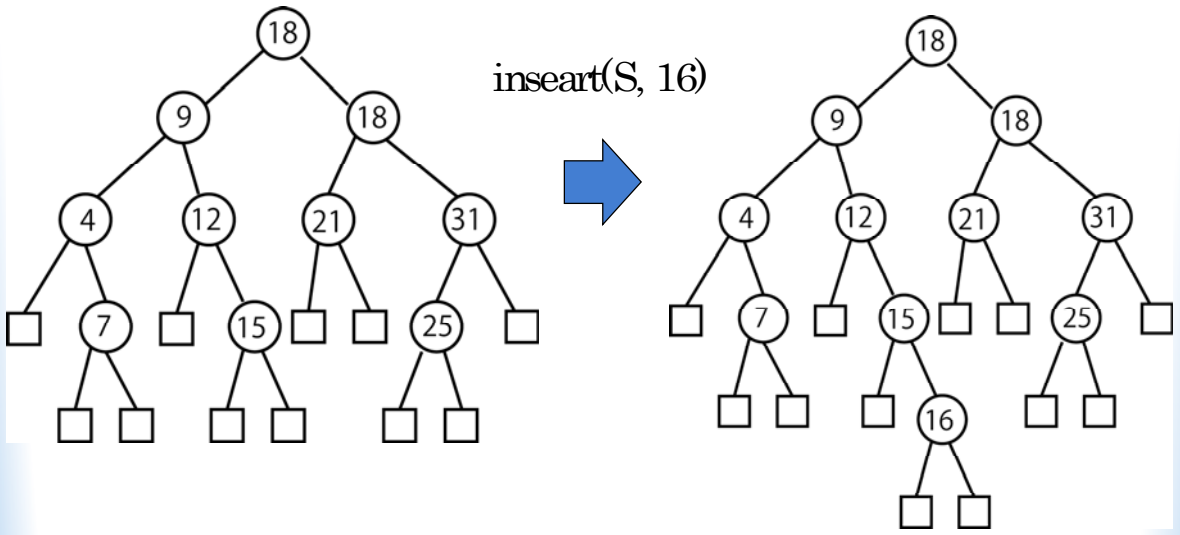
Step1: vが外点ならば，Noと出力して終了する.
Step2: (vが内点である) key(v)==x ならばYesと出力して終了する.
Step3: (key(v)!=xである)もしkey(v)>xならばv ← left(v).
もしkey(v)<xならばv ← right(v). Step1へ戻る.

2分探索木における基本操作：挿入

insert(x,S)

1. member(x,S)を実行する.
2. もし、x in Sなら何もしない. そうでないなら、点vは外点である.
3. vを内点に替えて、key (v)←xとする.
4. vの左の子、右の子を外点として追加する.

inseart(S, x)の例



2分探索木における基本操作：削除(1)

delete(x,S)

Step1: xより小さいもののうちで最大値をキーとして持つ点vの探索

aはkey(a)==xの点とする.

(i) left(a)が内点の時：

- aの左子left(a)に行く. そこから右子を辿ることをできるだけ繰り返し, 外点に到達したとき, その直前の内点をvとする. key(v)がこの部分木の内にあるキーの最大値である（確認せよ）.

(ii) left(a)が外点の時：

- xはaを根とする部分木のキーの中の最小値ということになり, 条件を満たす点は存在しないので, v ← aとする.

2分探索木における基本操作：削除(2)

Step2: キーの入れ換えと削除

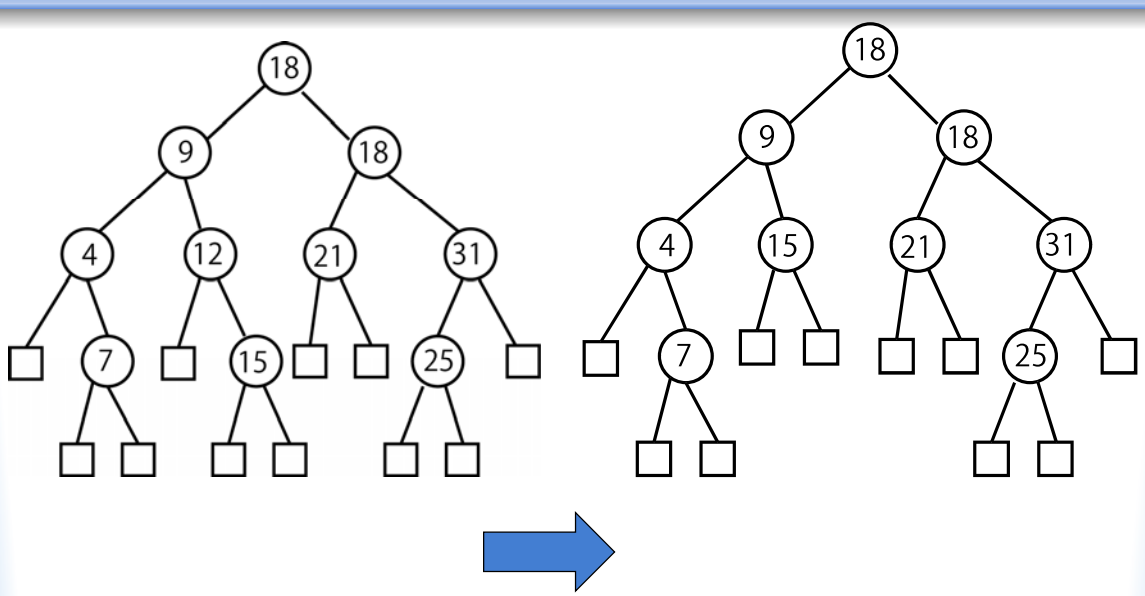
(i) left(a)が内点の時：この場合，right(v)は外点である（確認せよ）.

- key(a)とkey(v)を入れ換える.
- もしleft(v)が内点ならばright(parent(v)) ← left(v)とする.
- もしleft(v)が外点ならばv自身を外点に変える（xは削除する）.

(ii) left(a)が外点の時：v==aとなっている.

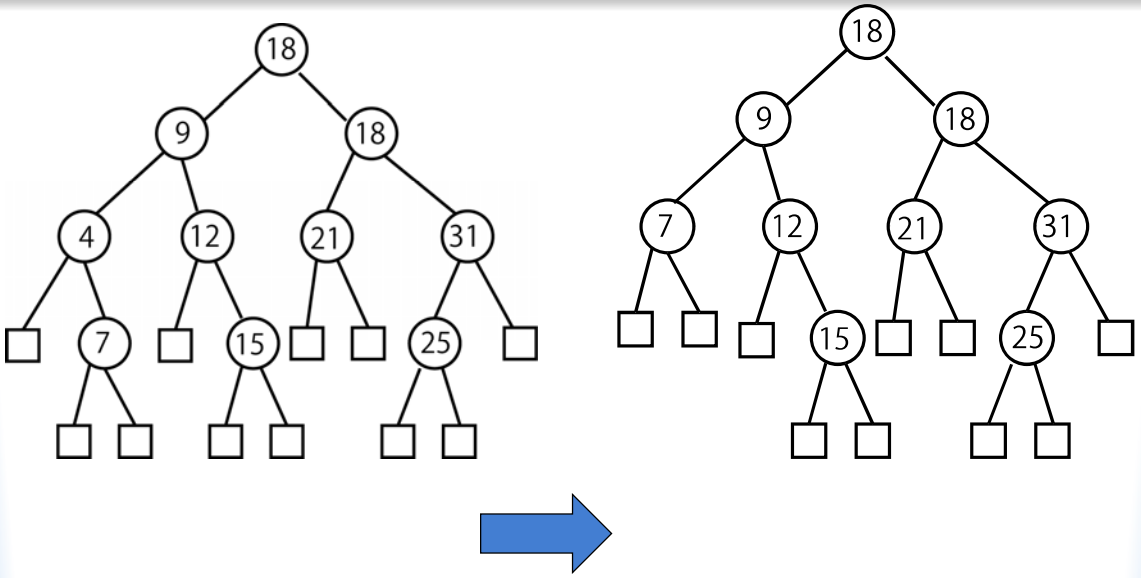
- もしright(v)が内点ならばright(v)をparent(v)の右子（vがparent(v)の右子の時）あるいは左子（vが左子の時）とする.
- もしright(v)が外点ならばv自身を外点に変える.

delete(S, x)の例(1)



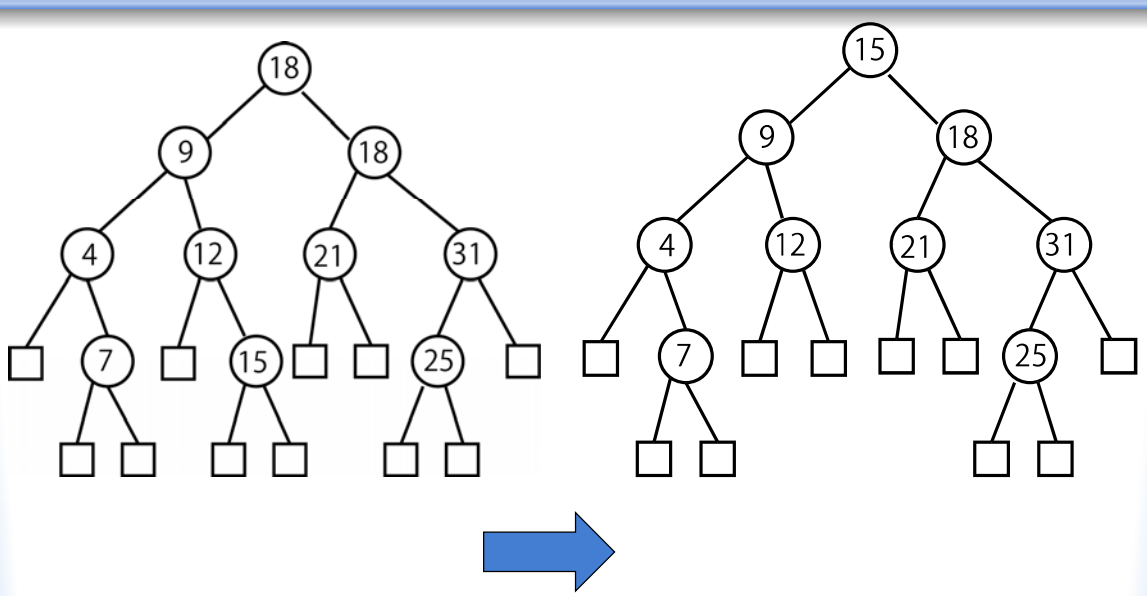
delete(S, 12)

delete(S, x)の例(2)



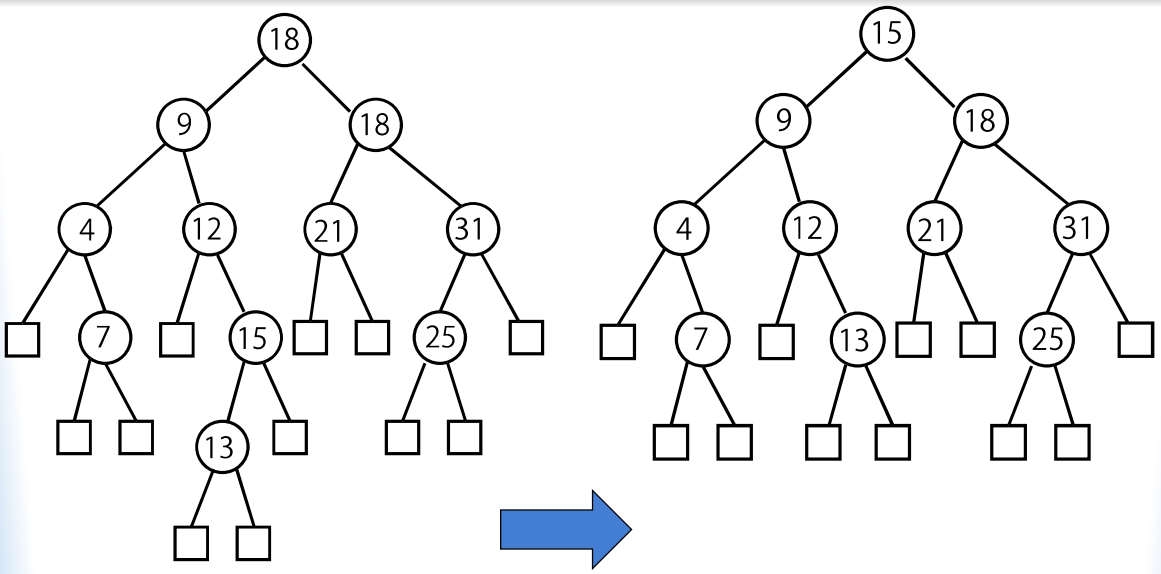
delete(S, 4)

delete(S, x)の例(3)



delete(S, 18)

delete(S, x)の例(4)



Inseart(S, 13); delete(S, 18)

2分探索木における基本操作:回転

時間計算量

- 2分探索木におけるmember, insert, deleteの各々は, いずれも最悪計算時間はO(height)である. ここでheightは2分探索木の高さである.

回転(rotation)

単回転(single rotation)

- 対称順を維持しながら2分探索木を変形する操作
- 右回転, 左回転

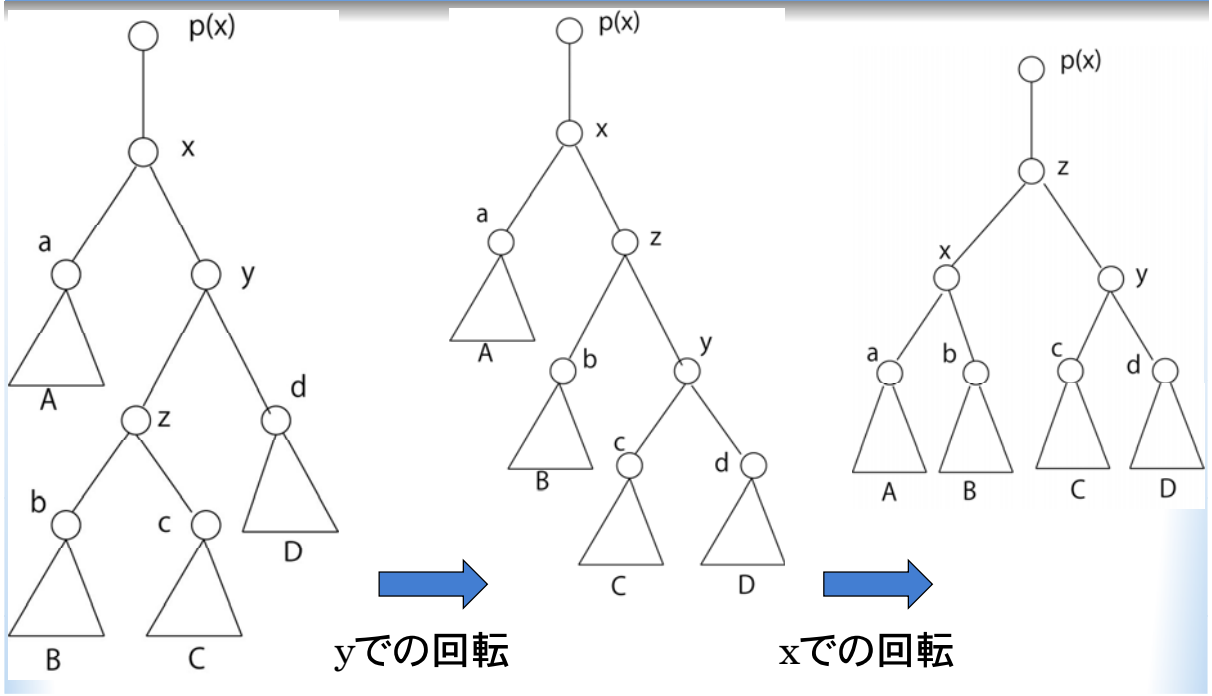
複回転(double rotation)

- 単回転の組合わせ

平衡2分探索木

- 木の高さがO(log n)で抑えられている木.
- 回転を利用することで実現できる.

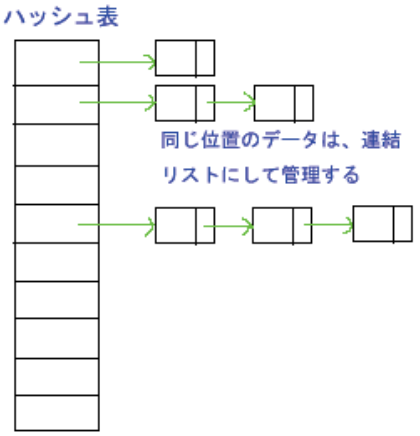
回転操作の例



ハッシュ法 (1)

ハッシュ法（チェイニング(chaining)方式）

1. xを区間[0..m−1] に入る整数に変換する関数h（ハッシュ関数）を準備する
2. m個のリストを準備し，各リストの先頭を指すポインタを配列A[0..m−1]に入れる.
3. Sの各要素xはA[h(x)]が指しているリストに入る.



ハッシュ法 (2)

member(x,S):

- h(x)を計算し，次にA[h(x)]が指しているリストを先頭の要素から順次調べ，xが入っているか否かを判定する.

insert(x,S):

- A[h(x)]が指しているリストの先頭にx を挿入する.

delete(x,S):

- A[h(x)]が指しているリスト内にxを見つけそれを除去する.

h(x)の値の分布が一様であれば，平均的に一つのリストの長さは n/mとなる.
m>|S|のようにmを選べば，各リストの長さは1以下となる.

基本操作の平均実行時間 O(1)

整列(ソーティング)

全順序集合Sが与えられたとき，S(|S|=n)の全ての要素をその全順序に従って並べ替えよ．

(1)バケットソート→O(m+n) (m:S中の最大整数値)

(2)選択法，(3)挿入法，(6)クイックソート→O(n^2)

(4)ヒープソート，(5)マージソート→O(n log n)

バケットソートと基数ソーティング

バケットソート(1)

n個の整数a_1, a_2, ..., a_nを考える．

- 各要素a_iは$a_i \leqq m$ (i=1, 2,..., n)とする．
- m個のバケツ(buckets) q(1), q(2), ..., q(m)を準備する．
  - 実際にはm個のキューを用いる．
- i=1, 2, 3,...の順に各要素a_iをバケツ$q(a_i)$に入れていく．
- バケツ内の要素を連接する．

計算時間

- バケツの準備(キューの生成)にO(m)
- n個の要素をバケツに入れる(キューに挿入する)はO(n)
- バケツ内の要素の連接にO(n)
- 全体ではO(m+n)

選択法(1)

選択法(selection sort)

1. Sの中の最小値を選ぶ.
2. 2番目に小さい値を選ぶ
3. ....
4. n番目に小さい値を選ぶ

ステップ									比較	入れ換え
	15	19	29	9	16	11	6	3		
1									7	1
	3	19	29	9	16	11	6	15		
2									6	1
	3	6	29	9	16	11	19	15		
3									5	1
	3	6	9	29	16	11	19	15		
4									4	1
	3	6	9	11	16	29	19	15		
5									3	1
	3	6	9	11	15	29	19	16		
6									2	1
	3	6	9	11	15	16	19	29		
7									1	0
	3	6	9	11	15	16	19	29	28回	6回

選択法(2)

時間計算量

- 比較回数=(n−1)+(n−2)+....+2+1=n(n−1)/2
- 入れ替え回数=<n−1
- 合計で, $O(n^2)$

バブルソート(1)

1. n番目とn−1番目を比較し、n番目の方が小さければ入れ替える。
2. n−1番目とn−2番目を比較し、n−1番目の方が小さければ入れ替える。
3. 同様の処理をn−1回繰り返すと、1番目にn個の中から最も小さい数字がくる。
4. 1番目を除いたn−1個の数に対して、上記の作業を行う。すると2番目にn−1個の中から最小の数字（つまりn個の中で2番目の小さい数字）がくる。
5. 同様の処理をn−1回繰り返す。

バブルソート(2)

- 15 19 29 9 16 11 6 3
- 3 15 19 29 9 16 11 6
- 3 6 15 19 29 9 16 11
- 3 6 9 15 19 29 11 16
- 3 6 9 11 15 19 29 16
- 3 6 9 11 15 16 19 29
- 3 6 9 11 15 16 19 29

バブルソート(3)

時間計算量

- もっとも小さいデータを一番左まで運ぶのにn−1回
- 2番目に小さいデータを左から2番目まで運ぶのにn−2回
- 3番目に小さいデータを左から2番目まで運ぶのにn−3回
- 比較回数=(n−1)+(n−2)+....+2+1=n(n−1)/2
- 合計で，$O(n^2)$

挿入法(1)

挿入法(insertion sort)

1. j=2,3,...,nなる各jに対して以下を繰り返す
2. A[1],...,A[j−1] までは既に整列されているとする(されている)
3. A[j]をA[1]からA[j−1]の適当な位置へ挿入してより長い整列された列をつくる

挿入法(2)

ステップ									比較	入れ換え
	15	(19)	29	9	16	11	6	3		
1									1	0
	15	19	(29)	9	16	11	6	3		
2									1	0
	15	19	29	(9)	16	11	6	3		
3									3	3
	9	15	19	29	(16)	11	6	3		
4									3	2
	9	15	16	19	29	(11)	6	3		
5									5	4
	9	11	15	16	19	29	(6)	3		
6									6	6
	6	9	11	15	16	19	29	(3)		
7									7	7
	3	6	9	11	15	16	19	29	26回	22回

挿入法(3)

時間計算量

- 比較回数の最大値は，各繰り返しで，「整列済の要素数」に一致する.
- 全体では，比較回数=1+2+3+...+(n−1)=n(n−1)/2
- 入れ替えの回数は比較回数以下である
- 合計では，$O(n^2)$

選択法との比較

- 最悪の時間計算量は等しい
- 整列法は部分的に整列されたところがある場合には比較回数がかなり減少する.
- 一般的に選択法より速い

ヒープソート(1)

ヒープソート

- Sの要素をキーとするヒープを構成する.
- findmin, deleteminを繰り返す.

計算時間

- log(n−1)+log(n−2)+...+log 2=log(n−1)!<= n log n
- O(n log n)

マージソート(1)

マージソート(mergesort)

Step1: 入力系列を2分割する.

Step2: 各々の系列をマージソートで整列する.

Step3: ソートされた系列をマージする.

マージソート(2)

ステップ		
	15 19 29 9 16 11 6 3	
1		2等分割
	(15 19 29 9) (16 11 6 3)	
2		2等分割
	(15 19) (29 9) (16 11) (6 3)	
3		2等分割
	(15) (19) (29) (9) (16) (11) (6) (3)	
4		統合
	(15 19) (9 29) (11 16) (3 6)	
5		統合
	(9 15 19 29) (3 6 11 16)	
6		統合
	(3 6 9 11 15 16 19 29)	

マージソート(3)

時間計算量

- $n=2^k$の場合に考える．そうでない場合も結果は一緒になる．
- n個の要素をマージソートするのに必要な計算量をT(n)とする．
- マージに要する時間計算量は，O(n)となる．

時間計算量T(n)は

$$\begin{cases} n \geq 2: & T(n) = 2T(n/2) + cn \\ n = 1: & T(n) = c' \end{cases}$$

となる．一般に，n≧２の時には，

マージソート(4)

$$
\begin{aligned}
&T(n) = 2T(n/2) + cn = 2\{2T(n/2^2) + c(n/2)\} + cn \\
&= 2^2 T(n/2^2) + 2cn = 2^2\{2T(n/2^3) + c(n/2^2)\} + 2cn \\
&= 2^3 T(n/2^3) + 3cn \\
&\cdots\cdots \\
&= 2^k T(n/2^k) + kcn \\
&= c'n + cn\log n \le c''n\log n
\end{aligned}
$$

したがって,

$$O(n\log n)$$

となる.

クイックソート(1)

クイックソート(quicksort)

- Step1: 入力系列から一つの要素x(ピボットと呼ぶ)を選ぶ.
- Step2: xを基準にして系列を次の二つの系列に分割する(partition).
- $\alpha 1$: a≦xなる要素からなる系列
- $\alpha 2$: a≧xなる要素からなる系列
- Step3: $\alpha 1, \alpha 2$をクイックソートで整列する.

クイックソート(2)

プログラムの設計

- 入力系列は，A[1], A[2], ..., A[n]に与えられているものとする．
- 但し，A[0]=−∞とする．
- ピボットの選択
  - 系列の左端A[left]，右端A[right]，中央位置A[center]の3要素を取り出す．ここで，center=[(left+right)/2]とする．
  - A[left]≦A[center]≦A[right]の順に並べ替える．
  - A[center]をピボットとする．すなわち，3要素の中央値(median)としている．
  - A[center]とA[right]を入れ換える．
- 分割の方法
  - A[left]から右方へ移動しながらxと系列の要素を比較して，A[i]≧xとなる最初の位置で止る．
  - A[right−1]から左方へ移動しながらxと系列の要素を比較して，A[j]≦xとなる最初の位置で止る．
  - A[right]=xなので，i≦nで，A[0]=−∞だからj≧0である．

クイックソート(3)

例

```
(left)            (center)                (right)
 15   29   19   11    6    9   16    3
  3   29   19 <11>    6    9   16   15
  3   29   19   15    6    9   16 <11>
i                                     j
  3   29   19   15    6    9   16 <11>
       i                   j
  3   29   19   15    6    9   16 <11>
       i                   j
  3    9   19   15    6   29   16 <11>
            i         j
  3    9   19   15    6   29   16 <11>
            i         j
  3    9    6   15   19   29   16 <11>
            j    i
  3    9    6   15   19   29   16 <11>
  3    9    6 <11>   19   29   16   15
```

```
( 3    9    6) (11) (19   29   16   15)
( 3   <6>   9) (11) (15 <19> 16   29)
                 i                   j
(3)   (6)   (9) (11) (15   29   16 <19>)
                                i    j
(3)   (6)   (9) (11) (15   29   16 <19>)
                                j    i
(3)   (6)   (9) (11) (15   16   29 <19>)
(3)   (6)   (9) (11) (15   16 <19> 29)
.....
```

クイックソート(4)

時間計算量(ただし，cutoff=0とする)

- 最悪時間計算量 $O(n^2)$
- 最良時間計算量 $O(n \log n)$
- 平均時間計算量 $O(n \log n)$
- 入れ換えの回数より比較の回数が多くなる.
- 長さnの系列の分割(partition)は，$O(n)$.

最悪時間計算量

- 分割がアンバランスになる時に最悪の時間計算量となる.
- $T(n)=O(n^2)$

最良計算時間

- バランス良く分割が行われた場合

クイックソート(5)

クイックソートの計算時間

(i) ピボットxの選び方により異なる.

- 時間をかけずにできるだけ全体の中央値を求めたい.
- (1)ランダムにxを選ぶ.
- (2)ランダムに何個か要素を選び，その中の中央値をxとする.
- (3)左端からみて最初に出会う異なる2要素の大きいものをxとする.
- (4)左端，右端とこれらの中間位置の3要素の中央値をxとする.

(ii) nを固定したとき，計算時間が$O(n^2)$となる入力もあれば，$O(n \log n)$となる入力もある. $O(n^2)$となるのは，n!の入力系列のうち，どの程度か？

- $O(n \log n)$より大きくなる系列は確率的に極めて少ない.
- nが大きくなるとn!の1%強程度である.
- ほとんどの場合，$O(n \log n)$となる.

ヒープ（順序付きキュー）

完全２分木(1)

用語の説明

- 根付き（無向）木
- 親，子，兄弟，祖先，子孫
- 葉(leaf)
- 点の深さと点の高さ，木の高さ
- 完全2分木

点の番号付けと親子関係

完全２分木の点数，高さ，深さの関係

- 点数nと高さhの関係は

$$2^0 + 2^1 + \ldots + 2^{h-1} + 1 \le n \le 2^0 + 2^1 + \ldots + 2^{h-1} + 2^h$$

$$2^h \le n \le 2^{h+1} + 1$$

$$h \le \log_2 n < h + 1$$

ヒープ

ヒープ

- スタック，キューは,要素が入ったタイミングで取り出される順番が決まる
- 入ったタイミングではなく，優先度が最大の要素を最初に取り出す

ヒープの定義

- 集合S={a_1, a_2, ..., a_n}の各要素間には全順序があるとする.
- n個の点を持つ完全d分木の各点にSの要素を１個ずつ割当てたもので，全ての点vが以下の条件を満たす.
- [ヒープ条件]vに割当てられているSの要素はvの子供に割当てられているどの要素よりも大きい（小さい）か等しい.

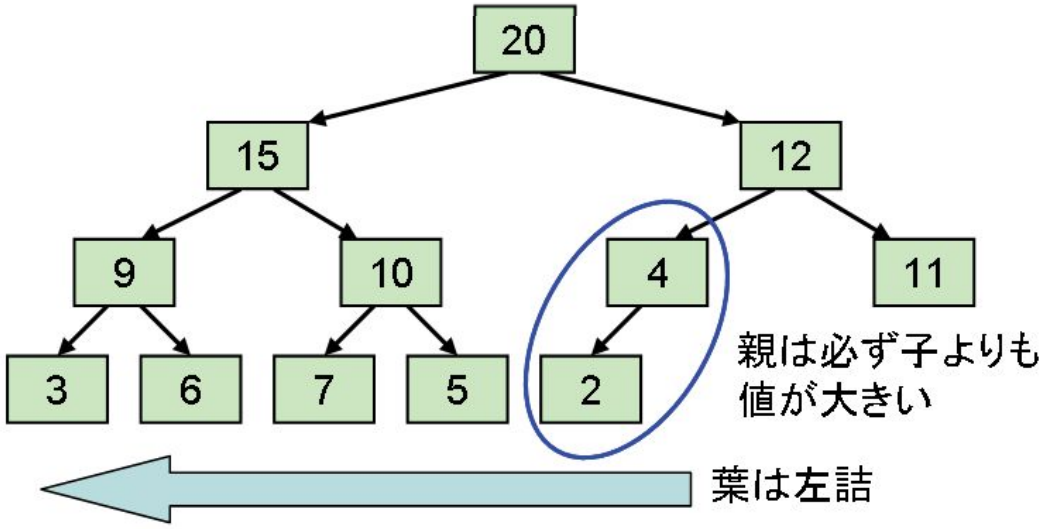
ヒープの根にはSの中の最大（小）要素が割当てられる.

基本操作

基本操作

- insert: 新しい要素の挿入
- findmax: ヒープ中の最大要素を見つける
- deletemax: ヒープ中の最小要素を除去する
- decreasekey: ヒープ中の一つの要素の値をある一定値減少させる

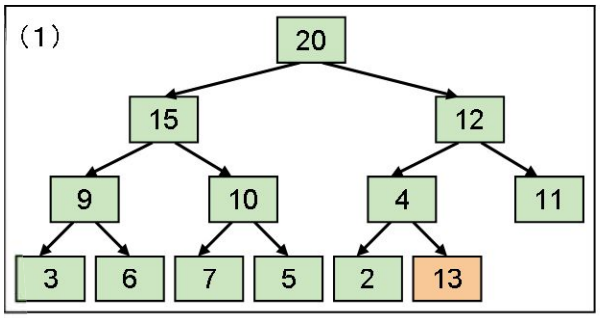
ヒープ

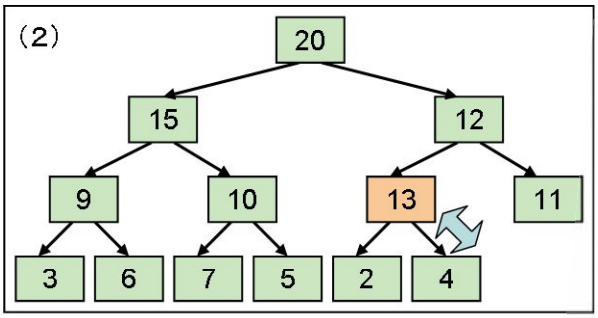


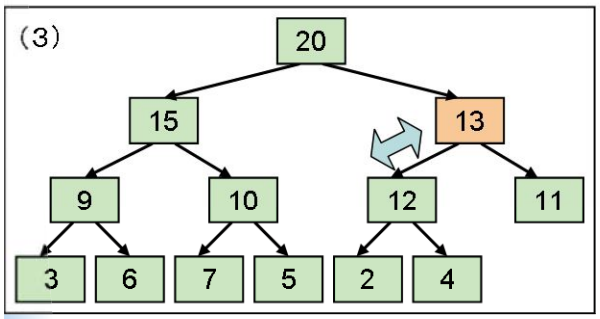
1	2	3	4	5	6	7	8	9	10	11	12
20	15	12	9	10	4	11	3	6	7	5	2

iの親は $[i/2]$
左の子は $2i$
右の子は $2i+1$

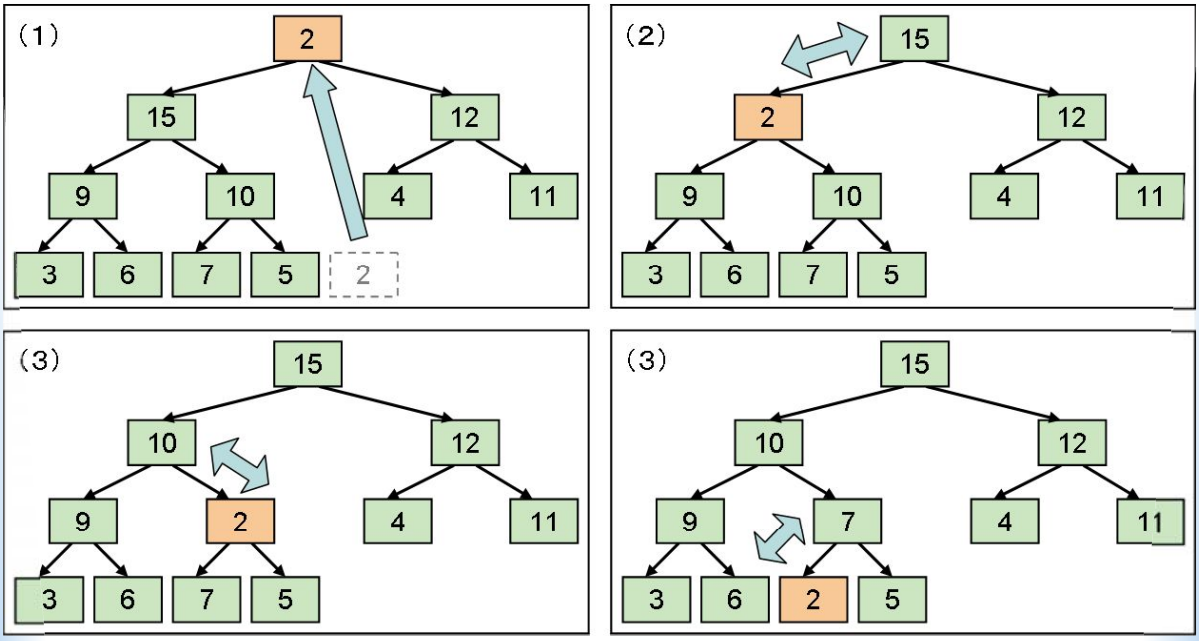
要素の挿入







最大要素の削除



基本操作の時間計算量

命題

- n個の点を持つ2ヒープの基本演算insert, findmax, deletemaxの各々の計算時間は,
  - insert, deletemaxが$O(\log n)$時間,
  - findmaxが$O(1)$時間である.

完全２分木の点数, 高さ, 深さの関係

$$2^0+2^1+\ldots+2^{h-1}+1 \le n \le 2^0+2^1+\ldots+2^{h-1}+2^h$$

$$2^h \le n \le 2^{h+1}+1$$

$$h \le \log_2 n < h+1$$

完全２分木の点数をn, その高さをhとすれば,

$$h \le \lfloor \log_2 n \rfloor$$

ヒープの構成法(1)

insertの反復

- Step1: 空のヒープを作る.
- Step2: insertを繰り返す.

時間計算量

- i個の点を持つヒープに1点を追加してヒープを再構成するのに必要な時間計算量は $O(1+\lfloor \log(i+1) \rfloor+1)$
- 総計算量は，一般のn(≧2)に対して，

$$1+\sum (1+\lfloor \log(i+1) \rfloor+1) \le 2n-1+\log n! \le 2n-1+\log n^n$$

- に比例する．すなわち，$O(n \log n)$

命題

- n個の点を持つ2ヒープはinsertの反復により0(n log n)の計算時間で構成できる.

ヒープの構成法(2)

Shift-downの反復（点数が固定されている場合を想定）

ヒープ化

- 2つのヒープがある時に，新しい1点vを根としこれらのヒープを部分木とするような完全２分木から，１つのヒープを構成する.
- height(v)+1に比例する計算時間でできる.

ヒープの構成

- Step1: 点数nの完全２分木を作り，各点の中にキーを適当に１個ずつ仮配置する.
- Step2: 完全２分木の配列番号を逆順に辿りながらヒープ化を繰り返す.

ヒープの構成法(3)

時間計算量

- 木の高さをhとする．深さiの点は高々2^i個存在する．その高さはh-i以下である．
- 深さiの各点から葉に向って多くてもh-i+1回のshift-downが繰り返される．
- 深さiは，h-1，h-2，...，0と変わる．
- よって，全体でのshift-downの実行回数は

$$\sum_{i=0}^{h-1}(h-i+1)*2^i=(h+1)\sum_{i=0}^{h-1}2^i-\sum_{i=1}^{h-1}i2^i$$
$$=(h+1)(2^h-1)-\{(h-2)2^h+2\}$$

- となり，O(n)である．
- 命題
- n個の点を持つ2ヒープはshift-downの反復によりO(n)時間で構成できる．

演習問題(1)

1. 次の計算時間をオーダーで表せ。C は正の定数である。

(1) $C + 2C + 3C + \cdots + nC$

(2) $C + 2^2C + 3^2C + \cdots + n^2C$

(3) $C\log 2 + C\log 3 + \cdots + C\log n$

2. 次の問に答えよ。

(1) 空のスタックに対して、push(3), push(5), push(2), pop(), pop()を実行した時のスタックの状況を図示せよ。

(2) 空のキューに対して、enqueue(1), enqueue(9), dequeue(), enqueue(3), dequeue(), dequeue()を実行したときのスタックの状況を図示せよ。

演習問題(2)

9 個の整数 25, 29, 7, 32, 18, 10, 16, 5, 2 がこの順番に入力されたとして、次の(1)〜(3)に答えよ。ただし、大小関係を逆転し「根」に最小値が来るようにせよ。

(1) insert を繰返すことによってヒープ g1 を構成せよ。

(2) g 1 において deletemin 実行後のヒープ g2 を示せ。

(3) g2 において 25 を 3 に減少させる decreasekey(25, 3)実行後のヒープ g3 を示せ。

演習問題(3)

入力系列 12, 3, 24, 30, 4, 9, 15, 8 に対して、選択法、バブルソート、挿入法、ヒープソート、マージソート、クイックソートを適用した時のデータの並びの変化を図示せよ。

演習問題(4)

1. 以下の(1)〜(4)に示す整数がこの順に入力データとして与えられた時、insert を反復して２分探索木を構成せよ。
   (1) 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
   (2) 20, 18, 16, 14, 12, 10, 8, 6, 4, 2
   (3) 6, 16, 8, 14, 10, 12, 2, 4, 18, 20
   (4) 10, 6, 16, 14, 18, 4, 2, 12, 20, 8

2. 以下で指示する２分探索木とキーに対して delete を実行せよ。
   (1) 前問(3)の２分探索木から 14
   (2) 前問(4)の２分探索木から 10
   (3) 前問(4)の２分探索木から 18

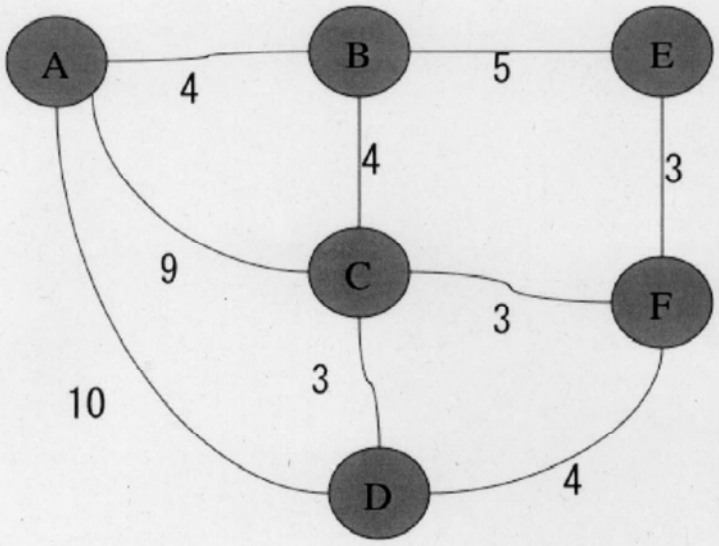
演習問題(5)

以下のグラフに関して次の質問に答えよ。

(1) 下記のグラフはオイラーグラフではない。その理由を述べよ。また、枝を追加してオイラーグラフにせよ。その時のオイラー閉路も求めること。

(2) 点Aを根とする最小木を求めよ。

(3) 点Aを出発点として他の各点までの最短路をダイクストラ法を用いて求めよ。



3.7　類似性発見

3.7.1 概要

講師：岡﨑 威生／琉球大学工学部情報工学科講師岡﨑

日時：2012 年 9 月 11 日　10:20～／12:50～／14:40 ～／16:30 ～

場所：琉球大学総合情報処理センター

内容：データ解析の基礎となる確率論と統計的推論からスタートし、データ解析言語 R を利用した実習を行う。次に基本的な配列解析であるアラインメントの原理と方法を演習を交えて理解する。最後に類似性に基づいたクラスタリング手法について、適用ノウハウと供に学ぶ。

3.7.2 準備

特筆事項なし

3.7.3 配布物

・講義スライド
・R 実行メモ
・gene.csv（遺伝子発現データ、R 実行メモ用データファイル）
・正規分布表
・t 分布表
・F 分布表
・BLOSUM50（スコアテーブル）
・PAM250（スコアテーブル）
・方眼紙（アラインメントワーク用）
・統計解析言語システム R の紹介（参考資料）

3.7.4 課題

・プリント（演習問題 12 題）

3.7.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.7.6 時間割

時間	内容
10:20～	【映像上映】 ・ セントラルドグマ ～ゲノム情報からタンパク質ができるまで～ / The Central Dogma 独立行政法人理化学研究所 ゲノム科学総合研究センター http://www.youtube.com/watch?v=EgweXtBCynE
10:38～	確率分布 ・ 確立とは 算術的確率／統計的確率 ・ 確率の性質 →R 演習を実施 ・ 条件付き確率 →R 演習を実施
11:00～	・ ベイズの定理 ・ 確率分布 ・ 分散と標準偏差 →R 演習を実施
11:47～	・ 標準化
12:00～	昼食
12:50～	確率分布（続き） ・ 正規分布（二項分布は省略） ・ 標本と母集団 ＊ 時間の都合上、統計的推定の項目は省略
13:25～	配列アラインメント ・ 動的計画による候補の探索 ・ ギャップペナルティ リニアギャップ／アフィンギャップ BLOSUM50／PAM250 ・ Needleman-Wunsch 法
14:50～	→アラインメント演習 ・ 相同性探索ツール BLAST ・ 統計的検定

	・ 位置の違い検定 ・ t 分布 数表 →位置の違いの検定演習
15:27～	・ ばらつきの違い検定（F 検定） ・ Wilcoxon 順位和検定
15:36～	相関／相関係数 ・ 相関係数の概念 ・ 相関係数の計算式 ・ ノンパラメトリック相関係数
15:57～	→相関係数演習
	休憩
16: 15～	クラスタリング【演習】 ・ クラスタ／グループ化／デンドログラム ・ 似ている度合いの測定方法 距離、類似度 ・ 階層的クラスタ構成 ・ クラスタ併合のルール 重心法、最短距離法、最長距離法 群平均法、Ward 法 ・ クラスタリングの留意点 数値の確認、距離の選択、併合ルールの選択
16:47～	→クラスタリング演習

特別講義V「類似性発見」
9月10～11日
バイオインフォマティクス人材育成講座

琉球大学 工学部 情報工学科
岡﨑　威生
okazaki@ie.u-ryukyu.ac.jp
http://www.ms.ie.u-ryukyu.ac.jp/~okazaki/

似ているってどういうこと？

- 生命情報データの解析キーワード
  - 同じ？
  - 近い？
  - 似ている？
- どうやって“客観的に”説明する？
  - 数学的に
  - 情報科学的に

確率・統計の役割

- 生命情報データを獲得した後
  - 正確？
  - いつでも同じ？

    不確定な現象　→「確率」的な見方が必要

- データをどうやって“客観的に”説明する？
  - 確率的な尺度に基づいて

    統計的手法

目　標

- データに不確定的要因が含まれることを意識し、確率的な視点で考察できる。
- 確率に基づいた統計的手法の意図を理解し、活用できる。
- データの類似性について解析する手法を理解し、ソフトウェアを活用して実践する。

予　定

16:20 – 17:50	計算環境、R、データ記述
10:20 – 11:50	確率分布、統計的推定
12:50 – 14:20	配列アラインメント
14:40 – 16:10	仮説検定、相関
16:20 – 17:50	クラスタリング

計算環境（ハードウェア）

Bioinformaticsで必要とする計算環境って？

ハードウェアの視点

- どんな計算をするの？
  大量データ（サイズ、数）を処理
  → 「高速性」が最重要
  → Super Computer
  - 京（世界一！）
  - 東京大学ヒトゲノムセンター
- PCでできないの？
  CPU, Memory, SDD, GPU　　並列処理・グリッド

計算環境（ソフトウェア）

ソフトウェアの視点

- Operating System
  ハードウェアをできるだけ計算に有効活用
  → Linux
  オープンソース（バザール方式）
  多くのDistribution（Redhat, Ubuntu, Debian...）
  CentOSのインストールデモ
- アプリケーション
  - Excel　取っつきやすい
  - 統計解析言語R　オープンソース、高機能、拡張性
    世界中の研究者が成果をライブラリとして追加
    ※別資料参照　　Rインストール

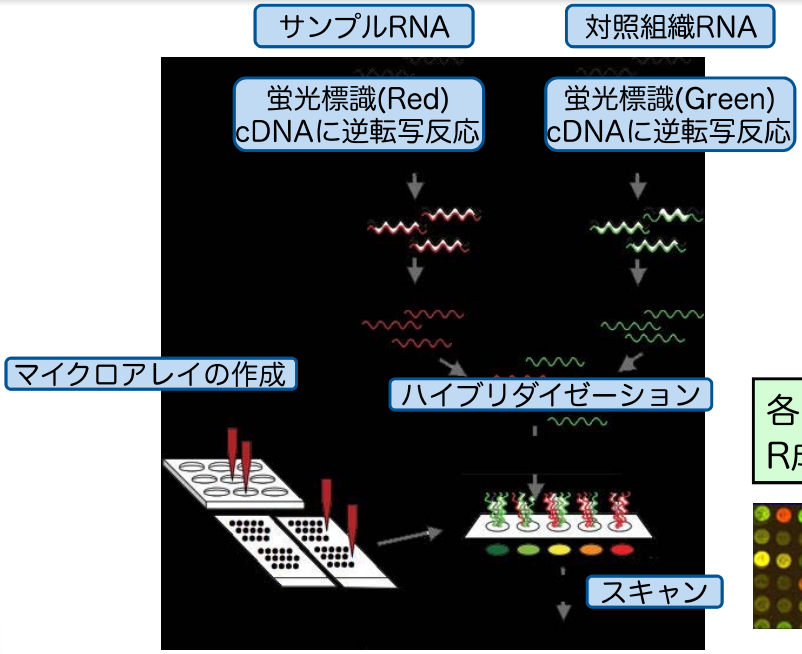
データ記述

EM8Z309	EM8Z310	EM8Z313
21.5	22.6	14.9
18.6	13.2	17.3
13.9	16.6	13.7
58.4	55.4	55.7
20.4	10.2	20.0
20.3	16.3	19.0
19.4	19.0	9.8
56.4	28.7	66.5
22.2	17.6	15.0
56.6	54.2	45.9

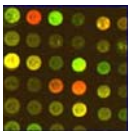
観測・実験結果データから、
何を考えることができるか？

- 平均
- 範囲
- 関係
- 予測
-

遺伝子発現データ



各スポットから
R成分とG成分の観測値



$$\log_2\left(\frac{R}{G}\right)$$

データの代表値

- 標本平均値：中心的傾向を示す

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 最頻値：度数が最多のカテゴリ値
- 中央値：大小順で順位が真ん中
- 最大値：大小順で最大
- 最小値：大小順で最小
- 四分位数：小さい順に並べて累積確率が
  0.25　第1四分位数
  0.75　第3四分位数

R演習

- csv形式（Excelファイル）をRで読み込む
  - 文字コード、ヘッダ、区切り記号
- データ表示
  - 数値表示、代表値計算

※データ入力の方法
  - Rの中で直接キータイプ入力
  - 外部ファイルから　⇐　各種DB、Excel

データのバラツキ

- 標本範囲：最大値から最小値を引いた値
- 標本分散：標本平均からの偏差の２乗の平均値

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 標本標準偏差：標本分散の平方根
- 四分位偏差：第3四分位数から第1四分位数を引いた値
  データの真ん中半分が収まる長さ

グラフ表現

調査・観測・実験で得られたデータは“チラバリ”を持つ

どんなチラバリかたをしているか？

- 数値で説明
  - データ全体を説明できるか
  - 特定の視点に偏ってないか
- 図で説明　→　グラフ
  - ヒストグラム
  - 散布図
  - ボックスチャート

  みんなが同じ受け止め方をするか？

確率とは

現象の起こりやすさ、確からしさの尺度

試行：実験／観測などの行為

事象：試行によって起こる結果

標本空間：試行によって起こる全ての場合の集合

- 算術的確率

  同程度の起こりやすさを仮定　$P(A) = \frac{\# A}{\# \Omega}$

- 統計的確率

  試行回数をn、事象Aの起こる回数を$n(A)$　$P(A) = \frac{n(A)}{n}$

  nが大きいときAの頻度がほぼ一定値

確率の性質

- $0 \le P(A) \le 1$
- $P(\Omega) = 1$
- Ω の部分集合 $A_1, A_2, \cdots$ が $A_i \cap A_j = \varnothing \quad (i \ne j)$ のとき

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

1) $P(A^C) = 1 - P(A)$
2) $P(\varnothing) = 0$
3) $A \subset B \quad \rightarrow \quad P(A) \le P(B)$
4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

条件付き確率

事象A,Bについて $P(A) > 0$ のとき、
条件Aの下でのBの条件付き確率を

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

1) $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$
2) $A_1, A_2, \cdots, A_n$ が互いに排反で、 $\bigcup_{i=1}^{n} A_i = \Omega$ のとき

$$P(B) = P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)$$

ベイズの定理

$A_1, A_2, \cdots, A_n$ が互いに排反で、 $\bigcup_{i=1}^{n} A_i = \Omega$ のとき

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}$$

$$= \frac{P(A_k)P(B \mid A_k)}{P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)}$$

ベイズの定理より、事象と条件を逆にした条件付き確率の計算が可能

事象の独立

事象A,Bが独立 $\Leftrightarrow$ $P(A \cap B) = P(A)P(B)$

$\Leftrightarrow$ $P(B \mid A) = P(B)$

$\Leftrightarrow$ $P(A \mid B) = P(A)$

Aという条件があってもなくてもBの確率は変わらない

Bという条件があってもなくてもAの確率は変わらない

A,Bが互いに確率的な影響を与えない

確率分布

- 確率変数：考察対象を変数化したもの
  確率を伴う変数
- 確率分布：確率変数に対して、事象とその確率の対応
  確率変数の値の確率的なチラバリ
- 分布関数：確率変数Xに対して　$F(X) = P(X \le x) \quad (x \in R)$

- 離散型確率分布
  確率変数のとりうる値が離散的　例）二項分布
- 連続型確率分布
  確率変数のとりうる値が連続的　例）正規分布

期待値

期待値：確率変数Xの「中心」を表わす値
　　←　確率という重みを考慮して算出

$$E[X] = \sum_k x_k P(X = x_k)$$ 離散的確率変数

$$= \int_{-\infty}^{\infty} x f(x)\, dx$$ 連続的確率変数

分散・標準偏差

分散：確率変数Xの「バラツキ」の尺度

$$V[X] = E[(X-m)^2] \quad m = E[X]$$

$$= \sum_k (x_k - m)^2 P(X = x_k) \qquad \text{離散的確率変数}$$

$$= \int_{-\infty}^{\infty} (x-m)^2 f(x)\,dx \qquad \text{連続的確率変数}$$

標準偏差： $\sigma = \sqrt{V[X]} = \sqrt{E[(X-m)^2}$

【性質】 $E[aX+b] = aE[X]+b \quad V[aX+b] = a^2 V[X]$

$$V[X] = E[X^2] - (E[X])^2$$

標準化

標準化：相対比較のための変換
確率変数Xに対して

$$Z = \frac{X-m}{\sigma} \quad m = E[X], \sigma = \sqrt{V[X]}$$

【性質】 $E[Z] = 0$

$$V[Z] = 1$$

「中心」と「バラツキ」を一定に揃えることで
相対比較を容易にする

二項分布

ベルヌーイ試行：
　1回の試行の結果、2つの事象のいずれかが起こり、
　繰り返し行なっても生起確率が変わらない。

二項分布：
　繰り返し回数nのベルヌーイ試行で、事象の発生回数X

$$P(X=k) = {}_nC_k p^k (1-p)^{n-k} \quad (0 < p < 1)$$

【性質】　$E[X] = np \quad V[X] = np(1-p)$

正規分布

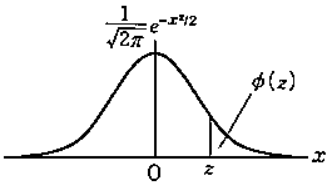
正規分布：
　自然科学や社会科学における多くの確率現象が該当

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

【性質】　$E[X] = \mu \quad V[X] = \sigma^2$
　　$Y = aX + b$ は正規分布
　　独立な正規分布の和は、正規分布

標準正規分布：標準化　$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

標準正規分布 数表



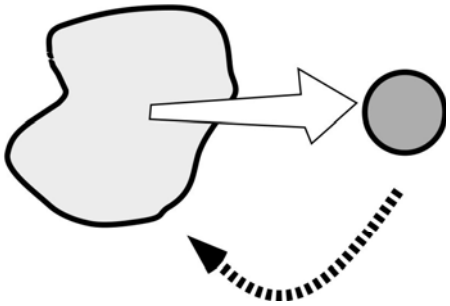
附表 2　正規分布　$\phi(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ の値

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

標本と母集団

母集団：調査（考察）対象の全体

標本：母集団から取り出した一部分　サンプリング



標本の情報から母集団の情報を推測

統計的推定

推定：

母集団から無作為に獲得した標本 $\{X_1, X_2, \cdots, X_n\}$ から
母集団のパラメータを推測する

標本確率変数 X_i：母集団分布と同じ確率分布　互いに独立

母集団のパラメータ

母平均：母集団の期待値　$E[X]$

母分散：母集団の分散　$V[X]$

統計量

統計量：標本確率変数の関数として表わされる確率変数

$$\{X_1, X_2, \cdots, X_n\} \quad \Rightarrow \quad Y = \varphi(X_1, X_2, \cdots, X_n)$$

よく利用される統計量

- 標本平均　$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 標本分散　$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- 不偏分散　$U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

統計量の期待値・分散

- 標本平均　$E\left[\overline{X}\right]=E[X]$　$V\left[\overline{X}\right]=\dfrac{V[X]}{n}$
- 標本分散　$E[S^2]=\dfrac{n-1}{n}V[X]$

$$V[S^2]=\frac{(n-1)^2}{n^3}\left(E[(X-E[X])^4]-\frac{n-3}{n-1}V[X]^2\right)$$

- 不偏分散　$E[U^2]=V[X]$

$$V[U^2]=\frac{1}{n}\left(E[(X-E[X])^4]-\frac{n-3}{n-1}V[X]^2\right)$$

良い推定量の条件

パラメータ θ の推定量 $\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)$ に対して

- 不偏性　中心が θ からズレていない　$E\left[\hat{\theta}\right]=\theta$
- 最小分散性　分散が小さいほうが望ましい
- 一致性　サンプルサイズが大きければ θ に近づく

標本平均：母平均に対して「不偏性」「一致性」
不偏分散：母分散に対して「不偏性」「一致性」

配列の類似性

塩基（DNA, RNA）やアミノ酸配列に対して

共通する遺伝子はどこ？　（遺伝子予測）
対象の遺伝子・タンパク質と似ているのは？
（相同性検索）
生命はどのように進化してきた？
（進化系統樹）

配列が似ていれば、機能も近いはずだ

配列アラインメントとは

アラインメントとは
☞　文字列を並べること

できるだけ同じ並びになるようにそろえる
☞　類似性を表すスコアを決めて、最大化

方法は？　総当たり
動的計画（Dynamic Programming）

配列アラインメント（総当たり）

次の２配列で、完全に一致する部分を探す

GCCACCAAGTTCCGCACATGCCGGATAGAAT

CCGGA

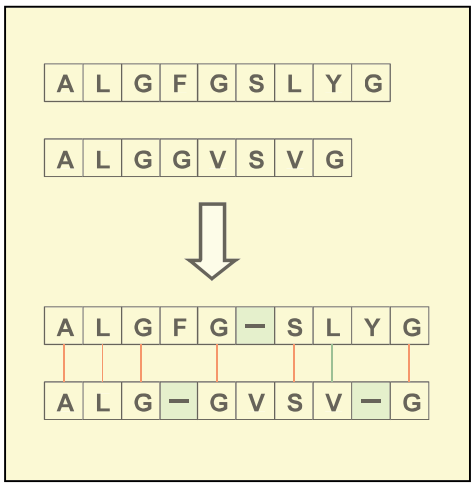
どうやって探す？

配列アラインメント（動的計画）

完全一致でなく、“おおよそ”一致

- ギャップがあってもいい
  - 挿入・欠失
  - 例）GACとGATC
- 違う文字が対応してもいい
  - 近い文字
  - 文字間の距離対応表が必要

総当たりで探すのは難しい



ギャップペナルティ

- リニアギャップ　⇒　$-gd$
  - ギャップ長　g
  - ギャップペナルティ　d
- アフィンギャップ　⇒　$-d-e(g-1)$
  - ギャップ開始ペナルティ　d
  - ギャップ伸張ペナルティ　e
  - ただし　$e<d$
  - 連続した挿入・欠失に対してリニアギャップより小さいペナルティ
    よく利用される値　$(d, e)=(12, 2), (11, 1)$

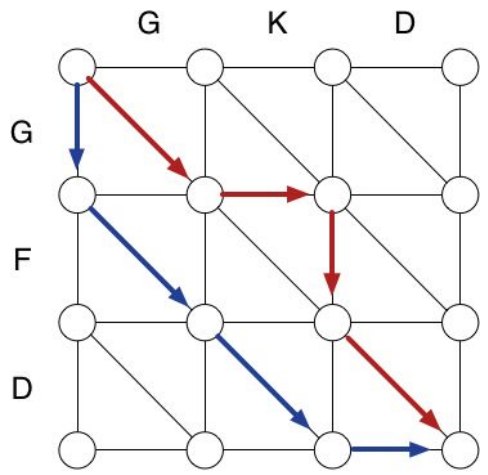
BLOSUM50（アミノ酸）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Needleman-Wunsch法

“できるだけ”高いスコアを見つける方法

- 格子状グラフを利用

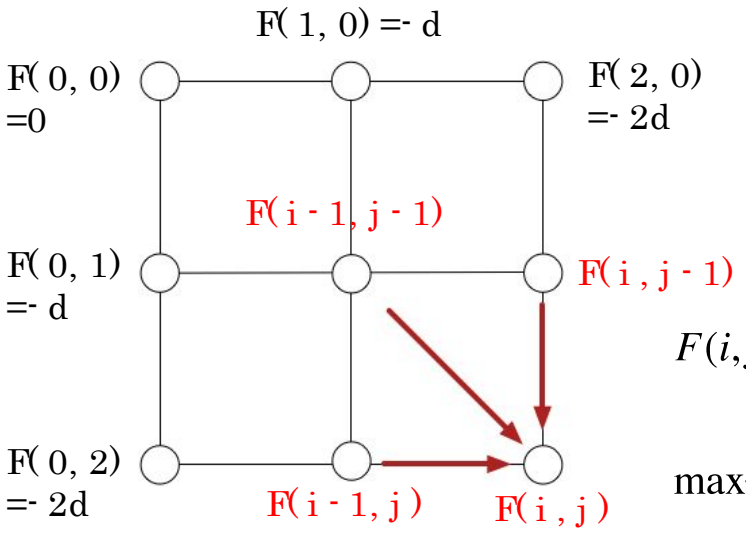


GK- D
G- FD

- GKD
GFD-

N-W法 格子状グラフの作り方

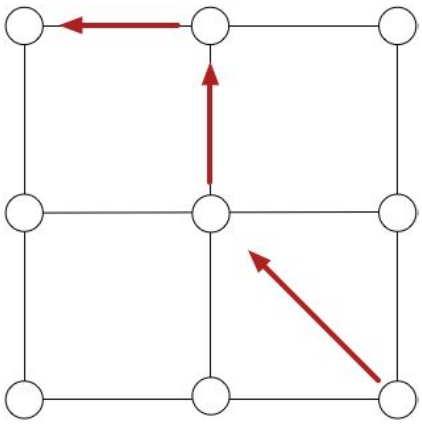
ギャップペナルティとスコア行列から



$$F(i,j) = \max\begin{cases} F(i-1,j-1)+S(x_i,y_j) \\ F(i-1,j)-d \\ F(i,j-1)-d \end{cases}$$

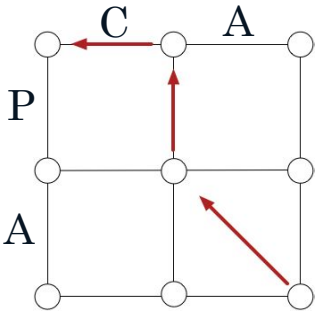
N-W法 トレースバック

- 右下（最大値）からスタート
- maxを与えた要素をたどる



N-W法 アラインメント書き出し

トレースバックの経路が
- 左上向きなら、両方の要素を対応
- 左向きなら、横列要素とギャップを対応
- 上向きなら、縦列要素とギャップを対応



A
A

- A
PA

C- A
- PA

アラインメント演習

次の２配列に対して、Needleman-Wunsch法で
アラインメントを実行しなさい。

A P E G L Q P L

A P D L Q A L

- スコア行列は、BLOSUM50を利用
- ギャップペナルティは、−5（リニアギャップ）

※スコア行列がかわると結果はどうなる？

PAM250（アミノ酸）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	−2	0	0	−2	0	0	1	−1	−1	−2	−1	−1	−3	1	1	1	−6	−3	0	0	0	0	−8
R	−2	6	0	−1	−4	1	−1	−3	2	−2	−3	3	0	−4	0	0	−1	2	−4	−2	−1	0	−1	−8
N	0	0	2	2	−4	1	1	0	2	−2	−3	1	−2	−3	0	1	0	−4	−2	−2	2	1	0	−8
D	0	−1	2	4	−5	2	3	1	1	−2	−4	0	−3	−6	−1	0	0	−7	−4	−2	3	3	−1	−8
C	−2	−4	−4	−5	12	−5	−5	−3	−3	−2	−6	−5	−5	−4	−3	0	−2	−8	0	−2	−4	−5	−3	−8
Q	0	1	1	2	−5	4	2	−1	3	−2	−2	1	−1	−5	0	−1	−1	−5	−4	−2	1	3	−1	−8
E	0	−1	1	3	−5	2	4	0	1	−2	−3	0	−2	−5	−1	0	0	−7	−4	−2	3	3	−1	−8
G	1	−3	0	1	−3	−1	0	5	−2	−3	−4	−2	−3	−5	0	1	0	−7	−5	−1	0	0	−1	−8
H	−1	2	2	1	−3	3	1	−2	6	−2	−2	0	−2	−2	0	−1	−1	−3	0	−2	1	2	−1	−8
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5	2	−2	2	1	−2	−1	0	−5	−1	4	−2	−2	−1	−8
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6	−3	4	2	−3	−3	−2	−2	−1	2	−3	−3	−1	−8
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5	0	−5	−1	0	0	−3	−4	−2	1	0	−1	−8
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6	0	−2	−2	−1	−4	−2	2	−2	−2	−1	−8
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9	−5	−3	−3	0	7	−1	−4	−5	−2	−8
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6	1	0	−6	−5	−1	−1	0	−1	−8
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2	1	−2	−3	−1	0	0	0	−8
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3	−5	−3	0	0	−1	0	−8
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17	0	−6	−5	−6	−4	−8
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	−2	−3	−4	−2	−8
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4	−2	−2	−1	−8
B	0	−1	2	3	−4	1	3	0	1	−2	−3	1	−2	−4	−1	0	0	−5	−3	−2	3	2	−1	−8
Z	0	0	1	3	−5	3	3	0	2	−2	−3	0	−2	−5	0	0	−1	−6	−4	−2	2	3	−1	−8
X	0	−1	0	−1	−3	−1	−1	−1	−1	−1	−1	−1	−1	−2	−1	0	0	−4	−2	−1	−1	−1	−1	−8
*	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	1

相同性探索ツール

アラインメントは計算が大変

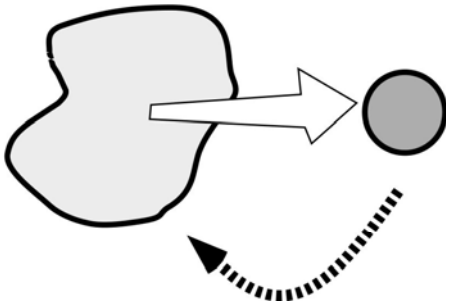
⇨　Web上で計算サービス

- BLAST　相同性探索のためのプログラム
  http://blast.ddbj.nig.ac.jp/

標本と母集団

母集団：調査（考察）対象の全体

標本：母集団から取り出した一部分　サンプリング



標本の情報から母集団の情報を推測

統計的推定

推定：

母集団から無作為に獲得した標本 $\{X_1, X_2, \cdots, X_n\}$ から
母集団のパラメータを推測する

標本確率変数 X_i：母集団分布と同じ確率分布　互いに独立

母集団のパラメータ

母平均：母集団の期待値　$E[X]$

母分散：母集団の分散　$V[X]$

統計量

統計量：標本確率変数の関数として表わされる確率変数

$$\{X_1, X_2, \cdots, X_n\} \quad \Rightarrow \quad Y = \varphi(X_1, X_2, \cdots, X_n)$$

よく利用される統計量

- 標本平均　$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 標本分散　$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- 不偏分散　$U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

統計量の期待値・分散

- 標本平均　$E\left[\overline{X}\right]=E[X]$　$V\left[\overline{X}\right]=\dfrac{V[X]}{n}$
- 標本分散　$E[S^2]=\dfrac{n-1}{n}V[X]$

$$V[S^2]=\frac{(n-1)^2}{n^3}\left(E[(X-E[X])^4]-\frac{n-3}{n-1}V[X]^2\right)$$

- 不偏分散　$E[U^2]=V[X]$

$$V[U^2]=\frac{1}{n}\left(E[(X-E[X])^4]-\frac{n-3}{n-1}V[X]^2\right)$$

良い推定量の条件

パラメータ θ の推定量 $\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)$ に対して

- 不偏性　中心が θ からズレていない　$E\left[\hat{\theta}\right]=\theta$
- 最小分散性　分散が小さいほうが望ましい
- 一致性　サンプルサイズが大きければ θ に近づく

標本平均：母平均に対して「不偏性」「一致性」
不偏分散：母分散に対して「不偏性」「一致性」

統計的検定

二つの標本が、同じ性質を持っているのか？

“同じだ”という予想 ⇒ 帰無仮説

帰無仮説が正しいかを検証 ⇒ 統計的検定

２種類の間違いの可能性

- 帰無仮説が正しいにも関わらず間違いと判定
- 帰無仮説が間違いにも関わらず正しいと判定

２番目の間違いはコントロールしにくい
１番目の間違いの可能性を条件として固定
⇒ 有意水準

位置の違いの検定①

正規分布に従う二つの標本が、同じ母平均を持つか？

$$x_1,\cdots,x_m \quad y_1,\cdots,y_n$$

- 母分散が等しい場合 ⇒ t検定

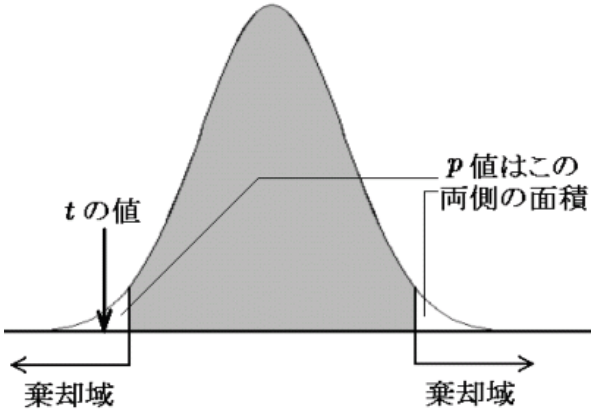
$$T=\frac{\bar{x}-\bar{y}}{\sqrt{\dfrac{\sum(x_i-\bar{x})^2+\sum(y_i-\bar{y})^2}{m+n-2}}}\times\frac{1}{\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}$$

自由度m+n-2のt分布

位置の違いの検定②

計算したTの値が

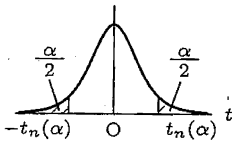
- 棄却域に入っていれば、“棄却する” ⇒ 位置は異なる
- 棄却域に入っていなければ、“棄却できない”



t分布 数表

t 分布表

$P(|T| \geqq t_n(\alpha)) = \alpha$ となる $t_n(\alpha)$ の値



n \ α	.500	.400	.300	.200	.100	.050	.020	.010	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.60
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.92
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

位置の違いの検定③

- 母分散が等しいとは限らない場合 ⇒ Welch検定

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{m(m-1)} + \dfrac{\sum (y_i - \bar{y})^2}{n(n-1)}}}$$

自由度 f のt分布

$$f = \frac{\left(\sum (x_i - \bar{x})^2 \Big/ m(m-1) + \sum (y_i - \bar{y})^2 \Big/ n(n-1) \right)^2}{\left(\sum (x_i - \bar{x})^2 \Big/ m(m-1) \right)^2 \Big/ (m-1) + \left(\sum (y_i - \bar{y})^2 \Big/ n(n-1) \right)^2 \Big/ (n-1)}$$

位置の違いの検定演習

次の２つのデータに対して、位置の違いがあるか検討せよ。

x	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
y	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4

有意水準：0.05
正規分布を仮定

1. Excelでデータ入力（csv出力）
2. Rで読み込み
3. 検定実行

バラツキの違いの検定

正規分布に従う二つの標本が、同じ母分散を持つか？

- F検定

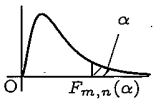
$$T = \frac{\sum (x_i - \bar{x})^2 \Big/ (m-1)}{\sum (y_i - \bar{y})^2 \Big/ (n-1)}$$

自由度(m-1,n-1)のF分布

F分布 数表

F 分布表

$P(F \geqq F_{m,n}(\alpha)) = \alpha$ となる $F_{m,n}(\alpha)$ の値 ($\alpha = 0.05$)



n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73

順位を利用した検定

t検定やF検定は分布に関する“仮定”が必要

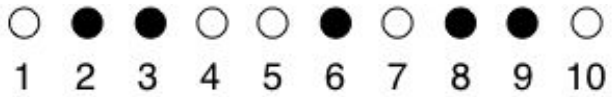
- 正規分布である前提

分布の型を条件にしないでズレの有無を検定したい

⇩

順位を利用

- ２群を大きさ順に並べたとき、
  ズレがなければ順位は偏らないはず！



Wilcoxon順位和検定

二つの標本間にズレはないか？

$$x_1,\cdots,x_m \quad y_1,\cdots,y_n$$

二標本を小さい順に並べ、一方の順位を足し合わせる

$$R_x = \sum_{i=1}^{m} R_i \quad R_i : \text{順位}$$

Rxが大きすぎたり小さすぎたりすると帰無仮説が疑わしい

Rxより極端な値をとる確率が有意水準以下なら棄却

相関

二つの項目の関連性は？

傾向の類似性

⇨　規則正しさを持って増減をともにする関係

☞　相関関係　正の相関、負の相関

※因果関係があるとは限らない

相関係数

Pearson積率相関係数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- −1から+1までの値
- +1に近いとき　⇨　正の相関
- −1に近いとき　⇨　負の相関
- 0に近いとき　⇨　無相関　関連性がない？

ノンパラメトリック相関係数

順位を利用した相関係数

- 分布が歪んでいたり、外れ値がある場合など

- Spearman順位相関係数

$$\rho = \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n(n^2-1)}$$

- Kendall順位相関係数

$$\tau = \frac{(A-B)}{\frac{n(n-1)}{2}}$$

順位の大小関係
A：一致する組数
B：不一致組数

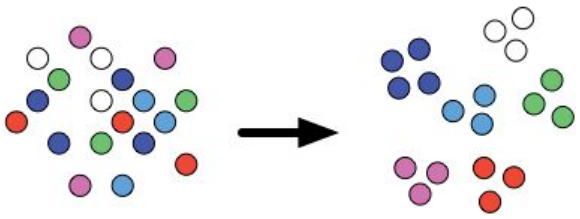
相関係数演習

遺伝子発現データの各処置間の関連性

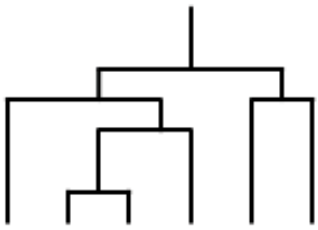
- 相関係数（Pearson、Spearman、Kendall）
- 対数変換するとどう変わる？
- プロットして確認

クラスタリング

類似している個体や変数を分類し、グループ化
（クラスタ）



階層構造があれば
⇨　デンドログラム



遺伝子発現解析
　各遺伝子の変化状況を測定し
　類似する遺伝子群を発見

似ている度合い

- 個体・変数がどれだけ離れているか　⇨　距離
  距離が小さいと、よく似ている
  - ユークリッド距離
  - 標準ユークリッド距離
- 個体・変数が類似している度合い　⇨　類似度
  類似度が大きいと、よく似ている
  - ピアソン相関係数
  - コサイン類似度

距離

距離を測る対象要素をn次元ベクトル表現

$$\mathbf{x} = (x_1, \cdots, x_n) \quad \mathbf{y} = (y_1, \cdots, y_n)$$

- ユークリッド距離

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- 標準ユークリッド距離

各次元の距離が等質でないとき

$$d = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{\sigma_i} \right)^2}$$

σは各次元の標準偏差

類似度

- ピアソン相関係数

二つのベクトルを確率変数とみて相関係数を類似度

$$sim = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^{n} (y_i - \mu_y)^2}}$$

μは各ベクトルの平均

- コサイン係数

ベクトルの向きの近さを類似度

$$sim = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

遺伝子発現データによく利用

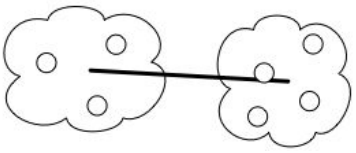
階層的クラスタ構成

n個の個体・変数間の非類似度行列から構成

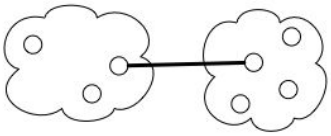
1. n個のクラスタからスタート
2. 最も類似性の高いクラスタを併合
3. クラスタが1つになれば終了
   そうでなければ４へ
4. ２で作られたクラスタと他クラスタの非類似度を計算して、非類似度行列を更新し、２へ

クラスタ併合ルール①

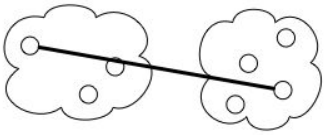
- 重心法
  各クラスタの重心間距離



- 最短距離法
  ２つのクラスタ内データで、最短距離の組み合せ



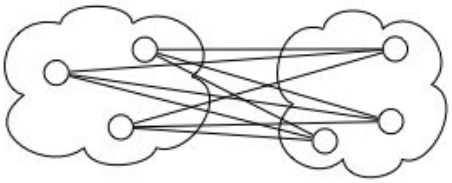
- 最長距離法
  ２つのクラスタ内データで、最長距離の組み合せ



クラスタ併合ルール②

- 群平均法

  ２つのクラスタ内データの全組み合せの平均距離



- Ward法

  クラスタ内のデータの平方和を最小化
  よく利用されるが、ユークリッド距離に限定

$$D(C_i,C_j) = \sum_{k=1}^{\#C_i+\#C_j} d(\mathbf{x}_k,\mathbf{m}_{i\oplus j}) - \sum_{k=1}^{\#C_i} d(\mathbf{x}_k,\mathbf{m}_i) - \sum_{k=1}^{\#C_j} d(\mathbf{x}_k,\mathbf{m}_j)$$

クラスタリングの留意点

- 数値の確認

  データの特性を考える
  必要な変数変換（対数、比など）

- 距離の選択

  データ特性に合致した距離を選択

- 併合ルールの選択

  生成されるクラスタの特性を考える

距離と併合ルール組み合せを複数実行し、
結果を比較検討することが必要

クラスタリング演習

目的：遺伝子発現データの遺伝子クラスタリング
出力：遺伝子のデンドログラム

手順：

1. ExcelデータをRで解析できるように変換(csv)
2. Rで読み込み、対数変換
3. 距離と併合ルールを決めてクラスタリング実行
4. 結果をデンドログラムに出力
5. どの遺伝子がどのグループに入っているか？
   クラスタリング結果の違いは？

まとめ

- 類似性に関する解析方法を紹介
  - 配列アラインメント
  - クラスタリング
  - 分布特徴量の比較
  - 相関係数
- 便利な解析ツール
  - Webサービス
    - Web API for Biology (WABI) http://xml.nig.ac.jp/
  - 統計解析言語 R

R実行メモ

【準備】

作業ディレクトリを設置　"N:¥bio"

データファイル"gene.csv"を"N:¥bio"に置く

【R起動】

・デスクトップ上のアイコンを使って起動する。

・作業ディレクトリの設定

　メニュー/ファイル/ディレクトリの変更を選択し、"N:¥bio"を指定する。

・作業ディレクトリの確認

　R画面上のコンソールで次のコマンドを入力する。

　getwd()

【解析手順】

1．データセットを準備する。

　Excel等で作成したファイルを読み込む。(csv形式が便利)

2．解析に必要なデータ整形をする。(対数変換など)

3．解析実行

これらはコンソール上でキーボードからコマンド入力によって行なう。

【データ入力】

・csvファイルを読み込む。

　csvファイル名：gene.csv、データセット名：genedata

　csvファイルに各レコードラベルが付いているときはrow.namesを指定

　genedata <- **read.table("gene.csv", header=T, sep=",", row.names="Name")**

・入力後にデータセットの内容を確認する。

　genedata

・データセット内の各変数を指定するときは$で区切る。

　genedata$EM8Z309

【特徴量表示】

・簡単のために、genedataの4変数を独立したデータセットに格納しておく。

　genedata1 <- genedata$EM8Z309

　genedata2 <- genedata$EM8Z310

　genedata3 <- genedata$EM8Z313

　genedata4 <- genedata$EM7ZC11

・標本平均値

mean(genedata1)

・中央値

median(genedata1)

・最大値

max(genedata1)

・最小値

min(genedata1)

・第１四分位数

quantile(genedata1, prob=0.25)

・第３四分位数

quantile(genedata1, prob=0.75)

・標本範囲

range(genedata1)

・標本分散値

不偏分散値は var() であるが、標本分散値は(n-1)/n 倍しなければならない。

var(genedata1)*(length(genedata1)-1)/length(genedata1)

・標本標準偏差

sqrt(var(genedata1)*(length(genedata1)-1)/length(genedata1))

・四分位偏差

IQR(genedata1)

【グラフ表示】

・ヒストグラム

hist(genedata1)

・散布図

plot(genedata)

・ボックスチャート

boxplot(genedata, range=0)

【二項分布の確率関数】

・X のとりうる値を準備する。

x <- seq(0, 10, 1)

・棒グラフで表示する。

barplot(dbinom(x, 10, 0.3))

【正規分布の確率密度関数】

・Xのとりうる値を準備する。

```
x <- seq(-10,10,0.1)
```

・曲線で表示する。

```
curve(dnorm(x,0,1),from=-10,to=10)
```

【統計的検定】

（Excelでデータを入力し、csv形式で書き出しておく。data.csv）

・t検定（二標本両側検定　片側検定の場合は"greater"、"less"を指定する。

```
t.test(dataset$x, dataset$y, alternative="two.sided", var.equal=T)
```

・ウェルチ検定

```
t.test(dataset$x, dataset$y, alternative="two.sided", var.equal=F)
```

・F検定

```
var.test(dataset$x, dataset$y)
```

・Wilcoxon順位和検定

```
wilcox.test(dataset$x, dataset$y, alternative="two.sided")
```

【相関係数】

・"pearson"、"spearman"、"kendall"が選択できる。

複数変数を含むデータの場合は、全組み合せの相関係数を算出する。

```
cor(dataset, method="pearson")
```

【クラスタリング】

・追加パッケージのインストールをする。

メニュー/パッケージ/CRANミラーサイトの設定を選択　Japan(Hyogo)を選択

メニュー/パッケージ/パッケージのインストールを選択　amapを選択

・追加したパッケージを有効化する。

```
library(amap)
```

・対数変換する。

```
logdata <- log(genedata, base=2)
```

・データの標準化を実行する。（各変数の平均を0分散を1に変換）

```
normaldata <- scale(logdata)
```

・距離行列を生成する。

"euclidean"（ユークリッド距離）、"maximum"（最大距離）、"manhattan"（マンハッタン距離）、"correlation"（ピアソン相関）、"pearson"（コサイン係数）等が選択できる。

```
distmat <- Dist(logdata, method="euclidean")
```

・クラスタリングを実行する。

クラスタ間の距離は"complete"（最長距離）、"ward"、"single"（最短距離）、"average"（群平均）、"centroid"（重心）等が選択できる。

```
result <- hclust(distmat, method="average")
```

・結果をデンドログラム表示する。

```
plot(result, hang=-1, main="Clustering", sub="log&average ",
     xlab="", ylab="")
```

・デンドログラム上で指定した個数のグループに分ける。

```
rect.hclust=(result, k=10)
```

・指定したグループ数に分けたとき、各要素がどのグループに属しているか表示する。

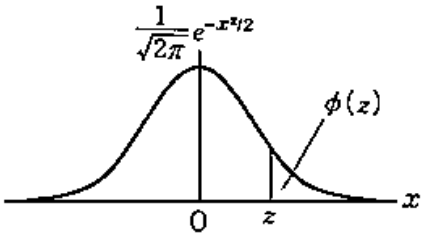
```
cutree(result, k=10)
```

・pdfファイルにグラフを描画するときは、最初に出力先ファイル名を指定して、描画し、閉じる。

```
pdf("result.pdf")
plot(result, hang=-1, main="Clustering", sub="log&average", xlab="",
     ylab="")
dev.off()
```

・ヒートマップを作成する。

```
d1 <- Dist(logdata, method="euclidean")
d2 <- Dist(t(logdata), method="euclidean")
c1 <- hclust(d1, method="average")
c1 <- hclust(d2, method="average")
heatmap(as.matrix(logdata), Colv=as.dendrogram(c2),
        Rowv=as.dendrogram(c1), scale="none", col=cm.colors(256),
        main="log&average", margin=c(8, 10))
```



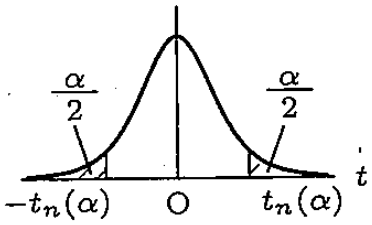
附表2　正規分布　$\phi(z)=\int_z^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$ の値

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003

スネデカー，コクラン(畑村，奥野，津村訳)：『統計的方法』(原書第6版)，(岩波書店，1972)および薩摩順吉：『確率・統計』(理工系の数学入門コース7)(岩波書店，1989)をもとに，さらに追加して作製.

t 分布表

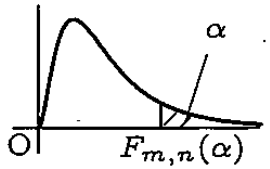
$P(|T| \geqq t_n(\alpha)) = \alpha$ となる $t_n(\alpha)$ の値



n \ α	.500	.400	.300	.200	.100	.050	.020	.010	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.60
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.92
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
31	.682	.853	1.054	1.309	1.696	2.040	2.453	2.744	3.633
32	.682	.853	1.054	1.309	1.694	2.037	2.449	2.738	3.622
33	.682	.853	1.053	1.308	1.692	2.035	2.445	2.733	3.611
34	.682	.852	1.052	1.307	1.691	2.032	2.441	2.728	3.601
35	.682	.852	1.052	1.306	1.690	2.030	2.438	2.724	3.591
36	.681	.852	1.052	1.306	1.688	2.028	2.434	2.719	3.582
37	.681	.851	1.051	1.305	1.687	2.026	2.431	2.715	3.574
38	.681	.851	1.051	1.304	1.686	2.024	2.429	2.712	3.566
39	.681	.851	1.050	1.304	1.685	2.023	2.426	2.708	3.558
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
45	.680	.850	1.049	1.301	1.679	2.014	2.412	2.690	3.520
50	.679	.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
240	.676	.843	1.039	1.285	1.651	1.970	2.342	2.596	3.332
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

F 分布表

$P(F \geqq F_{m,n}(\alpha)) = \alpha$ となる $F_{m,n}(\alpha)$ の値 ($\alpha = 0.05$)



n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

BLOSUM50

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

統計解析言語システムRの紹介

琉球大学工学部情報工学科　　岡崎 威生
okazaki@ie.u-ryukyu.ac.jp

1. はじめに

統計解析を支援するソフトウェアは、SPSS や SAS、BMDP など汎用目的に開発され世界的に広く使われているものから、特定の解析に限定して開発されたもの、あるいは Excel マクロにより実行するものなど様々である。統計解析を専門とする研究者は別として、多くの利用者は手段としての統計解析を効率よくかつ正確に実行するためにこれらの支援ソフトウェアを利用している。これまで GUI などが充実して操作しやすく計算結果の信頼性を求めると、商用ソフトウェアが選択されることが殆どであった。

オークランド大学の R.Ihaka とハーバード大学の R.Gentleman により統計学用のインタープリタ言語Rの開発がスタートした。Rはベル研究所で開発された統計解析言語Sに操作性を含めた外見を似せることを目的として設計されたものである。1995 年に GNU ライセンスのもとでオープンソースとして国際共同開発されるようになった。それ以降、全世界の統計学研究者や利用者らによる改良や拡張が施され、現在最も多くの最も先端の解析手法に対応したソフトウェアであると言っても過言ではない。

本報告では、R を利用した簡単なデータ記述と解析の実行手続きを、MacOSX 上での事例を使いながら紹介する。

2. Rの機能

Rには次の機能が備わっている。

・データ操作
・基本数学演算
・データ解析
・グラフィック
・簡潔なプログラミング言語

データ操作においては、SAS や SPSS などで作成されたデータセットや Excel ワークシートを直接利用することを可能にする。各種データベースとのインターフェースも準備されている。基本数学演算においては、行列操作や関数変換など解析の前処理に必要な操作を可能にする。データ解析はメインパートで、種々の解析手法を関数として準備している。グラフィック機能はデータ記述や解析結果の図示を行うもので、各種グラフ生成を行う高水準作図関数と、点や線などを描画する低水準作図関数が準備されている。画面上だけでなく、JPEG や EPS、PNG など多様な形式でファイルとして出力可能である。プログラミング言語機能に関しては、本来のインタラクティブな操作の繰り返しによる一連の手続きを、あらかじめ準備したシナリオに従い連続して実行させることを可能にする。また、新しい

統計解析手法の R 上での実装のために不可欠な機能でもある。

3．R の入手とインストール

R は The R Project for Statistical Computing (http://www.r-project.org/) というインターネット上のコミュニティによって管理されている。最新版のソースや、Linux や MacOSX や Windows 上で利用できるインストーラ付バイナリは、The Comprehensive R Archive Network (http://cran.r-project.org/) において配布されている。最近のバージョンでは日本語対応もされており、日本語 OS 上で利用する場合はメニュー等も日本語で表示されるようになった。本報告書執筆の時点での最新版は R-2.6.1 になる。

MacOSX (10.5) 上でのインストール例を示す。上記の CRAN サイトから R-2.6.1.dmg をダウンロードし、ダブルクリックするとディスクイメージが展開される。

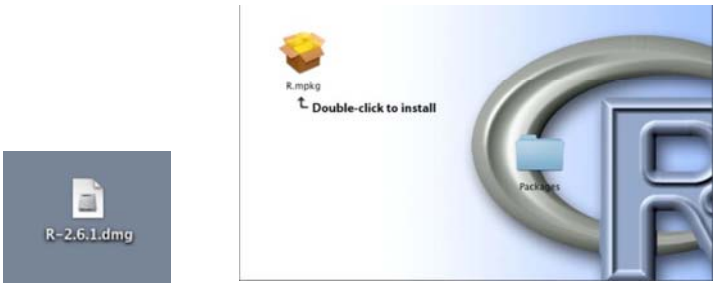


図 1. MacOSX 用インストールファイルと展開したディスクイメージ

ディスクイメージの中に含まれる R.mpkg をダブルクリックすることでインストールが始まる。途中で管理者パスワードの入力が必要になる。インストールが終了すると、アプリケーションフォルダ内に R の起動アイコンが表示される。このアイコンをダブルクリックすることで R が起動する。

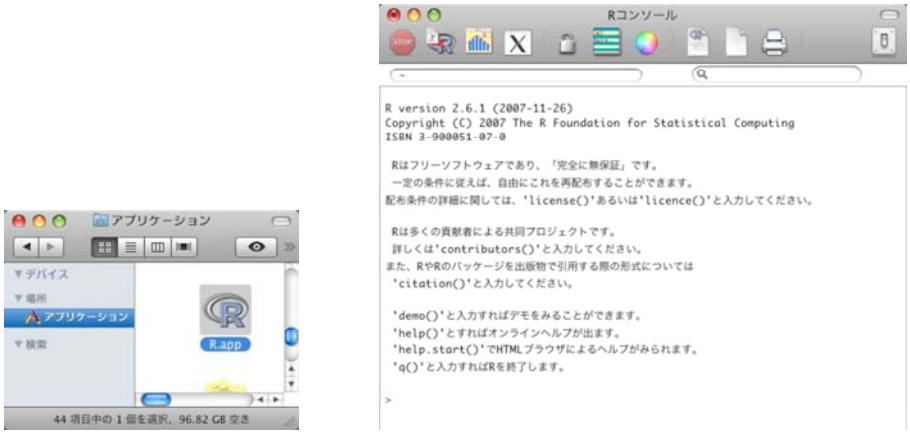


図 2. インストールされた R と起動時の画面

環境設定は特にカスタマイズする必要はないが、初期作業ディレクトリを特定のフォルダにしておくと便利であろう。ここではホームディレクトリ直下に R という名前のフォルダを作成し指定している。

図 3. 初期作業ディレクトリの指定

初期作業ディレクトリは、R が入出力するソースやデータのデフォルトの場所になる。

4. データ入力と操作

R にはサンプルデータセットが組み込まれている。このデータセット名と概説を表示するために、次のようにコンソールから入力する。

```
> library(help="datasets")
```

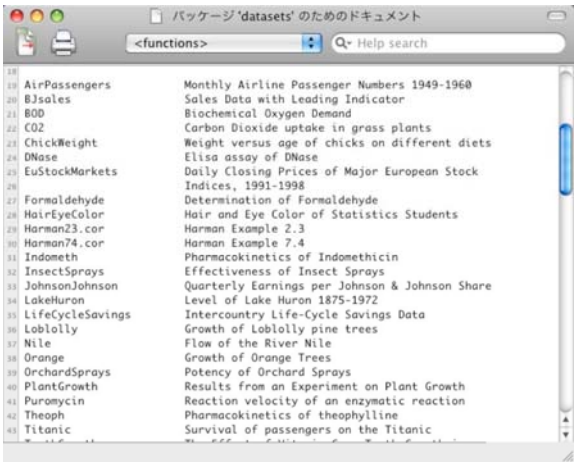


図 4. 準備されているデータセットの一覧

例えば分類問題でよく使われる iris データは、次のように入力することで表示できる。

```
> iris
```

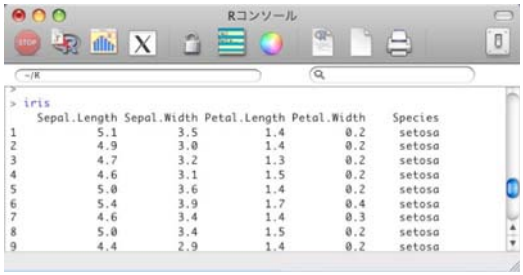
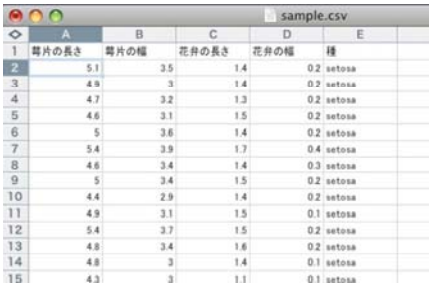


図 5. iris データの閲覧

このデータは左から、ID、萼片の長さと幅、花弁の長さと幅、種名を示している。

独自のデータセットを生成するには、R コンソールからプログラムで入力する方法と、テキストデータなどを準備しておいて読み込ませる方法がある。ここでは後者の例を示す。まず Excel など表計算ソフトを利用してデータを入力した後、csv 形式で保存する。ここではファイル名を sample.csv としておく。このファイルを初期作業ディレクトリに置く。



sample.csv

	A	B	C	D	E
1	萼片の長さ	萼片の幅	花弁の長さ	花弁の幅	種
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa

図 6. Excel でのデータ入力

このデータを R に取り込むため、次のように入力する。

```
> infile <- file("sample.csv", open="r", encoding="shift-jis")
> data <- read.table(infile, header=T, sep=",")
> close(infile)
```

これで、data というデータフレームに sample.csv ファイルの内容が格納された。encoding は、MacOSX でのデフォルトは utf-8、Windows でのデフォルトは CP932 である。ヘッダ部は変数名となり、データフレーム名$変数名で参照できる。例えば、花弁の長さのみを取り出したいときは次のように入力する。

```
> data$花弁の長さ
```

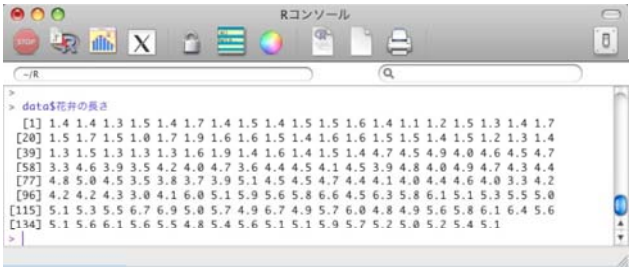


図 7.　変数「花弁の長さ」のみを書き出した結果

変数名に日本語の利用は認められるが、名前の先頭に数値を使ったり、名前の間に空白を含めることはできない。なお、欠測値は NA と表示される。

出力結果を R 以外で利用する場合、コンソール表示のコピー＆ペーストでも可能であるが、R から直接外部ファイルに書き出すこともできる。このときは write.table 関数を利用

する。

```
> outfile <- file("result.txt", open="w", encoding="shift-jis")
> write.table(data, outfile, header=T, sep=",")
> close(outfile)
```

初期作業ディレクトリに、ファイルが生成される。

5. データ記述

データフレームの簡単な統計量を調べるには、次のように入力する。

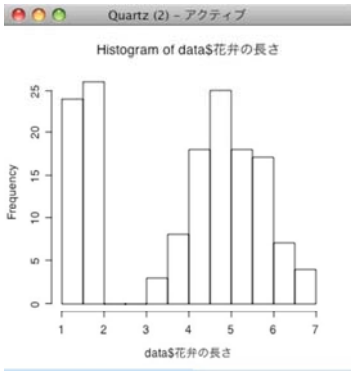
```
> summary(data)
```

図 8. データフレーム data の簡易集計結果

数値変数に対して、順に、最小値・第 1 四分位数・中央値・平均値・第 3 四分位数・最大値を表示している。カテゴリ変数には、それぞれのカテゴリの度数を表示する。

数値変数に対してヒストグラムを描画するには、次のように入力する。

カテゴリ変数に対して棒グラフを描画するには、次のように入力する。



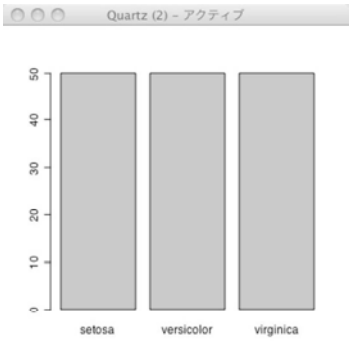


図 9. 花弁の長さのヒストグラム　　図 10. 種の棒グラフ

グラフはコンソールと別の新しいウィンドウに描画される。ファイルとして書き出したい場合は、次のように入力する。

```
> pdf(file="graphfile.pdf", width=400, height=400)
> barplot(table(data$種))
> dev.off()
```

二変量間の相関係数は cor.test 関数を利用して求めることができる。手法として、ピアソンの積率相関係数（pearson）とケンドールの順位相関係数（kendall）、スピアマンの順位相関係数（spearman）が選択できる。

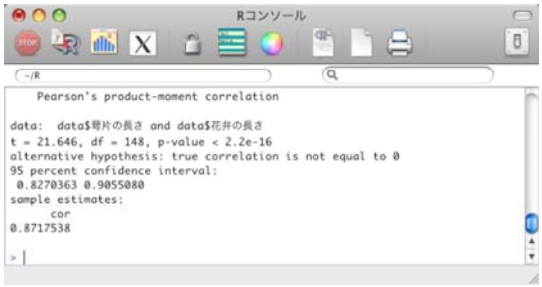


図 11. 変数「萼片の長さ」と「花弁の長さ」のピアソン積率相関係数

二変数の相関関係を散布図で描画するには、次のように入力する。

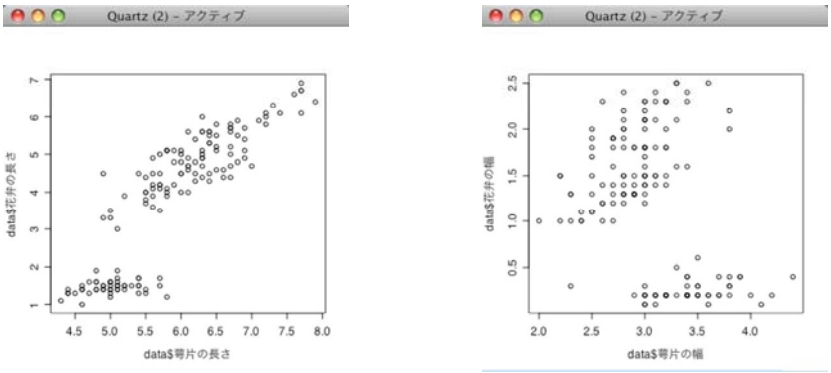


図 12. 二変数間の散布図

6. 線形判別分析の実行例

線形判別分析を行う場合は、MASS ライブラリに含まれる lda 関数を利用する。ここでは前述の iris データを利用し、4 つの特徴を説明変数にし、3 種類の種を判別することとする。まず、説明変数とグループ変数の形式を lda 関数が利用できるように変換する。

```
> 萼片長 <- data[,1]
> 萼片幅 <- data[,2]
> 花弁長 <- data[,3]
> 花弁幅 <- data[,4]
> 特徴1 <- cbind(萼片長, 萼片幅, 花弁長, 花弁幅)
> 特徴2 <- cbind(萼片幅, 花弁長, 花弁幅)
> 種 <- data[,5]
```

特徴 1 は全ての特徴の組み合わせ、特徴 2 は萼片の長さを除いたものである。次に線形判別分析を実行する。

```
> result1 <- lda(特徴1, 種)
> print(result1)
```

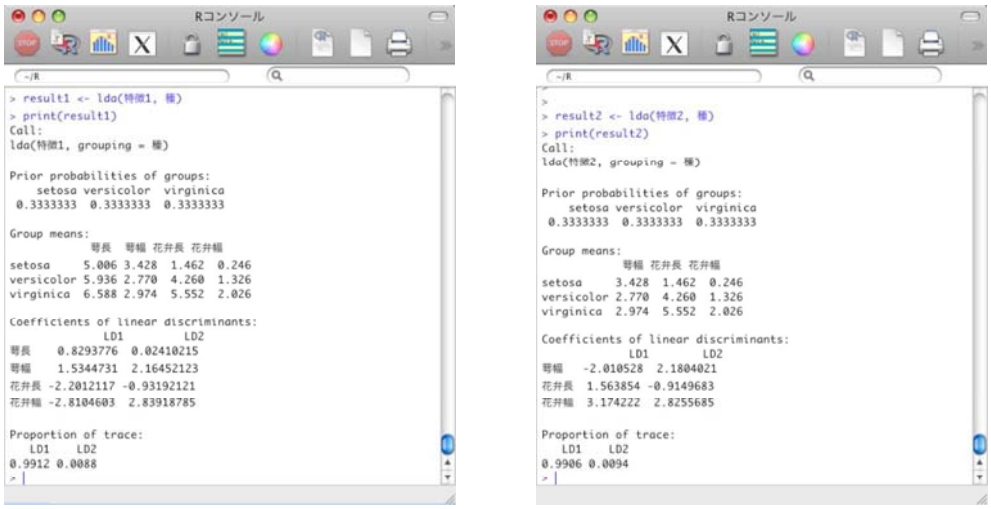


図 13. 線形判別分析の結果

この線形判別分析によって得られた判別ルールを利用し、特徴の大きさから種を予測する。predict 関数が予測を実行する。予測された種と実際の種の比較を table 関数により実行する。

```
> presult1 <- predict(result1)
> table(種, presult1$class)
```

図 14. 判別ルールの評価

両方の結果とも setosa が setosa と判別されたケースが 50、versicolor が versicolor と判別されたケースが 48、virginica が virginica と判別されたケースが 49 であることを示している。

7. おわりに

本報告では、統計解析言語 R の持つ機能の概略と簡単な使用法を紹介した。他の統計解析ソフトウェアの持つ殆どの機能を実現できる R の最大の特徴はオープンソースであるということであろう。多くのコントリビュータによって最新の解析手法が組み込まれ、その利用方法や事例もインターネット上で公開されている。そして何より無償で利用できるということである。R のインストール時には標準的な解析手法だけが、CRAN サイトで公開されているパッケージはシステムに含まれるパッケージインストーラを使うことで容易に追加できる。以前は Web サイトが唯一の情報入手源であったが、最近 R 解説書籍が多く出版されており初学者にも取り組みやすくなってきた。これまで他のソフトウェアを利用してきたユーザが、これを機会に R を利用し更なる R の進化に寄与してくれることを切に望む。

参考文献・URL

[1] "An Introduction on R", W.N.Venables, D.M.Smith and R Development Core Team, Network Thory Ltd. 2002

[2] "The R Tips データ解析環境 R の基本技・グラフィックスの活用集", 舟尾暢男著, 九天社 2005

[3] The R Project for Statistical Computing http://cran.r-project.org/

[4] The Comprehensive R Archive Network http://cran.r-project.org/

[5] RWiki http://www.okada.jp.org/RWiki/

[6] R による統計処理（青木繁伸）http://aoki2.si.gunma-u.ac.jp/R/

【データの代表値】10 ページ
・遺伝子発現データ（gene.csv）を用いて、EM8Z309、EM8Z310、EM8Z313、EM7ZC11 の代表値（標本平均値、中央値、最大値、最小値、第 1 四分位数、第 3 四分位数）を計算せよ。

【データのバラツキ】12 ページ
・遺伝子発現データ（gene.csv）を用いて、EM8Z309、EM8Z310、EM8Z313、EM7ZC11 のバラツキ指標値（標本範囲、標本分散、標本標準偏差、四分位偏差）を計算せよ。

【グラフ表現】13 ページ
・遺伝子発現データ（gene.csv）を用いて、EM8Z309、EM8Z310、EM8Z313、EM7ZC11 のヒストグラムを作成せよ。
・遺伝子発現データ（gene.csv）を用いて、全ペアの散布図を作成せよ。

・遺伝子発現データ（gene.csv）を用いて、EM8Z309、EM8Z310、EM8Z313、EM7ZC11 のボックスチャートを計算せよ。

【確率とは】14 ページ
・サイコロを無作為に 2 回ふるとき、2 回目に偶数の目が出る確率を求めよ。
・サイコロを無作為に 5 回ふるとき、「1」の目が 3 回出る確率を求めよ。

【条件付き確率】16 ページ
・ジョーカーを除く 52 枚のトランプから無作為に 1 枚カードを選ぶとき、A : ハートのカード、B : 絵札、として $P(B|A)$ を求めよ。

【ベイズの定理】17 ページ
・ある DNA 配列において、A・T・G・C の割合はそれぞれ、40%、30%、20%、10%であった。また、各塩基に対する読取りミスの発生する可能性は、7%、5%、4%、5%である。この配列に対して読み取りミスが発生したとき、その塩基が A である確率を求めよ。
・ある感染症のために 1 次検査を行ない、その結果に応じて要再検査となる場合を考 える。感染症の発生する確率は 0.01 とする。ここで感染症である場合は必ず 1 次検査で検出されることを仮定しておく。感染症に罹患していないものが要再検査となる場合もあり、その確率を 0.05 とする。感染症に罹患している事象を A、1 次検査の結果で要再検査となる事象を B とおくとき、$P(A|B)$ を求めよ。

【事象の独立】18 ページ

・ジョーカーを除く 52 枚のトランプから無作為に 1 枚カードを選ぶとき、A : ハートのカード、B : 絵札、としたとき事象 A と B は独立かどうか調べよ。

【期待値】20 ページ

・P(X=1) =0. 6、P(X=2) =0. 3、P(X=3) =0. 1 のとき、E[X] を計算せよ。

・バス停についてからバスの待ち時間を X とし、X の密度関数を

$$f(x) = \begin{cases} \frac{1}{10} & (0 < x \leq 10) \\ 0 & (x \leq 0, 10 < x) \end{cases}$$

とするとき E[X] を計算せよ。

【分散・標準偏差】21 ページ

・P(X= ﹣1) =0. 5、P(X=0) =0. 2、P(X=3) =0. 3 のとき、V[X] を計算せよ。

・E[X] =3、V[X] =5 のとき、Y=- 2X+3 に対して E[Y] と V[Y] を計算せよ。

【二項分布】23 ページ

・サイコロを無作為に 10 回ふったとき、「1」の目が 8 回以上でる確率を計算せよ。

・二項分布において、n=10、p=0. 3 のときの確率関数のグラフを作成せよ。

【正規分布】24 ページ

・正規分布 N(1, 4) の確率密度関数のグラフを作成せよ。

・確率変数 X が正規分布 N(3, 4) に従うとき、Y=(﹣X+2) / 3 の確率分布を求めよ。

・確率変数 X が正規分布 N(100, 16) に従うとき、P(94! X! 109) を計算せよ。

【良い推定量の条件】30 ページ

・母平均に対する推定量としての標本平均が不偏性を持つことを証明せよ。

3.8　配列と機能

3.8.1 概要

講師：有田 正規／東京大学理学部生物情報学科准教授

日時：2012 年 9 月 12 日　10:20～／12:50～／14:40 ～

場所：琉球大学総合情報処理センター

内容：DNA やタンパク質など生体分子の配列の相同性と機能について、BLAST による相同配列の検索や構造ドメイン、二次構造予想など配列の二次構造予測の実習を行う。また、酵素や代謝、タンパク質の相互作用など、KEGG や STRING の WEB ツールを活用して、それらのネットワークについて実習する。

3.8.2 準備

特筆事項なし

3.8.3 配布物

・講義スライド

・理系総合のための生命科学（参考資料／羊土社出版予定）

3.8.4 課題

※講義スライドに含む

3.8.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造(二次構造、三次構造、四次構造)、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

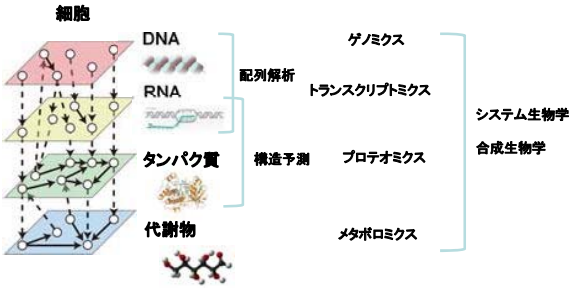
3.8.6時間割

時間	内容
10:20～	<u>イントロ</u> ・ バイオインフォマティクス概要 ゲノミクス トランスクリプトミクス プロテオミクス メタボロミクス
10:27～	<u>配列と遺伝子機能</u> ・ ゲノム解析の現状 ・ ショットガンシーケンス ・ シーケンスのカバー率 ・ 次世代シーケンス（NGS） ・ BLAST ・ DDBJ／EMBL／GenBank
11:41～	【演習】 Tyranosaurus rex の配列を BLAST 検索する ・ E-Value と P-value ・ BLAST Family Nucleotide BLAST、blastx、psi-blast など ・ アミノ酸置換行列 【課題】 課題へ取り組む
12:05～	<u>昼食</u>
12:50～	<u>課題の解説</u> <u>タンパク質の配列と機能</u> ・ タンパク質の立体構造 ・ 立体構造決定法 Fold It! ・ 立体構造データベースの分類 PDB、SCOP、CATH, Uniprot

13:28～ 13:40～ 14:02～	【演習】 タンパク質立体構造第一号である（ミオ）グロビンのスーパーファミリーを検索する <u>解析方法</u> ・ マルチプルアライメント ・ タンパク質の機能分類 【演習】 グロビンの機能解析／Gene Ontology ・ ヘモグロビンの仕組み ・ アロステリック効果 【課題】 課題へ取り組む
	休憩
14:42～ 15:40～ 15:57	<u>ネットワークという考え方</u> ・ 生物学におけるネットワーク 代謝ネットワーク、タンパク質相互作用ネットワーク 進化や遺伝子のネットワーク ・ ネットワークの定義 ・ 6 次の隔たり ・ ネットワークの特徴 スモールワールド、スケールフリー ・ べき分布と指数分布 ・ 平均頂点間距離、クラスター係数 ・ ネットワークのキーワード スケールフリーネットワーク、ネットワークモチーフ ・ ネットワークのモジュール ・ フィードバック ・ フィードフォワード ・ ネットワーク研究について

バイオインフォマティクス
スタンダードコース
（配列と機能１）

バイオインフォマティクスの概要



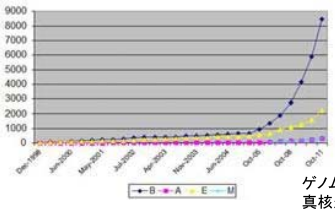
アウトライン

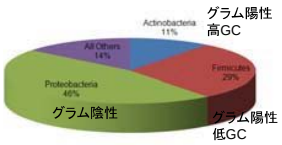
1. 配列と遺伝子機能
   - ゲノムのシーケンシング
   - 配列のデータベース
   - 配列の解析ツール
2. タンパク質のデータベース
   - タンパク質立体構造
   - 立体構造のデータベース
   - 構造と機能（モチーフ）
3. ネットワークという考え方
   - ネットワークの特徴（スケールフリー）
   - ネットワークモチーフ

ゲノム解析の現状

GenomesOnLine Database (GOLD)
http://www.genomesonline.org/

生物は、真核生物と原核生物に分けられる。
B バクテリア（細菌），A アーキア（古細菌），
E 真核生物（その中の哺乳類　M）





ゲノム解析で圧倒的に数が多いのはバクテリア。
真核生物の中では病原性の菌類。

ショットガンシーケンス

- 現在の技術では、数十から数百の断片（**リード**）しか読むことができない。
- 長いゲノムは、いったん短いリードに裁断して読んでから、ジグソーパズルのように再構成（**アセンブリー**）する。

ショットガン法とは？

1. ゲノム全体を水圧などで物理的に断片化し、大腸菌で増やせる**プラスミドベクター**に組み込む。この処理を**クローニング**という。（次世代機はクローニングしない。）
2. 理論的には、ゲノム断片がクローン中にランダムに挿入される。
3. クローン配列を両側から読む（**メイトペア　リード**）。
4. 読んだ結果をつなぎ合わせて、元の配列を復元する。

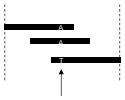
ショットガンシーケンス

要点

- ゲノム配列を完璧に復元することは難しい。
- リードをつなげた島のことを**コンティグ**と呼ぶ。
- クローンの両側から読んでおくと、復元が楽になる。

アセンブリーはどうやって？

- ランダムでも十分な数のリードを読めば、殆どの部分がカバーされるはずである。
- ただし、読み取りエラーや重複領域もあるので、リードを一意にはつなげられない。
- 数百万に及ぶリードをつなげてコンティグを作るソフトウェアを**アセンブラー**と呼ぶ。



SNPやミスマッチもありうる。
SNP: **s**ingle **n**ucleotide **p**olymorphism

シーケンスのカバー率

ゲノムサイズ *G* の生物をショットガン法でシーケンスする場合を考える。リードの数を *N* とする。読み取りのカバー率 (coverage) は、*a = NL/G*

区間 *L* 内に全くリードが存在しない確率は

$$\left(1-\frac{L}{G}\right)^N = \left(1-\frac{a}{N}\right)^N \xrightarrow{N\to\infty} e^{-a}$$

ゲノム配列を99.9%読む場合は、以下の式を解けばよい。

$$1-e^{-a} = 0.999$$

このとき、*a* = 6.9。 99.99%を読むには *a* = 9.2。

新型シーケンサーのIlluminaやSOLiDだと、1ラン（1週間）で数百 G bp の配列を出力する。理論的にはヒトゲノム (3 Gbp) を解読するのに十分な量である。

次世代シーケンス (Next Generation Seq)

第二世代	第三世代	第四世代？
emulsion PCR optical sequencing	single-molecule optical sequencing	semiconductor nano-pore
25-400 bp 20 Gb / day	1000 bp, skipping	100-200 bp 2 Gb / day?
DNA polymerase: Roche 454 Illumina HiSeq DNA ligase: AppliedBio Solid	DNA polymerase: Pacific Bioscience	Seminconductor: Ion Torrent Nanopore: Oxford Nanopore

いずれも、DNAのホスホジエステル結合を作成する際にでてくるリン酸や、水素イオンの量をマーカーとして塩基の判別をおこなう。

参考：遺伝子という概念

“gene” という言葉は「遺伝」を司る実体としてDNA発見以前から使われた。

EF Keller “The Century of the Gene”

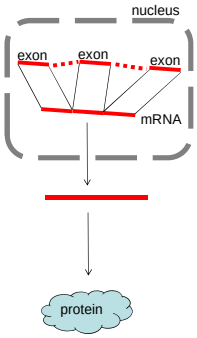
ヒトゲノム計画頃までの考え方

「遺伝子部分が特定機能のタンパク質をコードし、それらの相互作用が生命活動の設計図になる」

現在の考え方

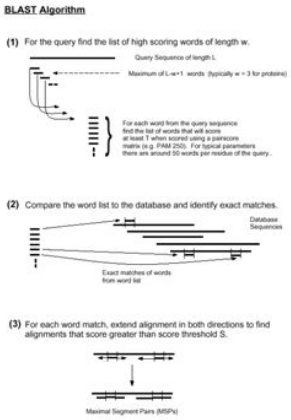
「タンパク質個々の働きは翻訳前後の修飾や、ノンコーディングRNAの影響も受けている」 ← 状況依存

ヒトの遺伝子数は2万2千と言われている。



BLAST

http://blast.ncbi.nlm.nih.gov/



動的計画法を基本としたアライメントは遅いので、配列断片（シード）を用いて高速化。

1.問い合わせ配列からシードを作成
2.シードがデータベース中にでてくる箇所を同定
3.一致したシードを左右に伸長
4.伸長した中で良いものを選択
high-scoring pair の評価

現在までに様々なオプションを開発。

DDBJ/EMBL/GenBank

http://www.ncbi.nlm.nih.gov/genbank/

読み取られた遺伝子配列は、ゲノムデータベースに登録される。
DDBJ ... DNA Data Bank of Japan
EMBL ... European Molecular Biology Laboratory (database)
GenBank ... NIH Genetic sequence database
3極構造で世界中の配列データを格納し、無償公開している。

レコードの例

```
LOCUS       NP_247669          253 aa        linear   BCT 28-JUL-2010
DEFINITION  nitrogenase reductase NifH2 [Methanocaldococcus jannaschii DSM 2661].
ACCESSION   NP_247669
DBSOURCE    REFSEQ: accession NC_000909.1
KEYWORDS    .
SOURCE      Methanocaldococcus jannaschii DSM 2661
ORGANISM    Methanocaldococcus jannaschii DSM 2661
            Archaea; Euryarchaeota; Methanococci; Methanococcales;
            Methanocaldococcaceae; Methanocaldococcus.
REFERENCE   1  (residues 1 to 253)
AUTHORS     Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., …
TITLE       Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii
JOURNAL     Science 273 (5278), 1058-1073 (1996)
PUBMED      8688087
REFERENCE …
```

演習: *Tyrannosaurus rex*

Tyrannosaurus rex の配列を blast 検索してみよう。

1. NCBIのウェブサイトを訪れる
2. 上部の検索スロットで、“Protein” “*Tyrannosaurus rex*” と入力
3. 適当な配列を選択 （他もOK）
4. GenBankデータを眺めた後、上部の FASTA をクリック
5. 配列データを copy

6. 画面最上部の resources -> Data&Software -> BLAST を選択
7. Protein blast というリンクを選択
8. 検索画面に先ほどの配列をペーストし、画面下部にあるサーチボタンで検索

T.rex protein

Psi-blast の結果、E-value 0.001-0.004 で collagen へのヒットが得られた。これは本当？

Sequences producing significant alignments with E-value BETTER than threshold

	Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
new	P0C2W2.2	RecName: Full=Collagen alpha-1(I) chain; AltName: Full=	54.7	54.7	5%	5e-05	100%	
new	Q9YIB4.1	RecName: Full=Collagen alpha-1(I) chain; AltName: Full=	55.1	55.1	5%	6e-05	94%	
new	P86209.1	RecName: Full=Collagen alpha-1(I) chain; AltName: Full=	50.8	50.8	5%	9e-05	94%	
new	P02457.3	RecName: Full=Collagen alpha-1(I) chain; AltName: Full=	52.4	52.4	5%	4e-04	94%	G M
new	XP_003222587.1	PREDICTED: collagen alpha-1(I) chain-like [Anolis carolin	50.8	50.8	5%	0.001	91%	U G M
new	XP_001367735.1	PREDICTED: collagen alpha-1(I) chain-like isoform 1 [Mor	50.8	50.8	5%	0.001	91%	U G M
new	XP_003768424.1	PREDICTED: collagen alpha-1(I) chain isoform 2 [Sarcopt	50.8	50.8	5%	0.001	91%	G
new	XP_003504008.1	PREDICTED: collagen alpha-1(I) chain-like [Cricetulus gri	50.8	50.8	5%	0.001	91%	G M
new	NP_445756.1	collagen alpha-1(I) chain precursor [Rattus norvegicus]	50.8	50.8	5%	0.001	91%	U G M
new	CAB01633.1	Collagen alpha1 [Rattus norvegicus]	50.8	50.8	5%	0.001	91%	G M
new	EHB16484.1	Collagen alpha-1(I) chain [Heterocephalus glaber]	50.8	50.8	5%	0.001	91%	
new	XP_003768423.1	PREDICTED: collagen alpha-1(I) chain isoform 1 [Sarcopt	50.8	50.8	5%	0.001	91%	G
new	XP_003414689.1	PREDICTED: LOW QUALITY PROTEIN: collagen alpha-1(I)	50.4	50.4	5%	0.002	91%	G M
new	P0C2W8.4	RecName: Full=Collagen alpha-1(I) chain; AltName: Full=	49.3	49.3	5%	0.004	91%	

Run PSI-Blast iteration 2 with max 500 Go

E-value と P-value

BLASTは、検索された類似配列の確からしさをE-valueで出力する。

E-value (expected value) とは、検索対象としたデータベースをランダム配列と仮定した際に何個ヒットするかを示す期待値である。この値は以下の式で計算される。

$$E = Kmne^{-\lambda S}$$

m, n: 比較する配列の長さ、S: 類似配列のスコア、 *K, λ*: 定数

配列の長さが2倍になれば同程度のスコアを出す確率は2倍になる。

2倍大きなスコアを出すには同程度の類似が続けて2回出る必要があるので、その確率は指数的に減少する。

P-value (probability value) とは、スコア*S* (期待値 *E*) のヒットが出る確率を示す。ランダム配列におけるヒットがポアソン分布に従うとすると以下の式で計算される。 $P = 1 - e^{-E}$

つまり、E-valueが大きいヒットほど、P-valueは1に近づく。

BLAST Family

- nucleotide BLAST (blastn) 核酸配列で核酸DBを検索
- protein BLAST (blastp) アミノ酸配列でアミノ酸DBを検索
- blastx 核酸配列を翻訳し、アミノ酸DBを検索

- Tblast シリーズ 核酸を翻訳したDBを検索
- psi-blast blastpの発展形。検索結果からプロファイルを作成し、そのプロファイルを元に検索をしなおして類似配列を集める。
- megablast blastnの発展形。ゲノムレベルの長い配列を検索できる。

アミノ酸置換行列

実際のデータから推定した、アミノ酸の「置換されやすさ」を示す指標。

例： BLOSUM62: 62%以上保存されたアミノ酸のブロック置換行列 (BLOck SUbstitution Matrix)。 置換されやすいアミノ酸対は正、されにくい対は負の値を持つ。

アミノ酸の分類

- A,M,I,L,V 非極性の側鎖
- C,S,T 極性の側鎖
- N,Q アミド基
- F,Y,W 芳香族
- H,R,K 塩基性
- D,E 酸性

KがEと置換しても良い（スコア0）などはかなり意外。
他の有名な置換行列にはPAM行列がある。

NCBIによるBLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	-1	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	-1	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	-2	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	-2	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-3	-2	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	1	-1	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	-1	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	1	-1	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

参考： 生物の名前について

生物の学名とは、リンネの二命名法に従うラテン名のこと。

ヒト *Homo sapiens*

Homo の部分を属名、*sapiens* の部分を種小名という。動物も植物も、種と呼ばれるものは学名を持つ。

Escherichia coli 大腸菌 *Mus musculus* マウス

Saccharomyces cerevisiae 酵母 *Allium cepa* たまねぎ

- 同じ種かどうかは、古くは形態学的手法、最近はリボソーム配列やミトコンドリア配列を元に判断される。
- 同種とされるものは、リボソーム配列が97%程度一致している。
- （正式には）学名には命名者を併記せねばならず、時代と共に変化する。

学名を整理するプロジェクト GBIF http://www.gbif.org/

まとめ

- 大量のゲノム情報が無償公開されている
- ゲノムの読み取りは、短いリードをアセンブルする
- 読み取った配列はデータベース検索によって機能をアサインする（アノテーション）
- 配列検索ツールの標準はBLAST
- E-value によって相同性の確からしさを評価する
- これらの配列を集めたものがデータベースとして公開されている

課題1

1. ヒトゲノム全体に対して、ある配列をBLAST検索する場合、E-value の値がどのくらいであれば、ヒットが進化的に関連深い有意なものと考えられるだろうか。理由とともに簡単に説明せよ。

2. 統合TV http://togotv.dbcls.jp/20071016.html を見てBLASTと psi-BLAST の違いを確認しなさい。イテレーションはなぜ収束するのか、自分の考えを簡潔に述べよ。

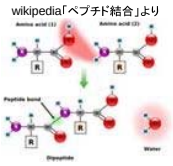
バイオインフォマティクス スタンダードコース
（配列と機能２）

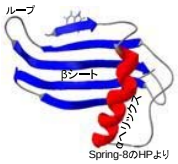
タンパク質の立体構造

遺伝子は、配列よりも立体構造のほうがよく保存される。

- アミノ酸の折りたたみ (folding) である**二次構造**には、**αヘリックス、βシート、ループ**がある。
- 二次構造が集まって、機能や構造の単位となる**ドメイン**を構成する。
- 複数ドメインからなる三次構造の決定には**X線結晶解析**や**NMR**（核磁気共鳴装置）**解析法**を用いる。

wikipedia「ペプチド結合」より



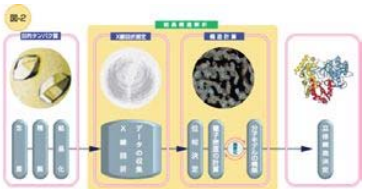


Spring-8のHPより

タンパク質の構造決定法

実験的な決定法
- X線の場合はタンパク質を結晶化し、回折のパターンから全原子の位置を決定する。NMRは溶液中の炭素や水素が共鳴するシグナルをもとに構造を予測する。

計算機による決定法
- 類似構造がわかれば局所的な最適化
- 類似構造を探すのに**スレッディング法**を用いる
- 構造がわからない場合は**ab initio法**で予測する



大型放射光施設Spring-8のページより

David BakerによるRosetta法は、ab initio法の代表格である。アミノ酸配列を与えられたら長さ9残基のフラグメント毎にPDBを検索して構造を予測する。
それらを組み合わせて全体構造を推定する。

Fold It!

ゲームとして立体構造問題を解くオンラインソフト

一般人を集めて、ゲームとして立体構造を解かせる。今は敗血症 (sepsis) を防ぐためのMannose-Binding Lectin問題が出題中。

立体構造のデータベース、分類

- PDB (Protein Data Bank)
  http://www.pdb.org/
  立体構造座標のデータベース。7万近い構造が登録されている。
- SCOP (Structural Classification Of Proteins)
  http://scop.mrc-lmb.cam.ac.uk/scop/
  4万タンパク質を手作業で分類。allα, allβ, α+β等の構造と、機能で分ける。
- CATH (Class, Architecture, Topology, Homologous superfamily)
  http://www.cathdb.info/
  タンパク質をドメインに切ってから半自動的に分類している。
- Uniprot ... 様々なタンパク質データの統合サイト
  http://www.uniprot.org/

演習：（スーパー）ファミリー

タンパク質立体構造第一号である（ミオ）グロビンの例

- グロビン
  - ヘモグロビン
  - ミオグロビン

1. SCOP にアクセス (Google “scop database”)
2. Top of the hierarchy に移動
3. All alpha proteins → Globin-like → Globin-like → Globins

ここまでくると、いろんなグロビンの構造がわかる

（ウマ、ヒト、マメ、クジラ、単細胞）

マルチプルアライメント

http://www.ebi.ac.uk/Tools/clustalw2/

複数の配列をアライメントする作業を多重（マルチプル）アライメントと呼ぶ。ペアワイズのアライメントを繰り返して全体を揃える方法が主流。（HMMなども使える）

高性能のツールはタンパク質の内側と外側を考慮するなど、複雑な評価関数を要する。代表的ツールは MAFFT。

参考：プロファイル、モチーフ

- プロファイルもプロフィールもプロフも profile
- タンパク質ファミリーに共通するアミノ酸配列パターンを意味する。同じ意味でモチーフという言葉もよく使われる。
- 配列のアライメントから、例えば以下のようなプロファイルを得たとしよう。（ [ ] はORの関係、x(2) は任意の2文字という意味）

  [FY]-C-x(2)-C-[AGT]-E-[MLI]-D

  [ ] の中はどれを選んでも一致と考える。

  F-C-x-x-C-A-E-M-D

  Y-C-x-x-C-G-E-L-D

進化的に相同でなくても、機能が同じならモチーフを共有しうる。

機能モチーフ、構造モチーフ

タンパク質の機能分類

- 細胞は、低分子化合物（アミノ酸、核酸、脂質など）、タンパク質、核酸配列（DNA,RNA）から構成される
- タンパク質には様々な機能がある
  - 構造タンパク質 … プロリンが多く含まれるケラチンは繊維として機能
  - 酵素タンパク質 … 食べ物からエネルギーを取り出したり、タンパク質を合成する「代謝」の主役
  - 膜タンパク質 … 細胞表面などに分布し、数々のシグナルを受け取る

  など
- 現在は、Gene Ontology により機能が整理されている

演習：グロビンの機能

グロビン遺伝子の機能を配列から調べてみよう。

1. Interproにアクセス http://www.ebi.ac.uk/interpro/
2. 検索画面で globin と入力

Globins are haem-containing proteins involved in binding and/or transporting oxygen. They belong to a very large and well studied family that is widely distributed in many organisms [1]. Globins have evolved from a common ancestor and can be divided into three groups: single-domain globins, and two types of chimeric globins, flavohaemoglobins and globin-coupled sensors. Bacteria have all three types of globins, while archaea lack flavohaemoglobins, and eukaryotes lack globin-coupled sensors [2].…

演習： Gene Ontology

GO は、3つのDirected Acyclic Graph から構成される

- **biological process** 代謝、細胞死、窒素利用などの「目的」
- **cellular component** 細胞膜、オルガネラ、シナプスなどの「部品」
- **molecular function** 酵素、調節因子、チャンネルなどの「機能」

1. http://www.geneontology.org/ で、”globin” を検索
2. 出てきた中で、 X associations というリンクをクリック

ヘモグロビンのしくみ

- ヘモグロビンは、赤血球の中で酸素を運搬するタンパク質。
- 4量体 （ホモテトラマー） を構成する。
- 肺の中で酸素を吸着し、末端組織中で酸素を離す。

↑

このプロセスを効率よくおこなうにはどうすればよいか？

酸素が多くある場合はなるべく多くと結合し、酸素が少ないときはできるだけ少ない数と結合すればよい。

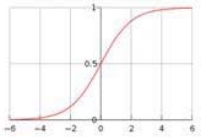
酸素と結合したヘモグロビンの割合 $$\theta = \frac{[L]^n}{K_d + [L]^n} = \frac{[L]^n}{(K_A)^n + [L]^n}$$

L : 酸素濃度 K_d: 定数

Hill 式で、n が [大きい 小さい] 場合にその役割をはたすことがわかる。

アロステリック効果

- シグモイド（S字）曲線に表されるように
  少ないときは反応せず
  多すぎるときはブレーキをかける
  働きを考えよう。
- これをアロステリック効果と呼ぶ。
  (allo 異なる + steric 形の)



- 数学的には Hill (ヒル) 式で記述できる。

反応した生体分子の割合（飽和率） $$\theta = \frac{[L]^n}{K_d + [L]^n} = \frac{[L]^n}{(K_A)^n + [L]^n}$$

(K_d: 乖離定数 [L]: 基質濃度)

$n > 1$ のとき： Lが増えるとθは急激に1に近づく

$n < 1$ のとき： Lが増えてもθはなかなか変化しない

参考：酵素のEC番号

- ほぼ全ての酵素反応は、国際生化学分子生物学連合によって階層的に分類され、**EC番号**が振られている。
- EC番号は反応式に基づいて4階層に分けられる。
- 最後の階層は単なる通し番号。

  - EC Class 1: Oxidoreductases 酸化還元酵素
  - EC Class 2: Transferases 転移酵素
  - EC Class 3: Hydrolases 加水分解酵素
  - EC Class 4: Lyases 除去付加酵素
  - EC Class 5: Isomerases 異性化酵素
  - EC Class 6: Ligases （ATPを用いる）合成酵素

まとめ

- 有用なタンパク質の立体構造はよく保存される
- 構造の分類からある程度機能も予測できる
- 立体構造は実験的に同定し、それを元に構造予測も行われる
- モチーフという概念でタンパク質の機能予測が可能
- ただし、配列から簡単には理解できない機能が多い
  - アロステリック効果

課題2

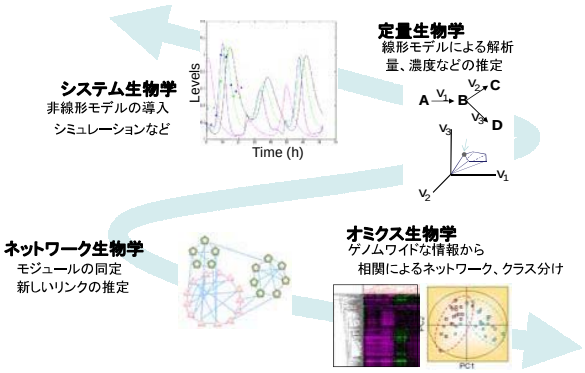
- PDBデータベースの「今月の分子」にある hemoglobin を読み、感想を述べよ。

http://www.pdbj.org/mom/index.php?l=ja&p=041

- 短縮型グロビン(truncated globin)の発見は業界に驚きを与えた。標準的なものは150残基なのに対し、短縮型は120残基しかない。SCOPの分類等で truncated globinを調査し、どのような生物が持っていてどんな機能なのか、簡潔に述べよ。

バイオインフォマティクス スタンダードコース （配列と機能３）

ネットワーク、システム解析へ



生物学におけるネットワーク

- **代謝のネットワーク (Metabolic Network)**
化合物がどのようにして合成されるかを記した経路
水道管のように、物質（水）の流れをあらわす

- **タンパク質相互作用のネットワーク (PPI Network)**
相互作用するタンパク質を結んでできる経路
伝言ゲームのように、情報の伝達をあらわす

- **進化や遺伝子のネットワーク (Genetic Network)**
配列の相同性や、共発現関係を結んでできる経路
実際の物理化学的ネットワークとは異なっている

ネットワークの定義

グラフ G = {頂点集合 V , 頂点間を結ぶ辺の集合 E }

vertex (node)	頂点
edge (link)	辺
degree	次数（頂点からの辺数）
path from x to y	xからyまでの頂点列
connected component	連結な部分グラフ
cycle	両端点が一致するpath
tree	cycleのない連結成分
spanning tree	全頂点を含む木

注. 多重辺やループは考慮しない。

6次の隔たり

友達の友達はみな友達

全世界に友人が等しくランダムに散らばっていたら
n^6 : n = 10 … 100万人
n = 20 … 6400 万人
n = 30 … 1億9200万人

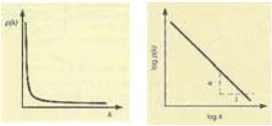
実際は
- 友人は地理的近傍に多い
- 友人の数は人により違う

Oracle of Bacon
http://oracleofbacon.org/
120万映画、150万俳優
1. お気に入り俳優名を入力
2. 何ステップか、確認

ネットワークの特徴

- スモールワールド
  - どんな頂点間も少ないステップ数でつながる
- スケールフリー
  - 共演数と人数をプロットすると、べき分布をなす
  - べき分布は $y = C x^{-\lambda}$ の形をとる（C, λ: 定数）

$\log y = \log (C x^{-\lambda})$
$= \log C - \lambda \log x$
つまり、log-logプロットが直線になる。



べき分布と指数分布

べき分布 $y = C x^{-\lambda}$
- dy/dx を計算すると$-\lambda y/x$
- 移項すると $dy/y = -\lambda\, dx/x$
  yの減少率は、xの減少率と傾きλで比例（xが大きくなるほど減少率は小さい）
- （正確には確率分布ではないが）ロングテール現象などのモデルに使える

指数分布 $y = C e^{-\lambda x}$
- dy/dx を計算すると$-\lambda y$
- 移項すると $dy/y = -\lambda\, dx$
  yの減少率はxの大きさに関係なく常に$-\lambda$
  (memoryless property)
- 時間に関係なく一定割合で壊れる部品などのモデルに使える

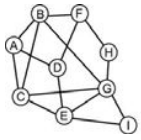
平均頂点間距離、クラスター係数

平均頂点間距離 AL

- ネットワークが連結（全部つながっている）と仮定し、全頂点間の最短距離の平均値
  - ネットワークは無向
  - ランダムなネットワークは比較的低い値 O(log n)

クラスター係数 C

- 点に隣接する頂点集合の間に引かれる辺の密度
  - n 頂点の間に引ける辺の数の最大値は n(n-1)/2
  - 全頂点間に辺があればクラスター係数は 1
  - ツリーのクラスター係数は 0



Albert, R, Barabasi, AL, *Reviews of Modern Physics* **74**, 47 (2002)

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460 902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

BAモデル

Preferential attachment "Rich gets richer."

- m_0個の頂点からスタート
- 頂点を順次追加する際、既存の頂点に次数に比例する確率で辺を作る。

特徴． ネットワークの進化をモデルしており、かつ、次数の分布がべき則 $p(k) \sim k^{-3}$ になる。

A. L. Barabasi, R. Albert "Emergence of Scaling in Random Networks", *Science* 286, 1999

ネットワークのキーワード

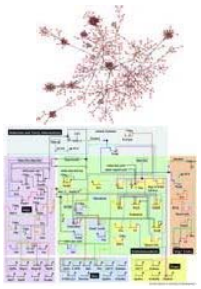
- **スケールフリーネットワーク**

次数分布がべき則に従う場合

少数の頂点が大変多くのリンクを持つ場合、それらを「**ハブ**」と呼ぶ。ハブの中にも更にハブがあるような階層構造をスケールフリー性とよぶ

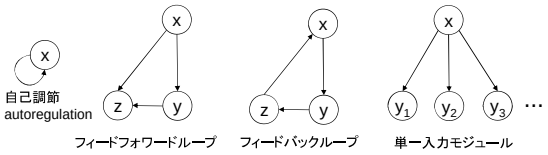
- **ネットワークモチーフ**

ネットワークの中に頻出する部分ネットワークのこと。たとえば代謝ネットワークにはフィードバック制御が多く、シグナル伝達にはフィードフォワード制御が多い



ネットワークのモジュール

- 遺伝子やタンパク質など何らかの「集合」が与えられ、集合に含まれる要素間に、結合するとか反応するなどの何らかの「関係」が定義されれば、ネットワークになる。
- 辺に方向性がある場合、→（矢印）で表現する。
- ネットワークはモジュールに分けられる。



ネットワークモチーフ

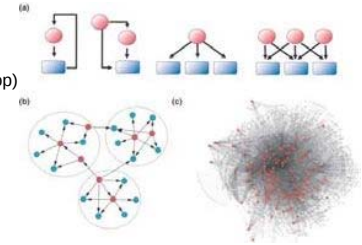
大腸菌や酵母の転写制御ネットワークはランダムではない

•3点のフィードフォワードが多い（ランダムの5倍）

•制御因子が複数遺伝子をコントロールしている

こうした特徴的なパターンをモチーフと呼ぶ。

中でもFFL (feed forward loop) は面白い特徴をもつ。



自己調節、フィードバック

- 大腸菌の転写ネットワーク要素のおよそ10%に存在

ほとんどは負のフィードバック （ブレーキ）

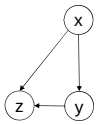
利点

- タンパク質の発現量を、生成速度と分解速度のみで制御するよりは、特定の値（最大値） に収束するスピードが速い
- 定常状態をロバストに保つことができる

注. 正のフィードバック （アクセル） も存在し、それらはどんどん増加するプロセスに多い。

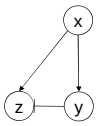
フィードフォワード

- フィードバック回路に比べ、存在意義がわかりにくい。
- X, Yが、AND関数でZに効いているとするなら、入力が持続的にONになったときのみ、ZがONになる。
  → 入力ノイズに対するロバストネス
- X, Yが、OR関数でZに効いているとするなら、入力が終了しても、Zは緩やかにONを続けられる。
  → 入力に対する遅延応答



YがZを抑制するように働いていたらどうなるか？

- XとYがANDでZに効いているとするなら、XがONになってからしばらくの間だけZもONになれる。
  → 入力情報のパルス化



まとめ

•ネットワーク研究には、自然界のネットワークの大域的な特徴を解析するものと、ネットワークの細部を解析するものとがある。
  –スケールフリーネットワーク
  –ネットワークモチーフ

•単に複雑なシステムとして捉えるのではなく、モジュールに分解する視点が重要
  –数理モデルより先に、思考実験をおこなう
  –思考実験の「穴」を数理モデルで補完

おわり

3.9　タンパク質の機能と構造

3.9.1 概要

講師：広川 貴次／産業技術総合研究所　生命情報工学研究センター
創薬分子設計チーム　チーム長

日時：2012 年 9 月 13 日　10:20～／12:50～／14:40 ～

場所：琉球大学総合情報処理センター

内容：立体構造やドメインのデータベースについて学び、PDB でみる立体構造を実習する。また INTERPRO によるモチーフ予測、膜貫通領域予測、触媒部位の予測等を実習する他、AUTDOCK によりリガンドのドッキングシミュレーション実習も行う。

3.9.2 準備

AutoDock の導入と動作確認 ※参照：AutoDock 導入手順および確認

・autodock インストール

・MGLTools インストール

・python インストール

docking フォルダの配布　▼docking 内のファイル

名前	種類	サイズ
1MK5_A_lig.pdb	PDB ファイル	3 KB
1MK5_A_rep.pdb	PDB ファイル	141 KB
autodock4.2.3_win32.exe	アプリケーション	963 KB
bitools.html	Chrome HTML Document	3 KB
MGLTools-1.5.4-Setup.exe	アプリケーション	35,189 KB
python-2.5.2.msi	Windows インストーラー パッケージ	11,028 KB

3.9.3 配布物

・講義スライド（※必要に応じて各自印刷）

3.9.4 課題

・kadai.pdf　（※各自印刷、回答して提出）

3.9.5 キーワード

生命科学分野
生物学基礎 細胞（生体膜、細胞内小器官、真核生物、原核生物）、細胞周期（体細胞分裂、減数分裂）、遺伝学（メンデルの法則、連鎖）、発生・分化、シグナル伝達（受容体、プロテインキナーゼ、細胞接着）、ウィルス、免疫、脳・神経
分子生物学 セントラルドグマ、複製（DNA）、転写（RNA、遺伝子発現、逆転写、スプライシング）、翻訳（タンパク質、tRNA、リボゾーム、遺伝暗号、翻訳後修飾）、RNA 遺伝子、生体エネルギーと代謝、エピジェネティクス（クロマチン構造）
バイオテクノロジー 遺伝子組み換え、分子クローニング、PCR、塩基配列決定法（サンガー法、次世代シークエンサー）、ゲノム（ショットガン法）、トランスクリプトーム（マイクロアレイ）、プロテオーム（質量分析、二次元電気泳動、酵母ツーハイブリッド法）、立体構造決定（X 線結晶構造解析、NMR、電子顕微鏡）、ChIP-seq 法
情報科学分野
コンピュータシステム 二進数、論理演算（論理積、論理和、排他的論理和）、数値演算（浮動小数点数、丸め誤差、桁落ち）、プログラミング言語（C 言語、Perl 言語）、ネットワーク（OSI 参照モデル、IP アドレス）、マークアップ言語（HTML、XML）
アルゴリズム データ構造（スタック、キュー、リスト、木構造）、探索（二分探索、ハッシュ表、木探索）、ソーティング、文字列比較
データベース技術 ３層スキーマ構造、データモデル（階層型、リレーショナル型）
確率・統計 確率分布・確率変数（母集団、標本、独立性、ベイズの定理）、データ解析（平均、分散、相関、回帰）、推定・検定（点推定、区間推定、帰無仮説、対立仮説、有意水準）
機械学習 決定木、k-最近接近傍法、ニューラルネットワーク、サポートベクトルマシン、感度、特異度、クロスバリデーション法、クラスタリング、K-平均法、自己組織化マップ（SOM）

バイオインフォマティクス分野
分子生物学データベース 文献 DB（PubMed）、ゲノム DB、核酸配列 DB、アミノ酸 DB、モチーフ DB（モチーフライブラリー）、立体構造 DB、代謝パスウェイ DB、多型ＤＢ、発現ＤＢ、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析 アラインメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアラインメント、グローバルアラインメント、Smith-Waterman 法、ペアワイズアラインメント、マルチプルアラインメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA 二次構造予測、遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC 含量、コドン使用頻度）、ゲノム比較（ゲノムアラインメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬）
タンパク質立体構造解析 化学構造（結合角、結合距離、二面角）、タンパク質高次構造（二次構造、三次構造、四次構造）、立体構造表現（コンタクトマップ、ラマチャンドランマップ、分子グラフィックス）、構造比較（重ね合わせ、RMSD、構造アラインメント、構造モチーフ、構造分類）、タンパク質構造予測（二次構造予測、ホモロジーモデリング、フォールド認識、スレッディング、３D-1D 法、分子動力学計算）、構造 DB
遺伝学・進化学 分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J 法）、最節約法、最尤法、同義置換、非同義置換）、ニューウィック形式、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLA タイプ）、Hardy-Weinberg の法則
トランスクリプトーム解析・プロテオーム解析 遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析、DNA マイクロアレイ、質量分析、二次元電気泳動、酵母ツーハイブリッド法
パスウェイ解析・システム生物学 ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

3.9.6 時間割

時間	内容
10:20～	<u>タンパク質の立体構造</u> ・ DNA からタンパク質へ ・ タンパク質の階層情報 ・ タンパク質立体構造データ ・ タンパク質階層構造 1 次構造、2 次構造、3 次構造 ・ タンパク質のグラフィックス表現 ・ 立体構造の重要性 ・ 活性部位を形成する立体構造の保存性 配列、ドメイン ・ 生体高分子の立体構造決定方法 ・ Protein Data Bank
10:42～	【演習】 タミフルとタンパク質の結合情報確認
10:55～	<u>タンパク質の立体構造予測</u> ・ 配列情報と構造情報のギャップ ・ タンパク質モデリングの考え方 相同性、類似性、ホモログ、アナログ ・ タンパク質モデリングの戦略 鋳型構造の探索、構造決定までのプロトコル ・ ロータマーライブラリ ・ ループモデリング ・ 分子力学計算による極小化 ・ 自動構築プログラム→モデル評価
11:13～	【演習】 構造不明のタンパク質配列について Verify3D で解析 ＊解析中に講義を進め、昼食後に確認 <u>鋳型構造の利用が困難な場合の解析</u> ・ Fragment Assembly 法

11:30～	【演習】 SWISS-MODEL にて、任意の配列を解析 <u>タンパク質の機能、構造解析</u> ・ タンパク質機能・構造のオントロジー 生物学的機能、タンパク質局在、分子機能解析
11:45～	昼食
12:50～	<u>タンパク質の立体構造予測</u> ・ タンパク質機能予測の必要性 バイオ繊細設計例 ・ 標的タンパク質同定 ・ タンパク質機能同定方法 Direct assay、Expression pattern、Genomic interaction など ・ 機能予測のための方法 生物学的機能解析、タンパク質局在解析、分子機能解析 ・ シグナルペプチドとタンパク質局在 ・ モチーフ配列 正規表現、PROSITE、確率プロファイル、隠れマルコフモデル
13:16～	【演習】 InterPro にて、任意の配列を持つタンパク質の局在を解析
13:23～	<u>局在解析例</u> ・ モチーフに依存しない解析例 ・ 膜タンパク質トポロジー予測 ・ 予測が難しい場合のアプローチ 分類学習、決定木、ニューラルネットワークなど
13:40～	【演習】 ・ PSORT II により、任意配列の局在解析 ・ TMHMM による膜貫通トポロジー予測 ・ SignalP によるシグナルペプチド予測は省略 ・ CASTp による活性ポケット探索
13:45～	・ KEGG Ligand リサーチ
14:00～	・ DrugBank による化合物探索 以降は、時間まで課題に取り組む

	休憩
14:44～	タンパク質-リガンド結合予測 ・ 創薬研究の推移 ・ 化合物の多様性 ・ 分子標的創薬のコンセプト ・ 分子標的創薬のコンセプト ・ ドッキング問題 ・ リガンドドッキングについて ・ ドラッがぶるポケット予測 ・ タンパク質-リガンドドッキング ・ スコアリングン形成
15:05～	【演習】 AutoDock サイト（ver4）をもちいて、任意のリガンドとレセプターの結合を確認

AutoDock 導入手順および確認

インストールに必要なファイル（3 つ）

・autodock4.2.3_win32.exe

・MGLTools-1.5.4-Setup.exe

・python-2.5.2.msi

１．実習用として C ドライブ直下に C:¥docking を作成する。

２．autodock4.2.3_win32.exe を実行し、インストールする。デフォルトのインストール先として、C:¥Windows¥system32¥が指定されているが、C:¥docking¥に変更しインストールする。

※フォルダを選択する際、C:¥docking¥system32¥と表示されるが、system32¥を削除し、C:¥docking¥とする。インストールが終了すると C:¥docking¥の中に次の 3 つが展開される。

・autodock4.exe

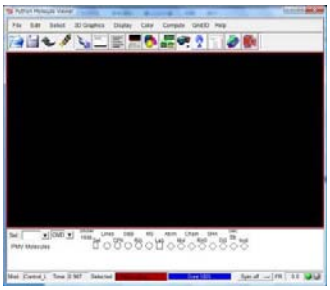
・autogrid4.exe

・cygwin1.dll

３．python-2.5.2.msi を実行しインストールを行う。インストール先は、C:¥docking¥を選択する。

４．MGLTools-1.5.4-Setup.exe を実行しインストールを行う。インストール先は、C:¥docking¥を指定する。インストール途中で表示される次のメッセージでは、Non commercial を選択し、Continue をクリックする。

インストール終了後、次のウィンドウが自動的に開く。

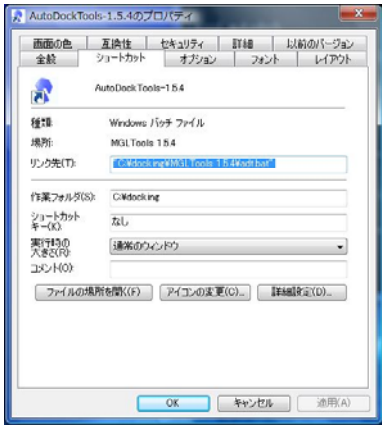


またデスクトップに上記 3 つのアイコンが生成される。

※沖縄高専メディアコンテンツ演習室の PC においては、C:¥ユーザー¥mi○○○（ログイン ID）¥デスクトップ¥の中に生成される。

５．AutoDockTools の作業フォルダの変更。

生成したショートカットアイコン“AutoDockTools1.5.4 を右クリック、プロパティを開き、作業フォルダを C:¥docking に変更する。



※グリッド計算およびドッキング計算

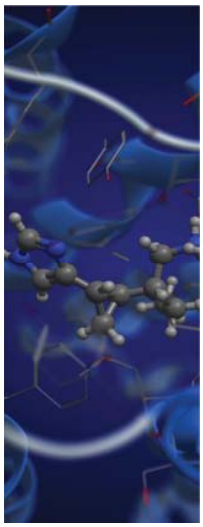
Windows のコマンドプロンプトでコマンドを実行して計算する。

コマンドプロンプトは、作業フォルダである docking フォルダを指定し、SHIFT キーを押しながら右クリック、「コマンドウインドウをここで開く」で開く。

C:¥docking>

が表示される。

以上



平成24年度沖縄県産業振興基金事業
「バイオインフォマティクス人材育成推進事業」
公益財団法人沖縄科学技術振興センター

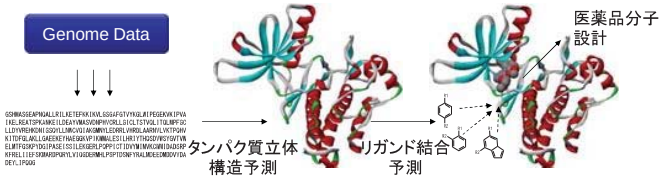
琉球大学　夏期集中講義

タンパク質の機能と構造

広川貴次

産業技術総合研究所
生命情報工学研究センター　創薬分子設計チーム
東京大学大学院　新領域創成科学研究科
北海道大学大学院　生命科学院

講義のねらい

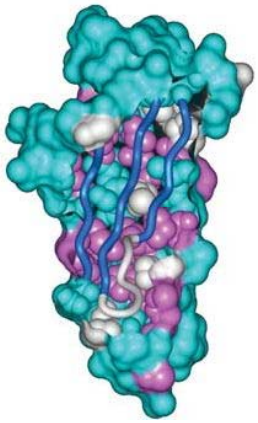


- 実用レベルのタンパク質配列・構造解析技術の習得
  - タンパク質機能予測
  - タンパク質立体構造予測
  - タンパク質-リガンド結合予測

スケジュール：2012年9月13日

時間	内容
10:20~11:50	第1章　タンパク質の立体構造（イントロダクション） 第2章　タンパク質立体構造予測
12:50~14:20	タンパク質立体構造予測（続き） 第3章　タンパク質機能予測
14:40~16:10	第4章　タンパク質-リガンド結合予測

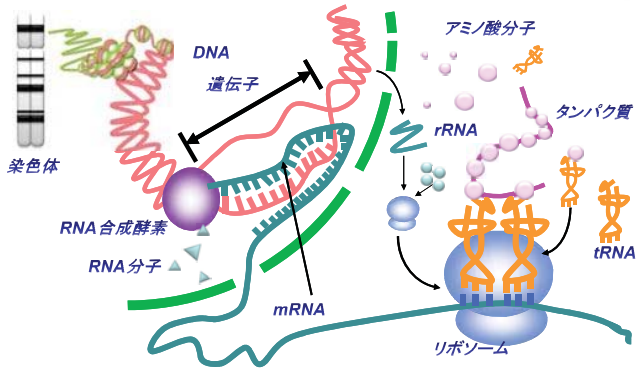
講義内容には、実習も含まれます。またWeb実習における予期せぬ障害（先方サーバーのダウン、メンテナンス）については、予めご了承ください



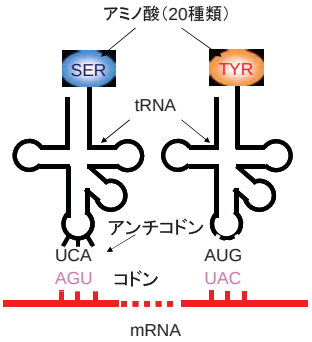
第1章

タンパク質の立体構造
Protein 3D-Structure

DNA配列からタンパク質へ



遺伝子≒タンパク質,tRNA,rRNA



2番目の塩基

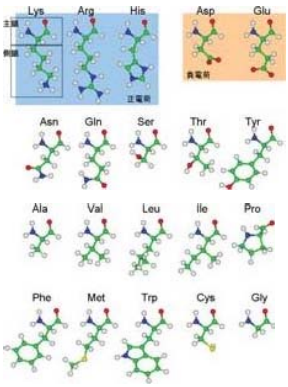
1番目の塩基	U	C	A	G	3番目の塩基
U	PHE PHE LEU LEU	SER SER SERS ER	TYR TYR STOP STOP	CYS CYS STOP TRP	U C A G
C	LEU LEU LEU LEU	PRO PRO PRO PRO	HIS HIS GLN GLN	ARG ARG ARG ARG	U C A G
A	ILE ILE ILE MET	THR THR THR THR	ASN ASN LYS LYS	SER SER ARG ARG	U C A G
G	VAL VAL VAL VAL	ALA ALA ALA ALA	ASP ASP GLU GLU	GLY GLY GLY GLY	U C A G

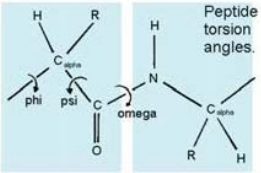
コドンとアミノ酸の対応表

アミノ酸残基／配列

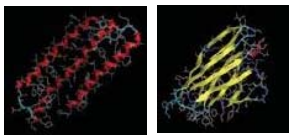
親水性アミノ酸残基

疎水性アミノ酸残基





ω角は、ペプチド結合形成箇所で基本的には、トランス構造を形成するのが殆どである



代表的な二次構造（左：ヘリックス、右：シート）

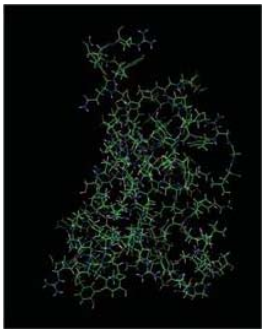
タンパク質の階層情報

プリオン遺伝子配列（DNA配列）

```
atgctggttctctttgtggccacatggagtgacctgggcctctgcaagaagcg
cccgaagcctggaggatggaacactgggggcagccgataccccggggcagggca
gccctggaggcaaccgctaccccacctcagggcggtggtggctgggggcagcct
catggtggtggctgggggcagcctcatggtggtggctgggggcagcccccatgg
tggtggctgggggacagcctcatggtggtggctgggggtcaaggaggtggcaccc
acagtcagtggaacaagccgagtaagccaaaaaccaacatgaagcacatggct
ggtgcagcagctggggcagtggtgggggggcctggcggctacatgctgggaag
tgccatgagcaggcccatcatacatttcggcagtgactatgaggaccgttact
atcgtgaaaacatgcaccgttaccccaaccaagtgtactacaggccccatggat
gagtacagcaaccagaacaactttgtgcacgactgcgtcaatatcacaatcaa
gcagcacacggtcaccacaaccaccaagggggagaacttcaccgagaccgacg
ttaagatgatggagcgcgtggttgagcagatgtgtatcacccagtacgagagg
gaatctcaggcctattaccagagaggatcgagcatggtcctcttctcctctcc
acctgtgatcctcctgatctcttttcctcatcttccctgatagtgggatga
```

アミノ酸配列

```
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGS
RYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGG
WGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTN
MKHMAGAAAAAGAVVGGLGGYMLGSAMSRPIIHFGSD
YEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCIT
QYERESQAYYQRGSSMVLFSSPPVILLISFLIFLIV
G（253残基）
```



アミノ酸配列が作る高次構造

タンパク質立体構造データ

KHMAGAAAAGAVVGGLGGYML-GSAMSRPIIHFGSDYEDRYYRENM

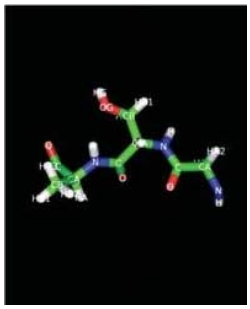
X Y Z

```
ATOM     91   N  GLY A 131       2.821   4.466   6.486
ATOM     92  CA  GLY A 131       2.745   5.753   7.168
ATO      [illegible] LY A 131       1.606   5.767   8.191
ATO      [illegible] LY A 131       0.893   4.770   8.360
ATOM     93   H  GLY A 131       2.455   4.395   5.548
ATOM     94 1HA  GLY A 131       2.594   6.541   6.435
ATOM     95 2HA  GI [illegible]     [illegible]   5.945   7.678
ATOM     96   N  SI [illegible]     [illegible]   6.899   8.878
ATOM     97  CA  SER A 132      0.493   7.011   9.998
ATOM     98   C  SER A 132     -0.965   6.730   9.606
ATOM     99   O  SER A 132     -1.342   6.647   8.431
ATOM    100  CB  SER A 132      0.644   8.417  10.596
ATOM    101  OG  SER A 132     -0.016   8.579  11.832
ATOM    102   H  SER A 132      2.048   7.679   8.719
ATOM    103  HA  SER A 132      [illegible]   6.286  10.751
ATOM    104 1HB  SER A 1[illegible]       605  10.778
ATOM    105 2HB  SER A 1[illegible]     [illegible]   9.157   9.891
ATOM    106  HG  SER A 132     -0.050   9.548  12.001
ATOM    107   N  ALA A 133     -1.822   6.688  10.622
ATOM    108  CA  ALA A 133     -3.251   6.451  10.517
```

原子名

アミノ酸名

残基番号



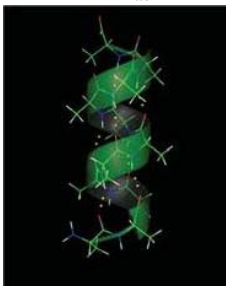
プリオンタンパク質の部分構造（ステックモデル）

タンパク質の階層構造

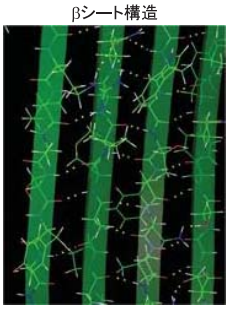
水素結合による二次構造形成

1結合＝2～10 kcal/mol
（共有結合＝100 kcal/mol）

αヘリックス構造

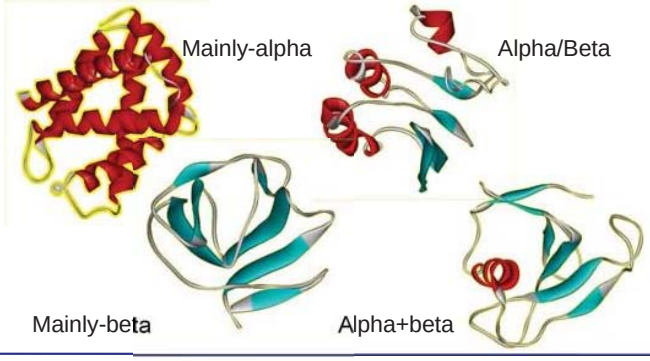


βシート構造

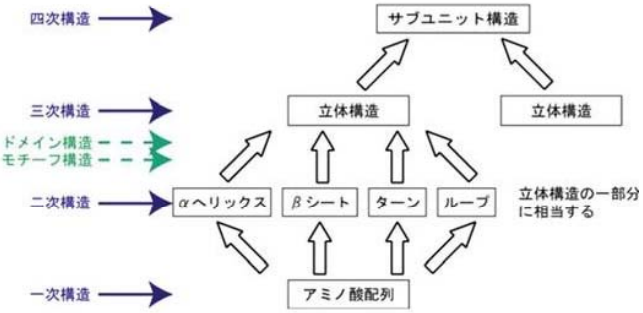


タンパク質の構造のドメイン構造（構造単位）

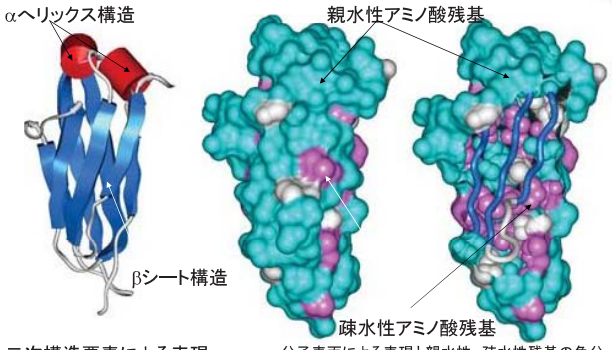
αwithα βwithβ αwithβ



タンパク質の階層構造



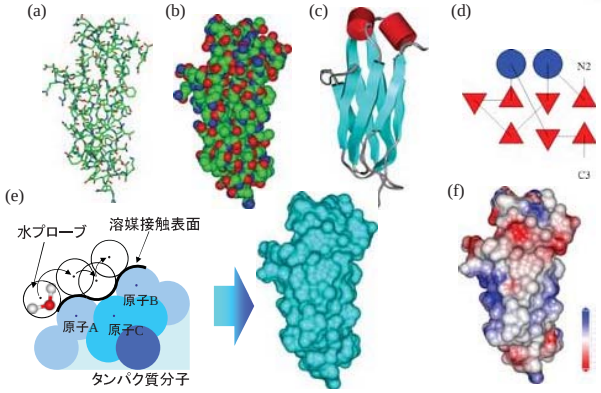
疎水性相互作用による三次構造の形成・安定化



二次構造要素による表現

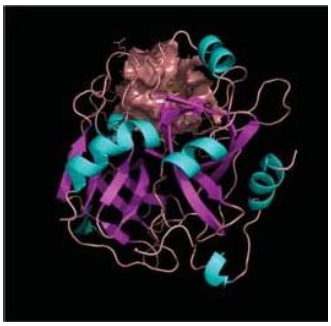
分子表面による表現と親水性、疎水性残基の色分け：右側は、内部構造表示

タンパク質の基本的なグラフィックス表現

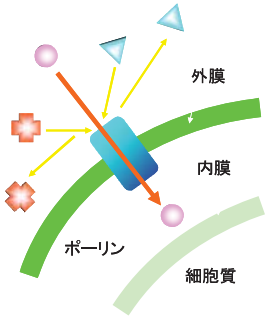


タンパク質の立体構造の重要性

- タンパク質の立体構造は、その機能の密接に関係している。
- 進化において、タンパク質のアミノ酸配列より立体構造の保存性が高い
- 既存の薬物のほとんどは、タンパク質と相互作用している。その相互作用形式や、薬物の化学構造は、タンパク質の立体構造からも推定できる可能性がある（鍵と鍵穴）

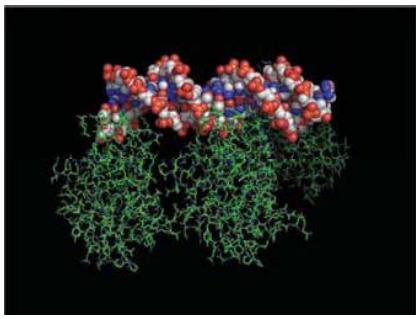


分子のふるい（外膜タンパク質ポーリン）



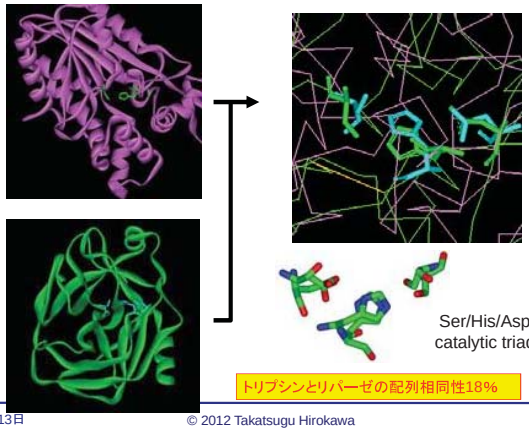
DNA結合タンパク質（癌抑制機能）

癌抑制機能を持つP53タンパク質：DNAとの結合モデル



ポワソンボルツマン方程式によるP53タンパク質表面の静電ポテンシャル表示：赤が負電荷、青が正電荷を示す

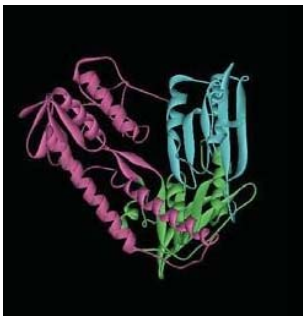
活性部位を形成する立体構造の保存性



トリプシンとリパーゼの配列相同性18%

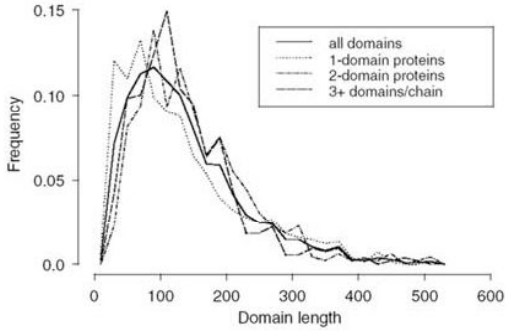
タンパク質ドメインの特徴 （参考）

- 全体構造への安定性に寄与が少ない独立した部分で密にコンパクトになっている
- 機能部分を含むユニット疎水性コア部分を1つ以上持つ
- ドメイン間では二次構造接触が少ない



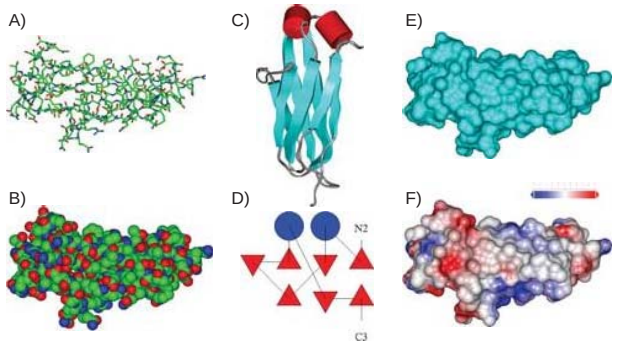
3つのドメインを持つタンパク質例：3grs

ドメイン長の特徴 （参考）



Wheelan et al., “Domain size distributions can predict domain boundaries”, Bioinformatics, 2000.より

タンパク質構造の可視化



ラインモデル（A）、CPKモデル（B）、カートゥンモデル（C）、TOPSモデル（D）、Connollyモデル（E）、表面静電ポテンシャルマッピングConnollyモデル（F）

生体高分子の立体構造決定方法

実験的手法

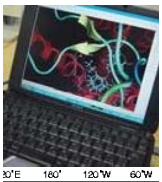
http://www.spring8.or.jp/index.html

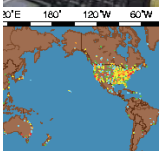
http://www.varianinc.com/

http://www.jeol.co.jp/

問題点
大量発現、精製
結晶化
界面活性剤分子

計算機的手法



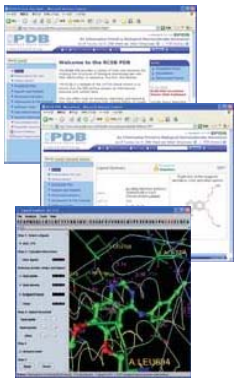


http://www.stanford.edu/group/pandegroup/folding/

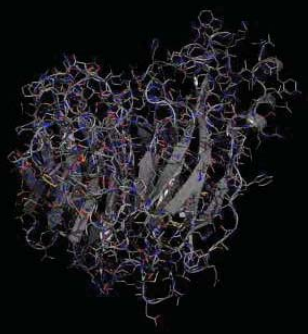
パソコン
並列計算機
グリッド環境
古典力学計算
バイオインフォマティクス etc

Protein Data Bank

- タンパク質、核酸、ウィルスなど生体高分子を中心とした実験的に決定された三次元構造データベース
  - 約67656件（2010年8月31日）
  - 低分子リガンドとのタンパク質複合体に関するデータも含む
- 一次構造、二次構造情報、原子座標、結晶化条件、構造因子、三次元イメージ
- 様々な情報へのリンクやプラグインの導入
  - 近年、特に低分子リガンド情報検索、可視化が充実（右図の下は、Ligand Explorerプラグインによる結合相互作用の可視化画面）



Neuraminidase/Tamiflu



Protein Data Bank実習

実習①

第2章

タンパク質立体構造予測
Protein Structure Prediction
Protein Modeling
Homology Modeling

配列情報と構造情報のギャップ

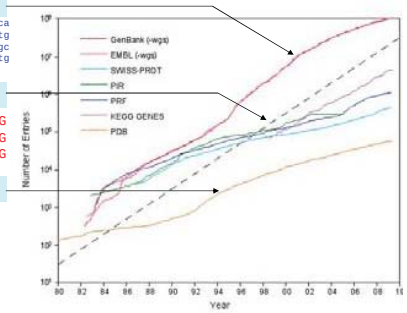
DNA配列

```
gtggctgggtcaaggaggtggcacccacagtcagtggaaca
agccgagtaagccaaaaaccaacatgaagcacatggctggtg
cagcagctgggcagtggtgggggggccttggcggctacatgc
tgggaagtgccatgagcaggcccatcatacatttcggcagtg
```

タンパク質配列

```
MANLGCWMLVLFVATWSDLGLCKKRPKPG
GWNTGGSRYPGQGSPGGNRYPPQGGGGWG
QPHGGGWGQPHGGGWGQPHGGGWGQPHGG
```

タンパク質構造



タンパク質モデリングの考え方

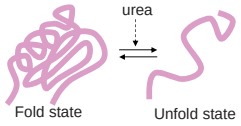
分子進化

ANCESTOR
(aldolase)

Homologous sequence / Similar structure

熱統計力学・物理化学

Anfinsen



Sequence to structure

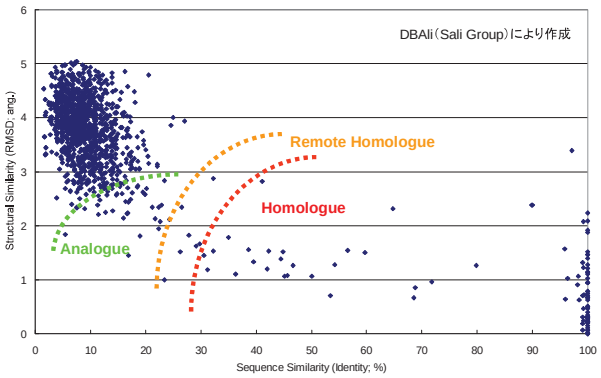
Levinthal

Time scale: ~msec

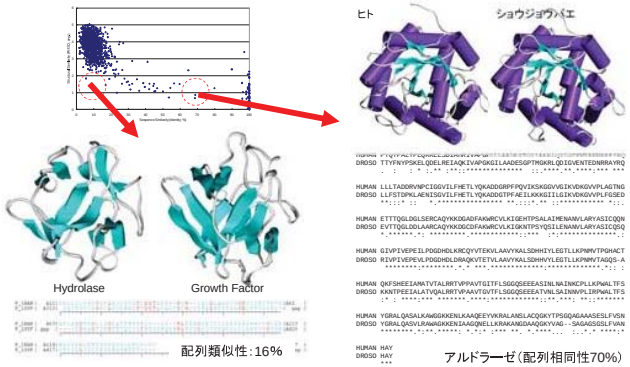
Not random diffusion

Folding pathway

配列の相同性と構造の類似性



ホモログとアナログ



タンパク質モデリングの戦略

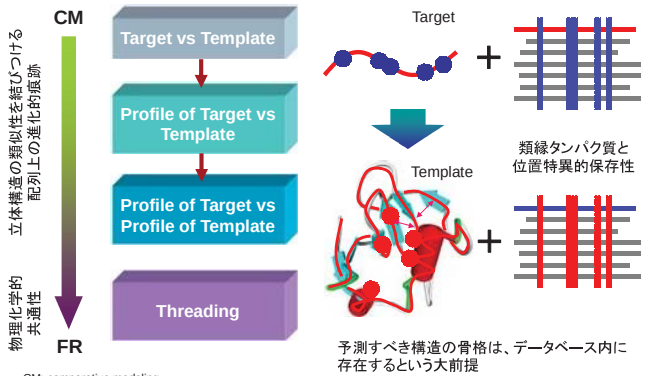
Search and Optimization

- **ホモロジーモデリングおよび構造認識法**
  - 鋳型構造への依存度大
  - 構造は、データベース中に存在
  - 少ない探索空間
  - 大きなタンパク質も対応可能
- **Fragment Assembly法**
  - 鋳型構造への依存度小
  - 新規な構造に対応
  - 膨大な探索空間
  - 小さなタンパク質（200残基以下）のみ対応

Potential Energy Surface

- **物理化学的ポテンシャル**
  - 動的変化の観測
  - 膨大な計算時間と精度
  - 対象範囲：数十残基
- **経験的ポテンシャル**
  - 出現頻度や、ペアポテンシャルなど
  - ヒューリスティック
  - 比較的良い構造が選出
  - 最適化に適したポテンシャル

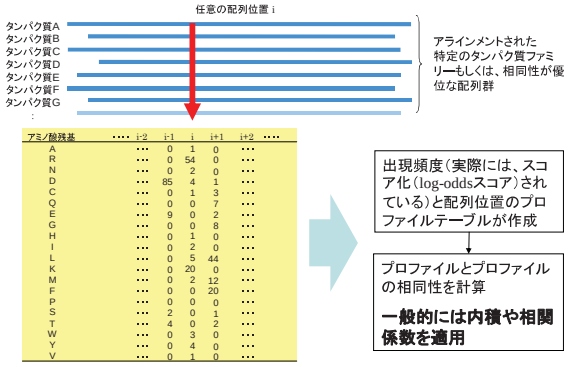
鋳型構造の探索



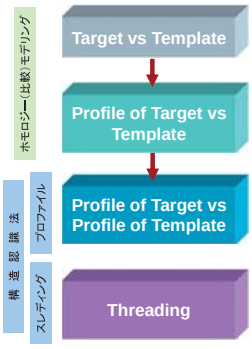
予測すべき構造の骨格は、データベース内に存在するという大前提

CM: comparative modeling
FR: fold recognition

プロファイルとは



代表的なプログラム（方法論）



BLAST、*FASTA*

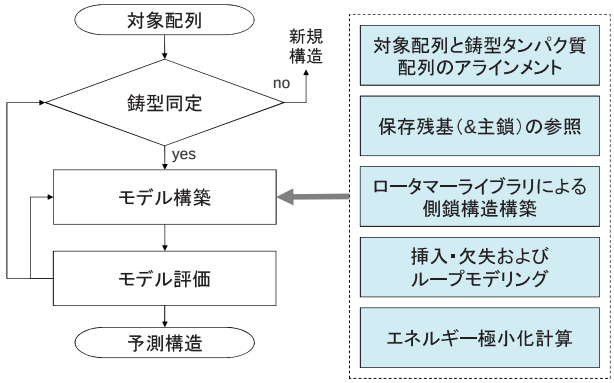
PSI-BLAST（ e-value< 10^{-10}）

Profile-Profile（&2D、3D情報を付加）方法

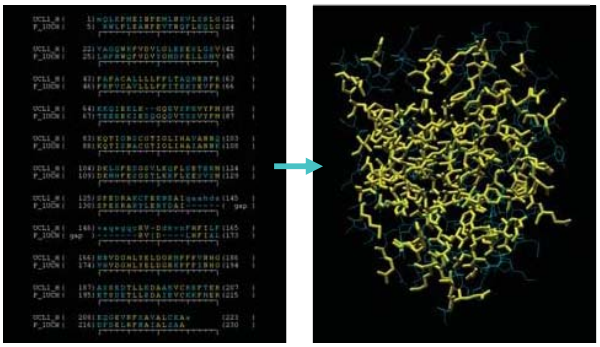
FFAS03、*FORTE*、*FUGUE*、*3D-PSSM*、*INBGU*、*Sam-T99*、*HMAP*

Sippl-Potential、*LIBRA*、*GenTHREADER*

構造構築までのプロトコール

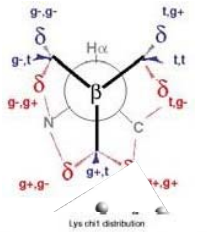


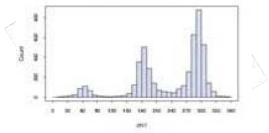
保存残基（&主鎖）の参照



ロータマーライブラリ

- タンパク質中に観測される二面角は、3種の二面角配座（g-, t, g+）がとる値に近いが、近傍原子との立体障害や側鎖間の相互作用の影響を受け、とりうる配座に偏りが生じる（Rotamer library）
- 主なRotamer library分類
  - Backbone-independent
  - Secondary structure-dependent
    - Alpha-helix, beta-sheet , coil
  - Backbone-dependent
    - φ、ψマップ





ループモデリング

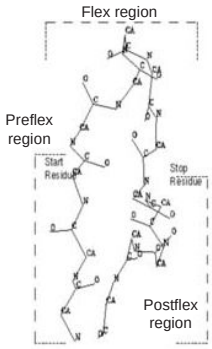
- ループ領域のpreflex, postflex領域について距離マトリックスを作成

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}, i, j \in pe, po$$

- PDBから抽出されたループ構造候補とのpreflex, postflexr両領域の距離マトリックス比較し、最小値をとる候補構造を割り当てる。

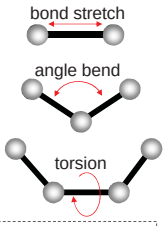
$$D_c^2 = \frac{2}{N(N+1)} \sum_{i \in pe, po} \sum_{j \in pe, po, j>1} \left(d_{ij} - d^c_{r_c(i) r_c(j)}\right)^2, i, j \in pe, po$$

- 分子力学/動力学計算により極小化計算を行う。



分子力学計算による極小化

- 原子から構成される分子を、調和力によって相互に作用しあう質量の集合体として捕らえる計算方法
  - ポテンシャルエネルギーを定義
  - 小分子から巨大分子まで適応
  - 相互作用項（ポテンシャル関数）や力場パラメータに違いによって様々な理論モデルが存在



$$V_{PEF} = \sum_{bonds} \frac{k_i}{2}(l_i - l_{0,i})^2 + \sum_{angles} \frac{k_i}{2}(\Theta_i - \Theta_{0,i})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

ポテンシャル関数例

自動構築プログラム

- 対象配列と鋳型タンパク質配列のアラインメント
- 保存残基（&主鎖）の参照
- ロータマーライブラリによる側鎖構造構築
- 挿入・欠失およびループモデリング
- エネルギー極小化計算

実績のあるツール
（benchmark analysis by SBC）

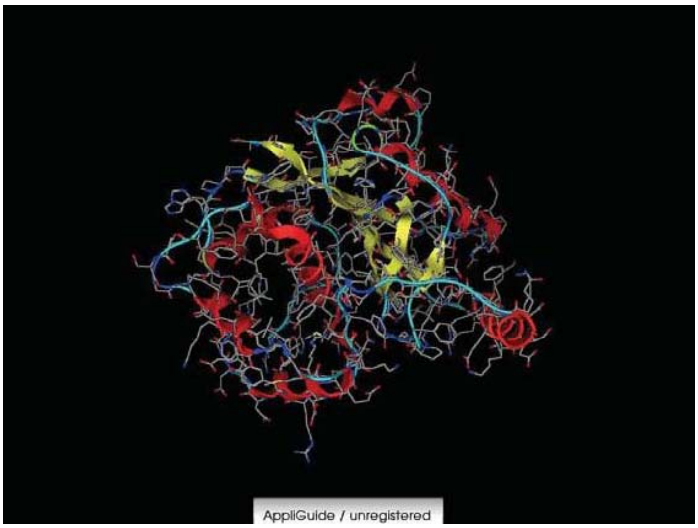
MODELLER（Sali, 1993）

SegMod（Levitt, 1992）

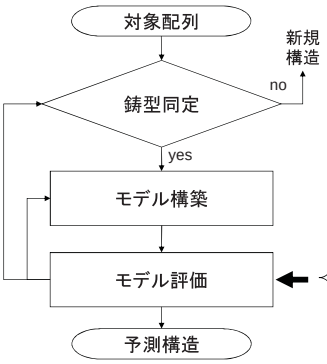
Nest（Honig, 2003）

SCRWL（Dunbrack, 1997）

Other methods:
Swiss-MODEL, 3D-Jigsaw, Builder



モデル評価



モデリングで最も重要なプロセス

Environment Based Method

Verify3D

Pair Potential Based Method

Prosa II or Prosa2003

Verify3D

3次元構造における残基の環境クラス

側鎖がタンパク質に埋もれている領域（溶媒への露出度）と残基の周辺の極性度によりアミノ酸残基を6クラスに分類（二次構造を込みで18種類）

E：溶媒に接している環境

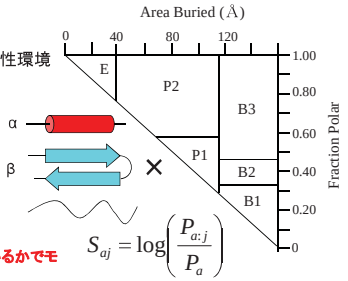
P1：部分的に埋もれている；わずかな極性環境

P2：部分的に埋もれている；極性環境

B1：埋もれている；疎水環境

B2：埋もれている；わずかな極性環境

B3：埋もれている；極性環境



$$S_{aj} = \log\left(\frac{P_{a:j}}{P_a}\right)$$

精度の高い結晶構造で統計データを収集。

予測構造がその統計値にどれくらい準じているかでモデル構造を評価

鋳型構造の利用が困難な場合

Search and Optimization

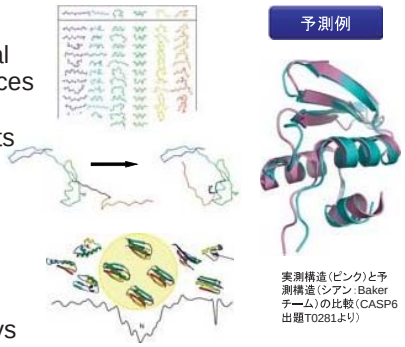
- **ホモロジーモデリングおよび構造認識法**
  - 鋳型構造への依存度大
  - 構造は、データベース中に存在
  - 少ない探索空間
  - 大きなタンパク質も対応可能
- **Fragment Assembly法**
  - 鋳型構造への依存度小
  - 新規な構造に対応
  - 膨大な探索空間
  - 小さなタンパク質（200残基以下）のみ対応

Potential Energy Surface

- **物理化学的ポテンシャル**
  - 動的変化の観測
  - 膨大な計算時間と精度
  - 対象範囲：数十残基
- **経験的ポテンシャル**
  - 出現頻度や、ペアポテンシャルなど
  - ヒューリスティック
  - 比較的良い構造が選出
  - 最適化に適したポテンシャル

Fragment Assembly法

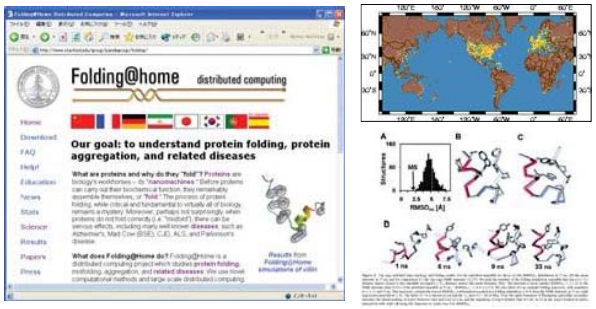
1. Select fragments consistent with local sequence preferences
2. Assemble fragments into models with native-like global properties
3. Identify the best model from the population of decoys



実測構造（ピンク）と予測構造（シアン：Baker チーム）の比較（CASP6 出題T0281より）

Dr. Strauss（Santa Fe,2002）より

物理化学的手法による予測（参考）



http://www.stanford.edu/group/pandegroup/folding/

ホモロジーモデリング（入門編）

実習②

- 入力情報：アミノ酸配列

```
MAALDSLSLFTSLGLSEQKARETLKNSALSAQLREAATQAQQTLGSTIDKATGILLYGLA
SRLRDTRRLSFLVSYIASKKIHTEPQLSAALEYVRSHPLDPIDTVDFEREGGVGVIVTPE
QIEEAVEAAINRHRRPQLLVERYHFNMGLLMGEARAVLKWADGKMIKNEVDMQVLHLLGPK
LEADLEKKFKVAKARLEETDRRTAKDVVENGETADQTLSLMEQLRGEALKFHKPGENYKT
PGYVVTPHTMNLLKQHLEITGGQVRTRFPPEPNGILHIGHAKAINFNFGYAKANNGICFL
RFDDTNPEKEEAKFFTAICDMVAWLGYTPYKVTYASDYFDQLYAWAVELIRRGLAYVCHQ
RGEELKGHNTLPSPWRDRPMEESLLLFEAMRKGKFSEGEATLRMKLVMEDGKMDPVAYRV
KYTPHHRTGDKWCIYPTYDYTHCLCDSIEHITHSLCTKEFQARRSSYPWLCNALDVYCPV
QWEYGRLNLHYAVVSKRKILQLVATGAVRDWDDPRLFTLTALRRRGFPPEAINNFCARVG
VTVAQTTMEPHLLEACVRDVLNDTAPRAMAVLESLRVIITNFPAAKSLDIQVPNFPADET
KGFHQVPFAPIVFIERTDFKEEPEPGFKRLAWGQPVGLRHTGYVIELQHVVKGPSGCVES
LEVTCRRADAGEKPKAFIHWVSQPLMCEVRLYERLFQHKNPEDPTEVPGGFLSDLNLASL
HVVDAALVDCSVALAKPFDKFQFERLGYFSVDPDSHQGKLVFNRTVTLKEDPGKV
```

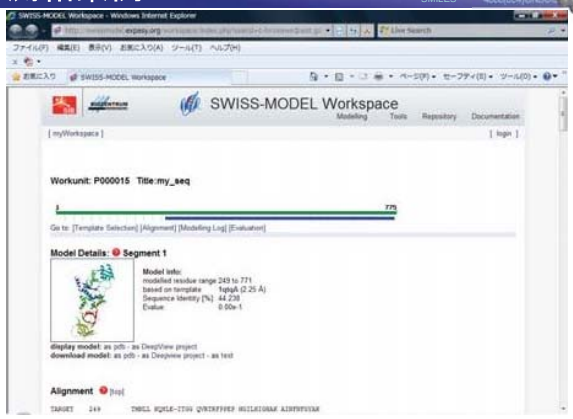
- 解析プログラム
  - SWISS-MODEL
    - 鋳型タンパク質の検索から主鎖、側鎖構築
    - 簡単な極小化計算
    - 結果は、リアルタイムで表示 or サーバーよりメールが届く

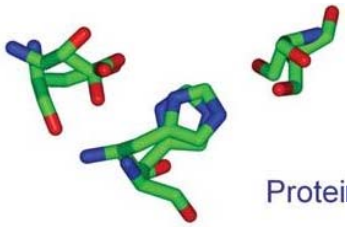
SWISS-MODEL入力手順

実習②

予測結果例

実習②

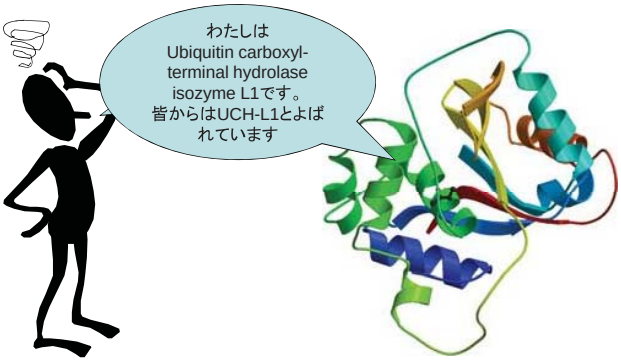




第3章

タンパク質機能予測
Protein Function Prediction
Annotation
Sequence/Structure Motif

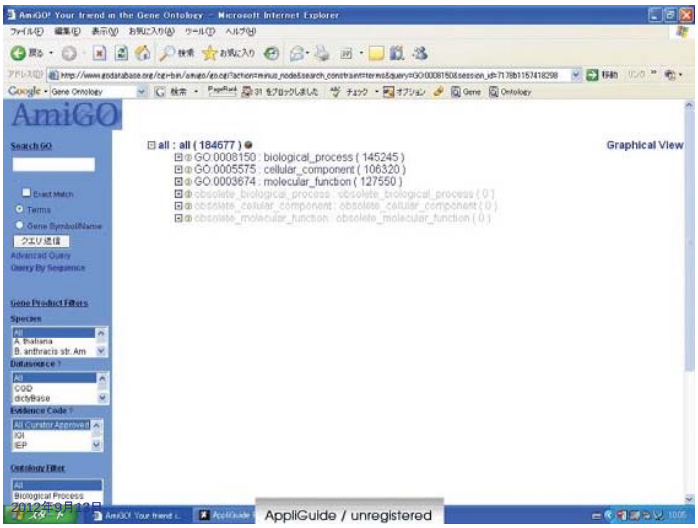
タンパク質機能情報って・・・



タンパク質機能・構造のオントロジー

- オントロジーとは？
  - ものごとを体系的に分類し、その関係を記述するもの
- タンパク質機能
  - なんのため？ (Biological process)
  - どこで？(Cellular component)
  - 何をしている？(Molecular function)
- タンパク質構造
  - 二次構造含有
  - フォールド
  - スーパーファミリー、ファミリー

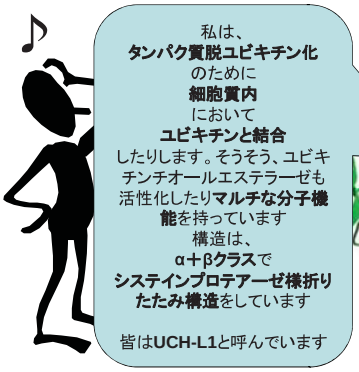
世界基準！



Gene Ontology：機能の体系的分類

- 生物学的機能（Biological Process）
  - *Text mining, Pathway, etc.*
- タンパク質局在（Cellar Component）
  - *Localization site prediction*
    - *Amino acid composition, Motif, hydrophobicity*
- 分子機能解析（Molecular Function）
  - *Sequence similarity, Fold similarity, Structure Prediction*
  - *Function prediction using structure imformation*

タンパク質の機能情報って（ふたたび）



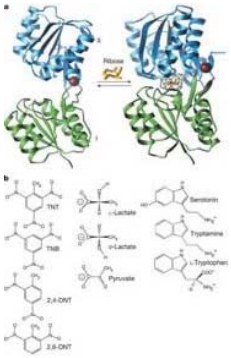
タンパク質機能予測の必要性

- 疾患志向
  - 文献、マイクロアレイ解析結果
  - 遺伝子とその機能は殆どが既知
- オーファンレセプター
  - レセプターであることは明らかであるが、結合するリガンドがわからないもの
- リガンド志向
  - リガンドが既知でそのレセプターを探索する場合（副作用タンパク質の同定等）
- 特定の遺伝子の改変
  - タンパク質工学的利用（バイオセンサー等）

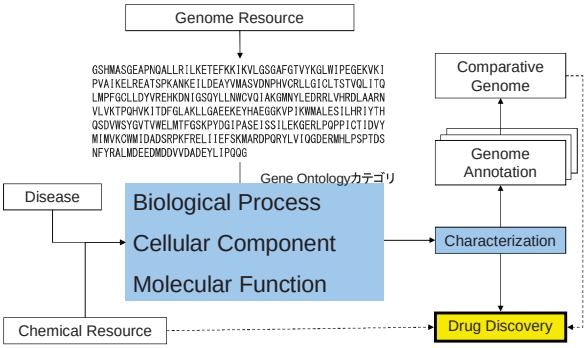
バイオセンサー設計例

Computational design of receptor and sensor proteins with novel functions, *Nature* **423**, 185 - 190 (2003)

- コンピュータによるリガンド結合部位の網羅的な変異モデルの作成により、新規リガンドを認識するセンサー受容体を作成。実験（シグナル伝達経路内に変異モデルを組み込み、遺伝子調節をモニター）による検証。
- Wild-typeに結合しているリガンドを基にポケット形成残基の位置を同定（12-18残基）し、配列ミュータント（1015-1023）とその構造（1045-1068）を生成、ドッキング計算。リガンドは、TNT、L-Lactate、Serotoninの三種を利用。それぞれ極性、疎水性、自由度などが異なるようなものを選択。
- DEE理論により側鎖モデリングとドッキングを実施。1タンパク質vs 3つのリガンド計算は、20プロセッサで3日間。



標的タンパク質同定

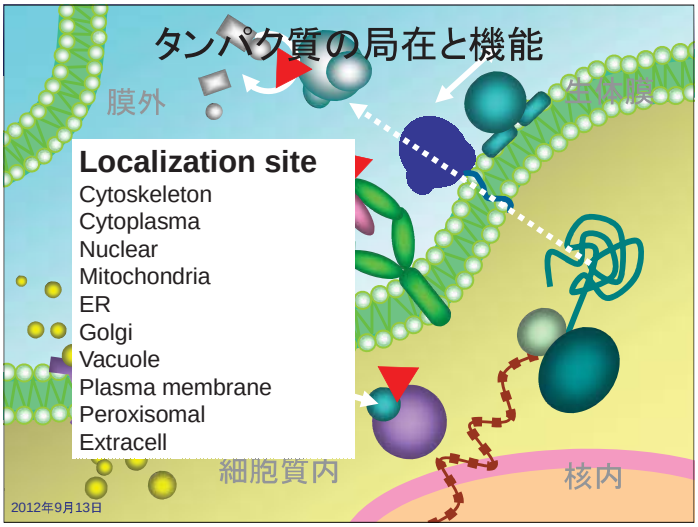


タンパク質の機能同定方法

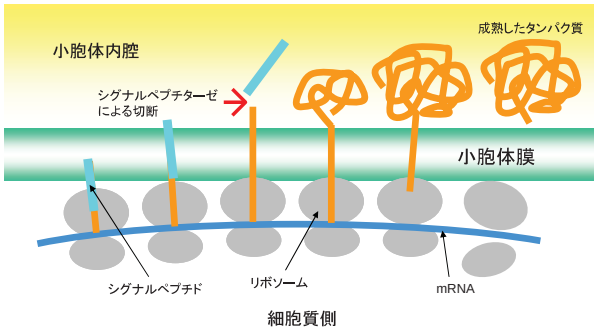
- *Direct assay*
  - Enzyme assays, *in vitro* reconstitution, immunofluorescence, etc.
- *Expression pattern*
  - Microarray data, northern, western blots, etc.
- *Genetic interaction*
  - Suppressors, synthetic lethals, functional complementation, etc.
- *Mutant phenotype*
  - Gene mutation/knockout, anti-sense experiments, RNAi experiments, etc.
- *Physical interaction*
  - 2-hybrid interaction, co-purification, co-immunoprecipitation
- ***Computational analysis***

機能予測のための方法

- タンパク質局在（Cellular Component）
  - Localization site prediction
    - Amino acid composition, Motif, Hydrophobicity
- 分子機能解析（Molecular Function）
  - Sequence similarity, Fold similarity, Structure Prediction
  - Function prediction using structure information
- 生物学的機能（Biological Process）
  - Text mining, Pathway, etc.



シグナルペプチドと分泌型タンパク質



シグナル配列の特徴

Protein Engineering **10**, pp.1–6, 1997
Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites
Henrik Nielsen, Jacob Engelbrecht, Søren Brunak and Gunnar von Heijne

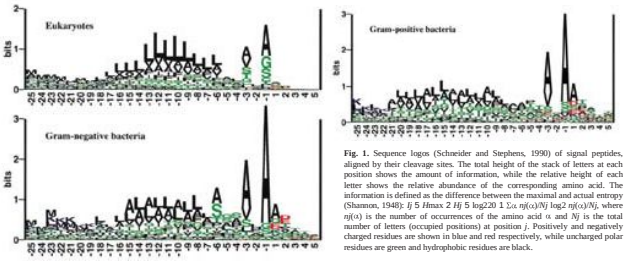
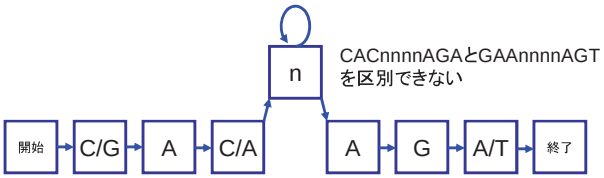


Fig. 1. Sequence logos (Schneider and Stephens, 1990) of signal peptides, aligned by their cleavage sites. The total height of the stack of letters at each position shows the amount of information, while the relative height of each letter shows the relative abundance of the corresponding amino acid. The information is defined as the difference between the maximal and actual entropy (Shannon, 1948): $I_j = H_{max} - H_j = \log_{220} 1 - \sum_\alpha n_j(\alpha)/N_j \log_2 n_j(\alpha)/N_j$, where $n_j(\alpha)$ is the number of occurrences of the amino acid α and N_j is the total number of letters (occupied positions) at position j. Positively and negatively charged residues are shown in blue and red respectively, while uncharged polar residues are green and hydrophobic residues are black.

モチーフ配列（正規表現）

位置番号	1	2	3	4	5	6	7	8	9	10	11
遺伝子A	C	A	C	a	a	a	c	g	A	G	A
遺伝子B	C	A	C	a	t	g	g	-	A	G	T
遺伝子C	C	A	A	t	c	t	a	-	A	G	A
遺伝子D	G	A	C	c	g	c	t	-	A	G	A
遺伝子E	C	A	C	a	c	t	-	-	A	G	A



モチーフ記述例（PROSITE）

タンパク質ファミリーやドメインのデータベースで、各々のファミリーの間で特徴的に保存されているモチーフを正規表現で記述

X	いずれかのアミノ酸
[AGL]	Ala or Gly or Leu
{CH}	CysとHis以外
-	区切り
x(3)	x-x-x
x(1,3)	x or x-x or x-x-x
<	N端　（C端は>）
.	パターン記述終了

PROSITEモチーフ例

[RK]-x(2,3)-[DE]-x(2,3)-Y.

マッチするモチーフ配列例

KLRACEDEEYRE

モチーフ配列（確率プロファイル）

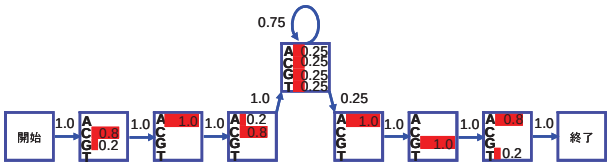
位置番号	1	2	3	4	5	6	7	8	9	10	11
遺伝子A	C	A	C	a	a	a	c	g	A	G	A
遺伝子B	C	A	C	a	t	g	g	-	A	G	T
遺伝子C	C	A	A	t	c	t	a	-	A	G	A
遺伝子D	G	A	C	c	g	c	t	-	A	G	A
遺伝子E	C	A	C	a	c	t	-	-	A	G	A

	1	2	3	4	5	6	7	8	9	10	11
A	0	1.0	0.2						1.0	0	0.8
C	0.8	0	0.8			0.25			0	0	0
G	0.2	0	0						0	1.0	0
T	0	0	0						0	0	0.2
-	0	0	0	0	0	0.2	0.75	0	0	0	0
--	0	0	0	0	0	0	0	1.0	0	0	0

繰り返し構造（例えばnの任意の繰り返し）を表現できない

モチーフ配列（隠れマルコフモデル：HMM）

位置番号	1	2	3	4	5	6	7	8	9	10	11
遺伝子A	C	A	C	a	a	a	c	g	A	G	A
遺伝子B	C	A	C	a	t	g	g	-	A	G	T
遺伝子C	C	A	A	t	c	t	a	-	A	G	A
遺伝子D	G	A	C	c	g	c	t	-	A	G	A
遺伝子E	C	A	C	a	c	t	-	-	A	G	A



InterPro（モチーフ検索サイト）

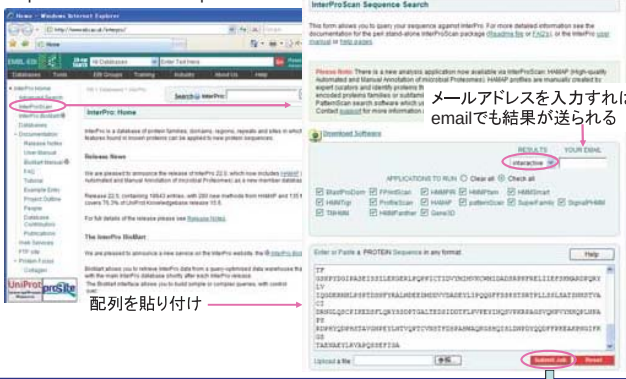
モチーフデータベース、ドメインデータベース、構造分類データベースなどを参照して対象配列への機能アノテーションを行う

InterPro（モチーフ解析）

実習③

http://www.ebi.ac.uk/interpro/

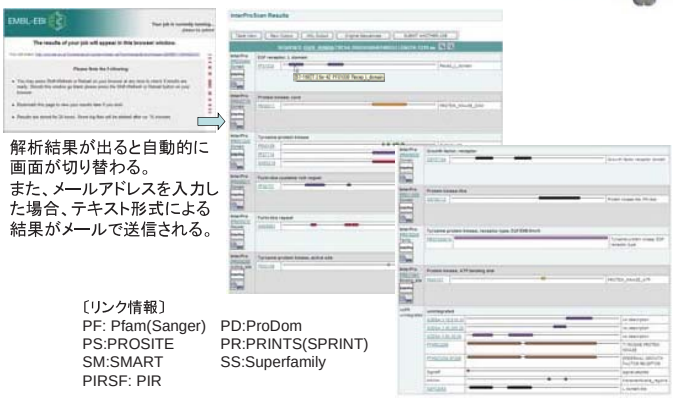
例題配列：EGFR_HUMAN.fasta



メールアドレスを入力すれば、emailでも結果が送られる

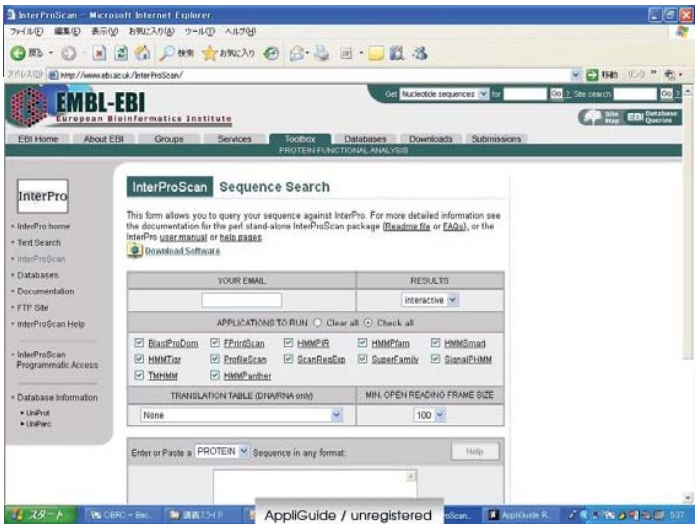
InterPro 解析結果

実習③

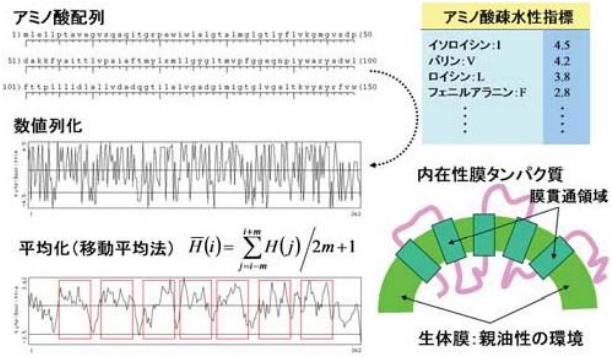


解析結果が出ると自動的に画面が切り替わる。
また、メールアドレスを入力した場合、テキスト形式による結果がメールで送信される。

〔リンク情報〕
PF: Pfam(Sanger)　PD:ProDom
PS:PROSITE　PR:PRINTS(SPRINT)
SM:SMART　SS:Superfamily
PIRSF: PIR



モチーフに依存しない解析例



膜タンパク質トポロジー予測

膜タンパク質トポロジー予測＝膜貫通領域（個数、領域）＋ループの細胞質内外予測

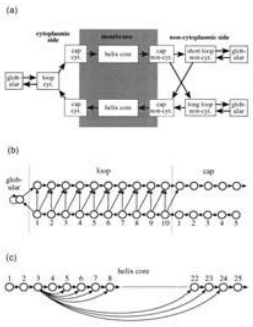
※膜タンパク質判別＝膜貫通領域が2本以上とされることが多い

J. Mol. Biol. (2001) **305**, 567–580

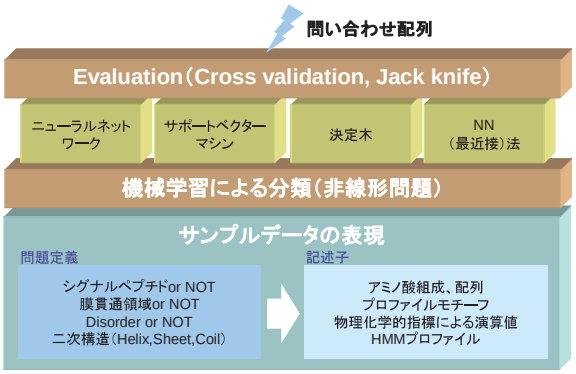
Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes

Anders Krogh[1]*, Björn Larsson[1], Gunnar von Heijne[2] and Erik L. L. Sonnhammer[3]

Figure 1. The layout of the hidden Markov model. (a) The overall layout. Each box corresponds to one or more states in the HMM. Parts of the model with the same text are tied, i.e. their parameters are the same. Cyt. represents the cytoplasmic side of the membrane and non-cyt. the other side. (b) The detailed structure of the inside and outside loop models and helix cap models. (c) The structure of the model for the helix core modelling lengths between five and 25, which translates to helices between 15 and 35 when the caps are included.



予測が難しい場合のアプローチ



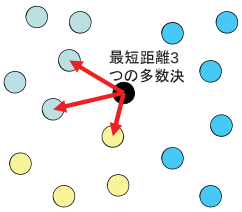
分類学習

代表サンプルを定める方法
- K近傍法、最短距離法

分割線に区切る方法
- 判別分析、ベイズの決定境界、マハラノビス汎距離

決定木による方法
ニューラルネットワーク
サポートベクターマシン

$$P^{LS}(i) = [a_1, a_2, a_3, \ldots a_N]$$



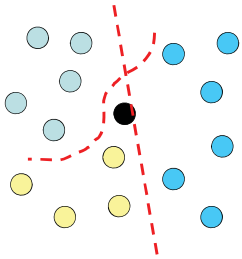
分類学習（続き）

代表サンプルを定める方法
- K近傍法、最短距離法

分割線に区切る方法
- 判別分析、ベイズの決定境界、マハラノビス汎距離

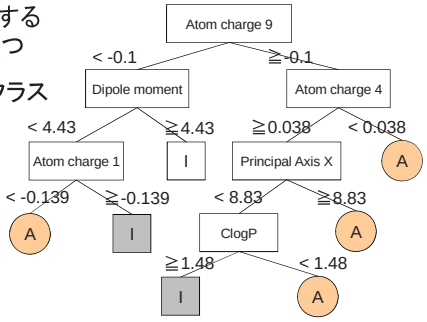
決定木による方法
ニューラルネットワーク
サポートベクターマシン

$$P^{LS}(i) = [a_1, a_2, a_3, \ldots a_N]$$



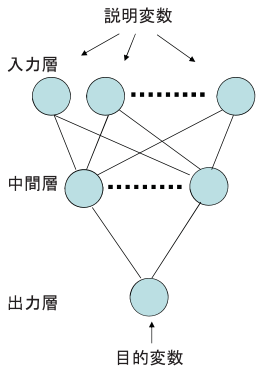
決定木

- クラスタリングの手順を視覚的に表現する機械学習法の一つ
- 新しいデータのクラス分類に利用



ニューラルネットワーク（NN）

- Back Propagation型
  - 階層型ネットワーク構造
  - 教師ありのNN
  - 実際の出力と希望の出力の誤差が最小になるように結合荷重を変更しながら学習（逆誤差伝播法とも呼ばれている）
  - 入力層、中間層、出力層からなる。中間層にシグモイド関数により非線形変換をしている。
  - ネットワークは、入力x_1, x_2, …, x_nに対して中間層を経て出力層から値を出力する



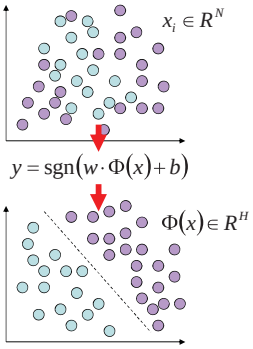
サポートベクターマシン（SVM）

- カテゴリ認識の方法
- 超高次空間への写像を行い、線形識別可能な問題にする

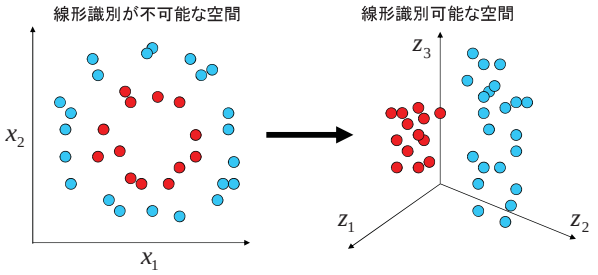
$$x_i \in R^N$$

$$y_i = 1 \text{ or } -1$$

カテゴリ（またはクラス）に対応

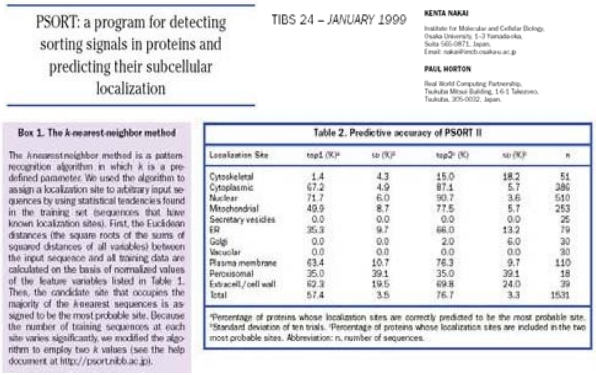


高次元への写像例



$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

kNN識別による局在予測



PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization

TIBS 24 – JANUARY 1999

KENTA NAKAI

PAUL HORTON

Box 1. The k-nearest-neighbor method

The k-nearest-neighbor method is a pattern-recognition algorithm in which k is a predefined parameter. We used the algorithm to assign a localization site to arbitrary input sequences by using statistical tendencies found in the training set (sequences that have known localization sites). First, the Euclidean distances (the square roots of the sums of squared distances of all variables) between the input sequence and all training data are calculated on the basis of normalized values of the feature variables listed in Table 1. Then, the candidate site that occupies the majority of the k-nearest sequences is assigned to be the most probable site. Because the number of training sequences at each site varies significantly, we modified the algorithm to employ two k values (see the help document at http://psort.nibb.ac.jp).

Table 2. Predictive accuracy of PSORT II

Localization Site	top1 (%)[a]	sp (%)[b]	top2[c] (%)	sp (%)[b]	n
Cytoskeletal	1.4	4.3	15.0	18.2	51
Cytoplasmic	67.2	4.9	87.1	5.7	386
Nuclear	71.7	6.0	90.7	3.6	510
Mitochondrial	49.9	8.7	77.5	5.7	253
Secretary vesicles	0.0	0.0	0.0	0.0	25
ER	35.3	9.7	66.0	13.2	79
Golgi	0.0	0.0	2.0	6.0	30
Vacuolar	0.0	0.0	0.0	0.0	30
Plasma membrane	63.4	10.7	76.3	9.7	110
Peroxisomal	35.0	39.1	35.0	39.1	18
Extracell./cell wall	62.3	19.5	69.8	24.0	39
total	57.4	3.5	76.7	3.3	1531

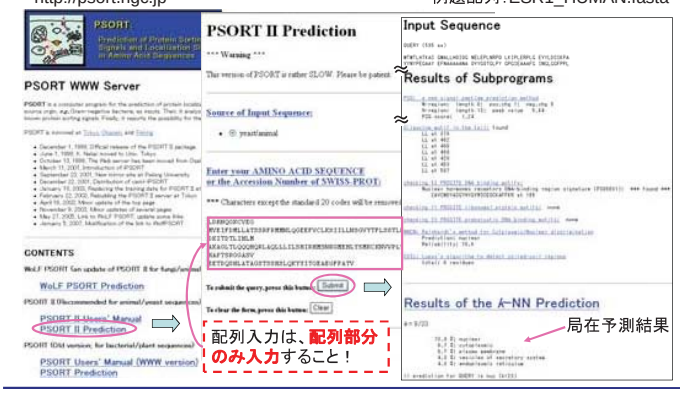
[a]Percentage of proteins whose localization sites are correctly predicted to be the most probable site. [b]Standard deviation of ten trials. [c]Percentage of proteins whose localization sites are included in the two most probable sites. Abbreviation: n, number of sequences.

PSORT II（局在解析）

実習④

http://psort.hgc.jp

例題配列：ESR1_HUMAN.fasta

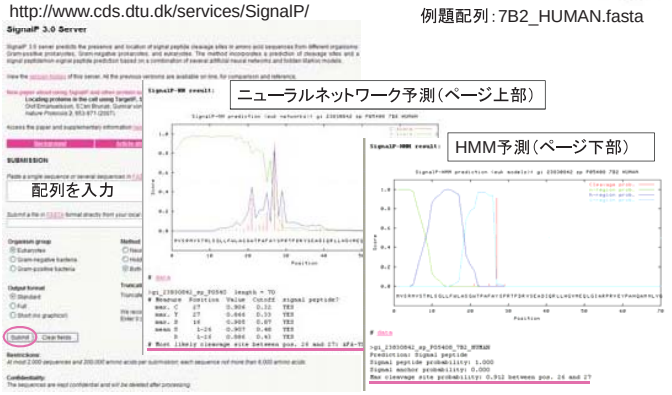


SignalP（シグナルペプチド予測）

実習④

http://www.cds.dtu.dk/services/SignalP/

例題配列：7B2_HUMAN.fasta

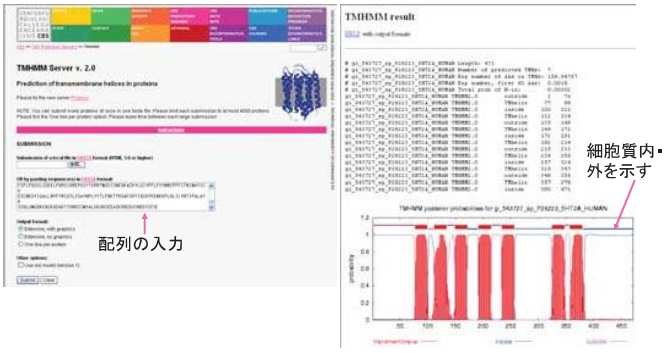


TMHMM（膜貫通トポロジー予測）

実習④

http://www.ebi.ac.uk/services/TMHMM/

例題配列：5HT2A_HUMAN.fasta



Disorder予測

Disorderが関連する機能
- DNA/RNA/protein recognition
- Modulation of specificity/affinity of protein binding
- Molecular threading
- Activation by cleavage

Disorder領域を配列組成、物理化学的指標、配列プロファイルで表現

ニューラルネットワーク、SVM等による分類学習と予測

代表的なツールにDISOPRED、PONDRなど

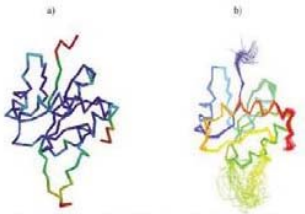


Fig. 3. The DISOPRED prediction for a structure solved by NMR is shown (Target 170). a) shows the DISOPRED prediction. Residues are coloured according to the estimated confidence that a residue is disordered, with red indicating the regions predicted with highest confidence and blue indicating the regions least likely to be disordered. b) shows the structural ensemble for the target. The local variability of the NMR ensemble clearly demarks regions which could be disordered, with good agreement with the disorder prediction. Colouring in this case is according to position in the sequence from residue 1 (blue) to residue 69 (red).

DISOPREDによるゲノム解析

Ward *et al*., *J Mol Biol*, 2004

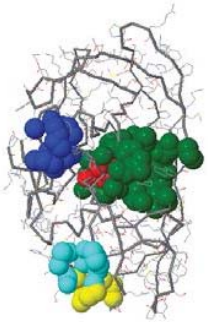
Table 2. Estimated disorder frequencies

Kingdom organism	Number of sequences	Disorder frequency	Length >30	Length >50
Archaea *Aeropyrum pernix*	1841	4.7	2.1	0.5
Archaea *Archaeoglobus fulgidis*	2409	2.8	0.9	0.2
Archaea *Halobacterium sp.*	2442	6.2	5.0	1.9
Archaea *Methanococcus jannaschii*	1784	2.8	1.0	0.3
Archaea *Pyrococcus abyssi*	1769	3.0	1.4	0.7
Archaea *Thermoplasma volcanium*	1497	3.2	1.0	0.3
Bacteria *Aquifex aeolicus* VF5	5288	6.4	5.7	2.0
Bacteria *Chlamydophila pneumoniae* AR39	1557	3.3	1.9	0.4
Bacteria *Chlamydophila pneumoniae* AR39	1111	6.2	4.8	2.3
Bacteria *Chlorobium tepidum* TLS	2248	5.1	3.3	0.5
Bacteria *Escherichia coli* K12	4247	4.6	2.8	0.8
Bacteria *Haemophilus influenzae* Rd	1650	4.4	3.8	1.3
Bacteria *Mycobacterium tuberculosis* H37Rv	3890	9.1	7.0	3.3
Bacteria *Neisseria meningitidis* MC58	2020	5.7	4.5	1.7
Bacteria *Salmonella typhi*	4714	4.9	2.7	0.9
Bacteria *Staphylococcus aureus*	2632	6.2	4.5	2.2
Bacteria *Synechocystis* species PCC 6803	3140	5.4	4.7	1.8
Bacteria *Thermotoga maritima*	1857	3.3	1.8	0.6
Bacteria *Treponema pallidum*	1035	6.1	6.4	2.6
Eukaryota *Arabidopsis thaliana*	21,482	16.8	33.8	19.0
Eukaryota *Caenorhabditis elegans*	20,506	15.9	27.5	15.6
Eukaryota *Drosophila melanogaster*	13,913	21.6	36.6	22.1
Eukaryota *Homo sapiens*	76,385	21.6	35.2	21.9
Eukaryota *S. cerevisiae*	6245	17.0	31.2	19.3
Archaea	11,742	3.8	2.0	0.7
Bacteria	35,389	5.7	4.2	1.6
Eukaryota	88,531	18.9	33.0	19.6
PDB (non-redundant at 95% sequence identity)	7169	3.2	0.5	0.1

The columns show the number of sequences, the percentage of residues predicted as being disordered and the percentage of chains with contiguous disordered segments of length greater than 30 and 50 residues, respectively.

結合部位の同定

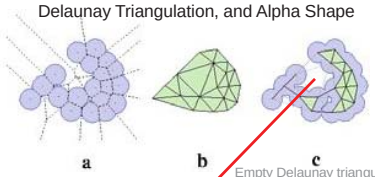
- タンパク質の立体構造情報（X線、NMR、予測構造含む）に基づき、グリッドやプローブ球などを用いて、表面形状から活性ポケットを同定するアプローチ
- 代表的なソフトウェア
  - LIGSITE
  - PASS
  - CASTp
  - PocketFinder



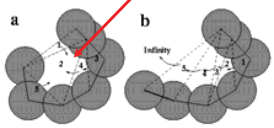
CASTpによる解析例

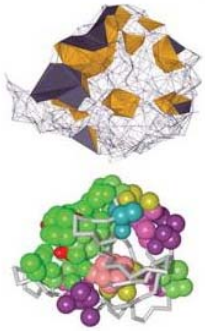
CASTpのアルゴリズム

1. Computational Geometry: Voronoi Diagram, Delaunay Triangulation, and Alpha Shape



2. Discrete Flow





staphyloccocal nucleaseのalpha shape モデル（上）とBinding site（下：緑のCPK）, Jiang *et al*.,1998より

CASTpを用いた統計解析例

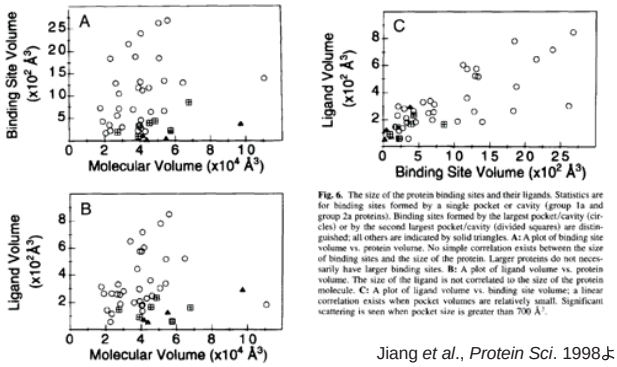


Fig. 6. The size of the protein binding sites and their ligands. Statistics are for binding sites formed by a single pocket or cavity (group 1a and group 2a proteins). Binding sites formed by the largest pocket/cavity (circles) or by the second largest pocket/cavity (divided squares) are distinguished; all others are indicated by solid triangles. **A:** A plot of binding site volume vs. protein volume. No simple correlation exists between the size of binding sites and the size of the protein. Larger proteins do not necessarily have larger binding sites. **B:** A plot of ligand volume vs. protein volume. The size of the ligand is not correlated to the size of the protein molecule. **C:** A plot of ligand volume vs. binding site volume; a linear correlation exists when pocket volumes are relatively small. Significant scattering is seen when pocket size is greater than 700 Å³.

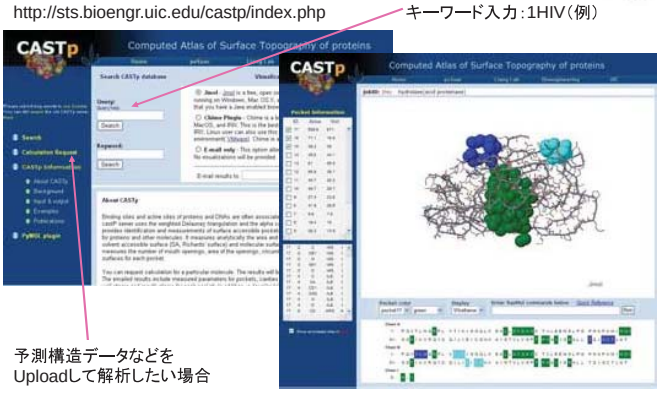
Jiang *et al*., *Protein Sci*. 1998より

CASTp（活性ポケット探索）

実習⑤

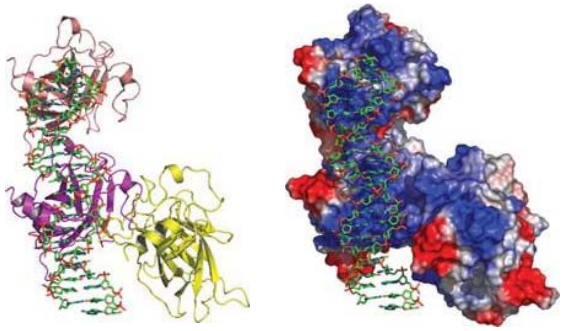
http://sts.bioengr.uic.edu/castp/index.php

キーワード入力：1HIV（例）



予測構造データなどを
Uploadして解析したい場合

静電ポテンシャル解析

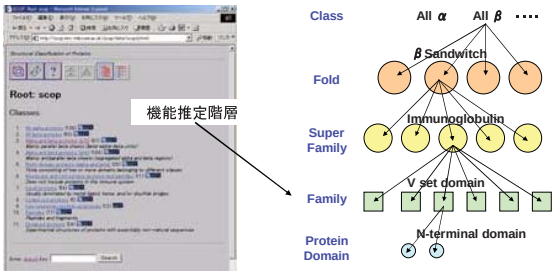


P53/DNA複合体（1TUP）の立体構造（左）と静電ポテンシャル解析結果（右）：共にPyMOLによるグラフィックス表示

構造分類データベース

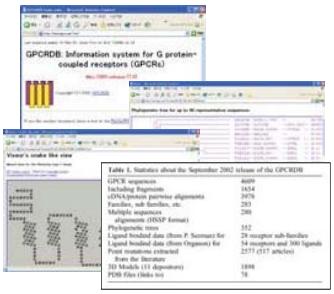
http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html

ホモロジーモデリングや、Threading法で用いたテンプレートタンパク質の分類を参照して、対象タンパク質の機能を推定

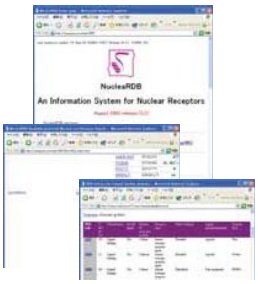


受容体セレクションデータベース

配列、構造、変異残基、リガンド結合定数、シグナルモチーフ情報など

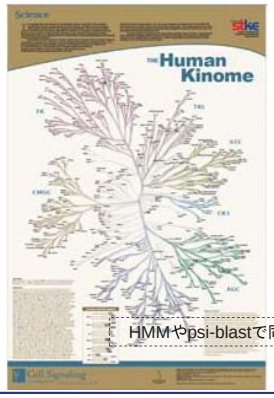


GPCRDB：Gタンパク質共役型受容体



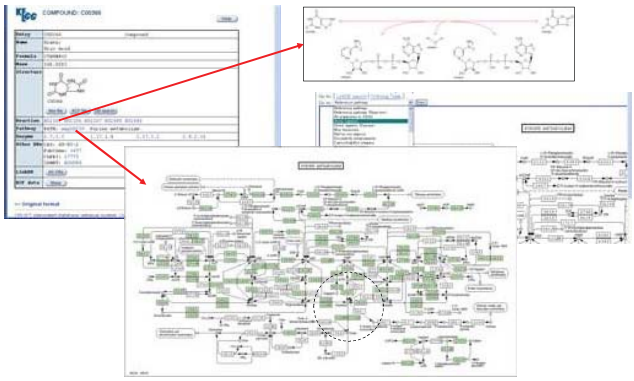
NucleaRDB：核内受容体

Kinome (Manning *et al.*, *Science*, 2002)



HMMやpsi-blastで同定

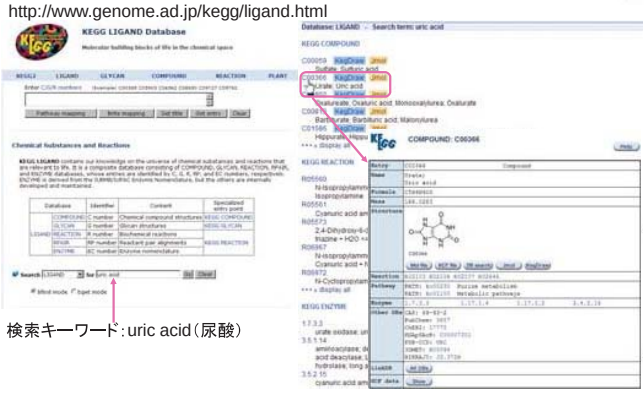
パスウェイデータベース（KEGG）



KEGG Ligand（リガンドデータベース）

実習⑥

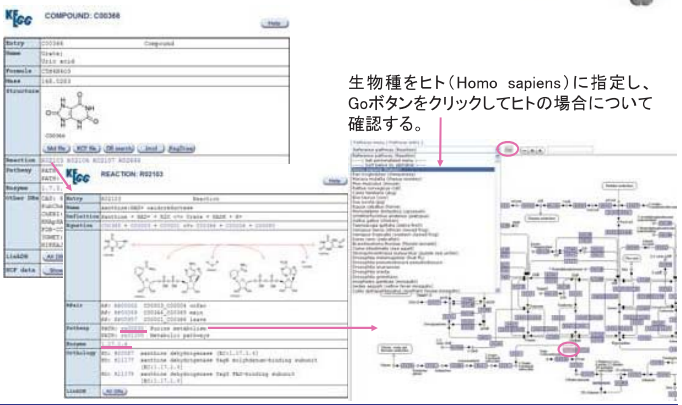
http://www.genome.ad.jp/kegg/ligand.html



検索キーワード：uric acid（尿酸）

パスウェイマップの表示（1）

実習⑥

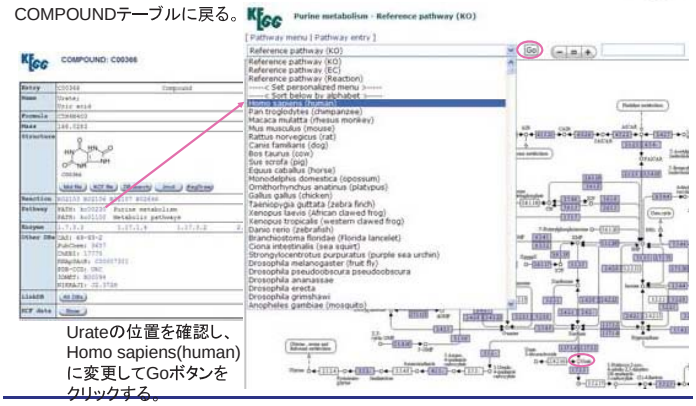


生物種をヒト（Homo sapiens）に指定し、Goボタンをクリックしてヒトの場合について確認する。

パスウェイマップの表示（2）

実習⑥

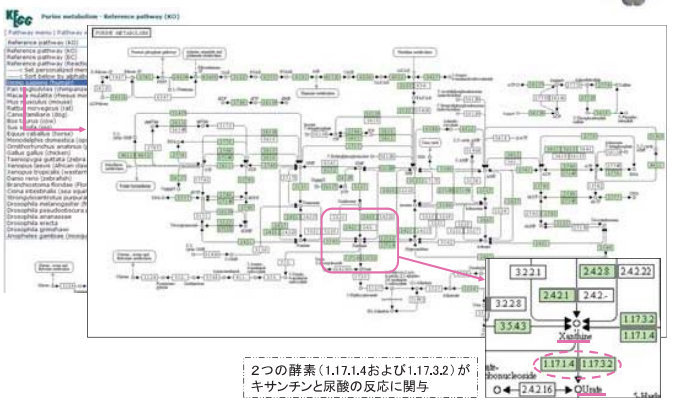
COMPOUNDテーブルに戻る。



Urateの位置を確認し、Homo sapiens(human)に変更してGoボタンをクリックする。

パスウェイマップの表示（3）

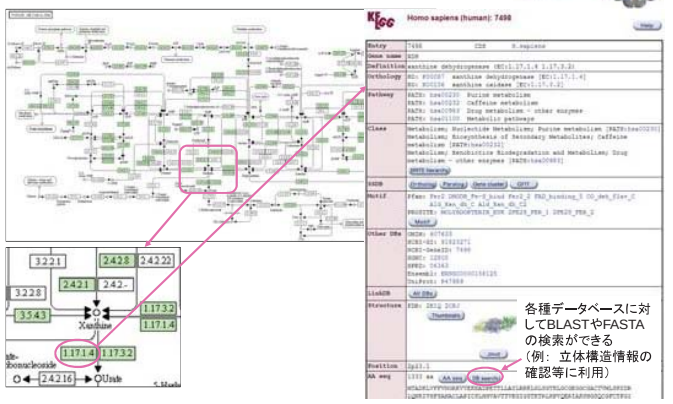
実習⑥



2つの酵素（1.17.1.4および1.17.3.2）がキサンチンと尿酸の反応に関与

関連酵素の情報

実習⑥



各種データベースに対してBLASTやFASTAの検索ができる（例：立体構造情報の確認等に利用）

DrugBank

- Wishart *et al*., *NAR*, 2006
- バイオインフォマティクス（タンパク質）とケモインフォマティクス（化合物）を融合した統合データベース
- 化合物（4100以上）とその標的タンパク質を関連付けし、DrugCardと呼ばれる独自のID管理で、様々なデータ項目を登録
- 化合物構造検索のインターフェースを実装したWebデータベース

Table 1. Summary of the data fields or data types found in each DrugCard

Drug or compound information	Drug target or receptor information
Generic name	Target name
Brand name(s)/synonyms	Target synonyms
IUPAC name	Target protein sequence
Chemical structure/sequence	Target no. of residues
Chemical formula	Target molecular weight
PubChem/KEGG/CMEB Links	Target pI
Swiss-Prot/GenBank Links	Target gene ontology
FDA/MSDS/RxList Links	Target general function
Molecular weight	Target specific function
Melting point	Target pathways
Water solubility	Target reactions
pKa or pI	Target Pfam domains
LogP or hydrophobicity	Target signal sequences
NMR/MS spectra	Target transmembrane regions
MOL/SDF/PDF text files	Target essentiality
MOL/PDB image files	Target GenBank protein ID
SMILES string	Target Swiss-Prot ID
Indication	Target PDB ID
Pharmacology	Target cellular location
Mechanism of action	Target DNA sequence
Biotransformation/absorption	Target chromosome location
Patient/physician information	Target locus
Metabolizing enzymes	Target SNPs/mutations

A more complete listing is provided on the DrugBank home page.

DrugBank（統合データベース）

実習⑦

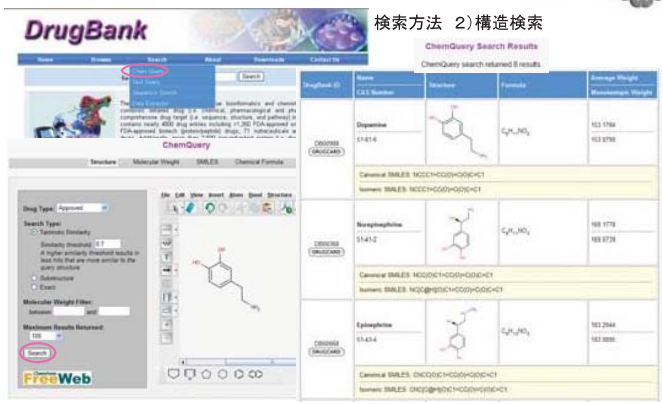
http://redpoll.pharmacy.ualberta.ca/drugbank/

検索方法　1）
キーワード入力（例dopamine）

DrugBank（統合データベース）

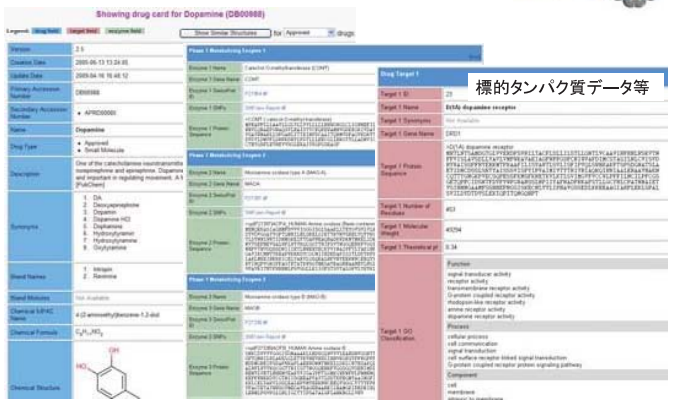
実習⑦

検索方法　2）構造検索



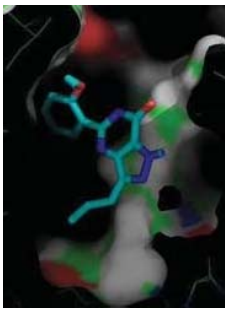
DrugBank（検索結果）

実習⑦



まとめ

- 大まかな標的タンパク質同定には、局在予測やドメイン（モチーフ）が有効
  - 分子機能の予測とあわせると信頼性が向上（例：Zn結合モチーフが同定され、なおかつ局在予測で核内と予測される場合など）
- 結合ポケットの同定技術で、構造ゲノミクスプロジェクトと関連して今後も重要となる
- 標的タンパク質同定を効率化させる統合化データベースの開発や検索条件の多様化



第4章

タンパク質-リガンド結合予測
Ligand Docking
Docking Simulation
Structure-Based Drug Design

創薬研究の推移

創薬研究：　生物系と化学物質の選択的相互作用の解明

世代	主なアプローチ、手法
第一世代（1940年代～）	抗生物質の発見・改良・合成
第二世代（1960年代～）	生理学的な研究に基づく受容体拮抗、イオンチャネル阻害、酵素阻害
第三世代（1970年代～）	分子生物学やバイオテクノロジーに基づくバイオ医薬品
第四世代（1980年代～）	遺伝子治療、転写因子阻害剤などによる難病への遺伝子レベルのアプローチ。化学物質については、コンビナトリアルケミストリ、ハイスループットスクリーニングなど
第五世代（2000年代～）	ヒトの遺伝子情報の解明、バイオインフォマティクスによるゲノム創薬。構造ゲノミクスによる受容体構造の活用。

化合物（創薬分子）の多様性

- **ヒトの遺伝子　≒　2万2千**
- **ヒト塩基配列　≒　30億**
- **ヒトの細胞数　≒　60兆**

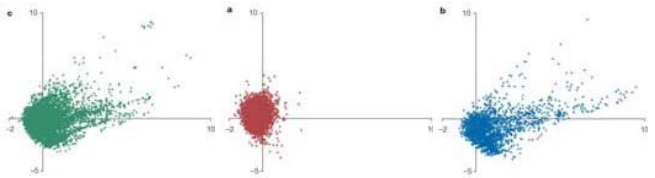
"the total number of possible small organic molecules that populate 'chemical space' has been estimated to exceed — an amount so vast when compared to the number of such molecules we have made, or indeed could ever hope to make, that it might as well be infinite."

Nature 432, 823, 2004より

Nature 432, 823 (16 December 2004)
Nature chemical space特集記事

宝（創薬分子）探しの旅

化合物空間における広がり比較



上市医薬品　　人が作った合成化合物　　天然物

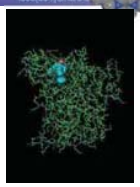
Nature 432, 824-828(16 December 2004)より

※キラリティや芳香族環の寄与がコンビケム合成だけでは不十分

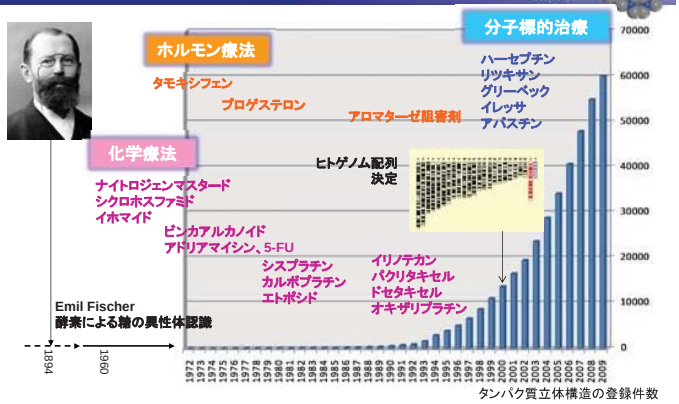
ただ、やみくもに探索しても新薬にたどり着けない

分子標的創薬のコンセプト

- 「鍵と鍵穴」理論
  - 酵素による糖の異性体認識（Emil Fischer, 1894）
- Structure-based drug design(SBDD)分野の発展
  - X線（主にリガンドとの共結晶構造）、NMR、計算機モデリング構造の活用
  - 計算機によるリガンド再挿入実験の高精度化（Friesner et al., J. Med. Chem., 2006）

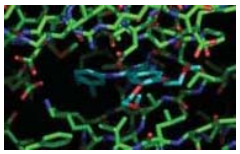


分子標的創薬の歴史（抗がん剤を例に）



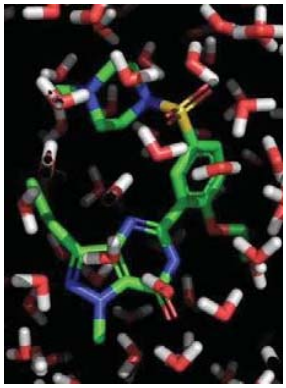
ドッキング問題

- 計算機的手法による非共有結合複合体予測
  - シグナル伝達、免疫応答、酵素触媒
  - 低分子化合物のもつ物理化学的、薬理学的な機能の分子認識機構から解明
  - 立体構造情報に基づく、新しいリード化合物発見、薬物候補の探索（structure-based drug design）
- タンパク質結合部位における自由エネルギー最安定のリガンド（DNA、RNA、タンパク質）結合状態を現実的な時間範囲内で検索、極小化する問題

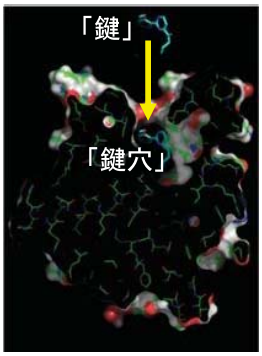


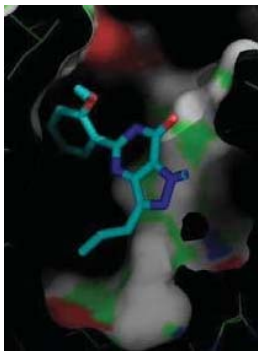
計算機による再挿入ドッキングの様子（青が天然でのリガンド結合状態）

でも難しい「鍵と鍵穴」理論（その壱）



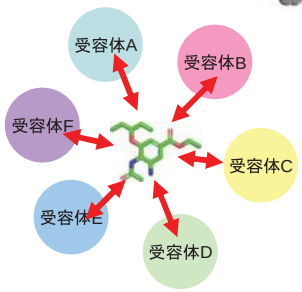
でも難しい「鍵と鍵穴」理論（その弐）





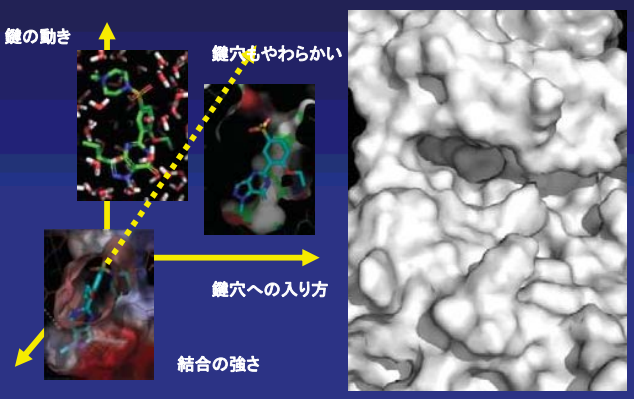
でも難しい「鍵と鍵穴」理論（その参）

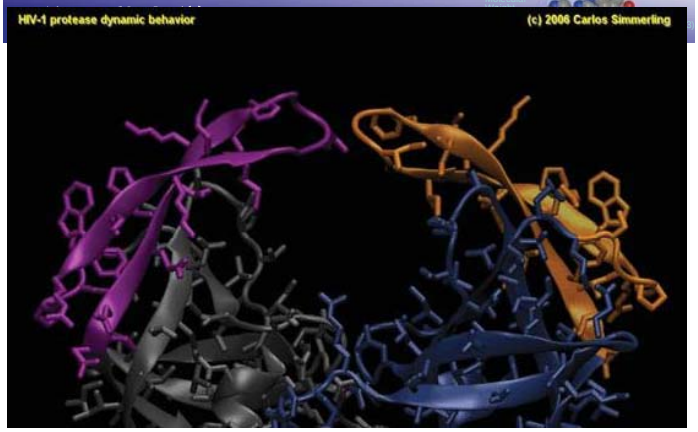
万能鍵があれば・・・・



本来の標的受容体以外にも作用すると副作用を引き起こす原因となる（選択性が大事）

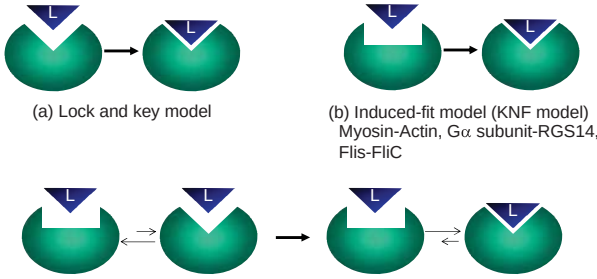
膨大な探索空間・・・・





SBDDにおける課題

Protein ligand binding mechanisms



(a) Lock and key model

(b) Induced-fit model (KNF model)
Myosin-Actin, Gα subunit-RGS14, Flis-FliC

(c) Pre-existing equilibrium model (MWC model)
Antibody Spe7-TrxShear3, Gplb-IX-V-Thrombin, NtrC-P

C.-S. Goh *et al*., *Curr. Opin. Struct. Biol*. 14, 2004.

リガンドドッキング基礎：探索とスコアリング

- ドッキング計算では「探索」と「スコアリング」の問題解決が常に要求される
- 探索
  - 相互作用分子間の可能な配座、配向空間を調べる
  - 何らかのrigid拘束条件を取りいれない限り、全探索は困難。精度と効率化が重要
  - 大局的極小構造近辺での確実なエネルギー極小化（精度）
  - 相互作用に無関係な探索空間の省略（一方で、安定配座付近での徹底的な探索）（効率化）
  - 組み合わせ問題のためのアルゴリズムや確率的手法による探索アルゴリズム

探索とスコアリング（続き）

- スコアリング
  - 真の結合自由エネルギーをスコア関数で近似
  - スコア関数
    - native/non-native結合状態の判別
    - 実験値と相関が観測できるような探索過程で得られた配座、配向のランク付け
    - 特定のタンパク質に対する複数のリガンドドッキングを実施した際の比較→バーチャルスクリーニングの本質
    - 複数のタンパク質 vs 複数のリガンドドッキング計算で正しい選択性を示すこと（クロスドッキング計算）

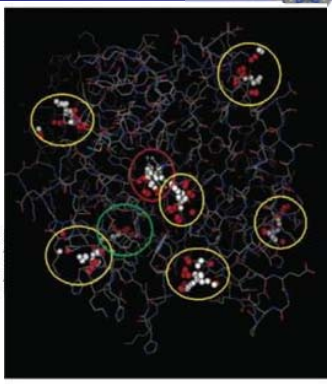
受容体の準備

全原子系で取り扱うことは困難。部分領域の指定や、偽分子などを用いた表現で近似

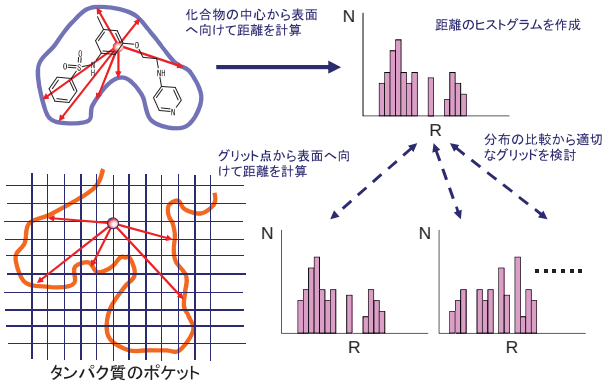
1. ドッキング部位の同定
   - 部位特異的変異実験、類似化合物の共結晶構造を参照
   - 活性部位同定プログラムを利用（CASTp etc）
2. タンパク質表面（形状）の表現・近似
   - 幾何的形状記述子（表面球、低密度点による表面（多面体等）の表現）
   - 幾何的形状記述子発展型（DOCKなどに用いられるポケット形状に相補的な球）
   - 相互作用特性を表現した記述子（LUDI、FlexX）
   - グリッド表現

ドラッガブルポケット予測：PLB index

- Propensity for Lligand Binding index（Soga et al., JCIM, 2007）
- 41の高品質X線構造をTraining Setとして活性部位を形成するアミノ酸残基の傾向指標を作成
- Test Setとして用いた804リガンドの結合ポケットの推定では、86%の高い判別率を達成
- ポケット候補はMOEのAlpha Site Finderを利用



タンパク質-リガンドドッキング

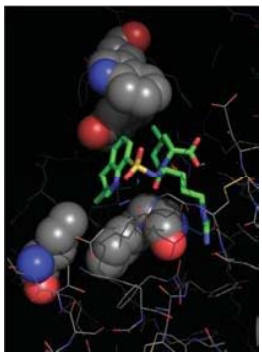


スコアリングの基本形

水素結合、疎水性相互作用など各相互作用項の総和で結合自由エネルギーを表現

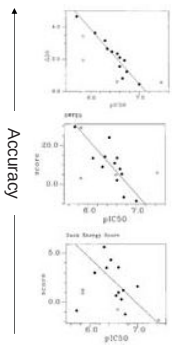
$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{H-bonds} f(\Delta r, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic\,int} f(\Delta r, \Delta\alpha) + \Delta G_{lipo} A_{lipo} + \Delta G_{rot} N_{rot}$$

化合物の自由度



リガンド - 受容体相互作用計算

- QM/FEP/TI
  - Quantum Mechanics
  - Free Energy Perturbation
  - Thermal dynamics integration
- Molecular Mechanics
  - Force field based
  - MM/PBSA
- Empirical
  - Empirical function
  - Knowledge-based



CPU time

Months

MD:Weeks

MD/MC:days

Minute

参考文献：Pealman *et.al.*, *J. Med. Chem.* 2001, **44**, 3417-3423.

まとめ

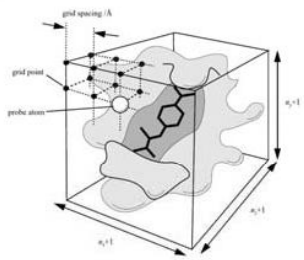
- ドッキング結果の考察ポイント
  - ドッキング（探索）が十分に行われているか
    - リガンド配座探索手法の検討
    - 受容体モデルの検討（特にモデリング構造の場合）
    - Induced Fit Dockingの検討
  - スコアリング
    - スコアの補正、統計解析の利用（コンセンサス法、相互作用パネル（福西ら））
    - 高精度計算の適用
  - 目的に応じた手法の使い分け（シード探索、リード探索、optimization）

AutoDock実習

- AutoDockサイト（ver.4）
  - http://autodock.scripps.edu/
- 実習手順
  - インストール
    autodock4.2.3_win32.exe　ドッキング計算プログラム
    MGLTools-1.5.4-Setup.exe　分子表示ツールGUI
    python-2.5.2.msi　GUI起動に必要
    - 実習用のインストール先は全てC:¥docking下を指定
  - 結晶構造を用いた再挿入実験
    Biotin-Streptavidin複合体を利用（PDB-ID: 1MK5）
    - 実習では、タンパク質と化合物を分離した1MK5_A_rep.pdbと1MK5_A_lig.pdbを予め用意
  - ドッキング計算結果の可視化

AutoDock 4.0

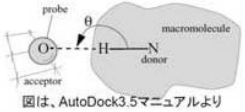
(Morris et al., J Comput Chem 19, 1998)



自由度は、化合物の内部回転自由度、並進移動、回転行列

遺伝的アルゴリズムによる探索

受容体はグリッドと呼ばれる格子点で表現さる。格子点では、予め複数の物理化学的性質をもつプローブ原子によってエネルギー計算されている。→高速化



図は、AutoDock3.5マニュアルより

AutoDock 4.0

(Morris et al., J Comput Chem 19, 1998)

スコア関数

$$\Delta G = \Delta G_{\mathrm{vdW}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \quad \text{vdW相互作用}$$

$$+ \Delta G_{\mathrm{hbond}} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{\mathrm{hbond}} \right) \quad \text{水素結合相互作用}$$

$$+ \Delta G_{\mathrm{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \quad \text{静電相互作用}$$

$$+ \Delta G_{\mathrm{tor}} N_{\mathrm{tor}} \quad \text{結合に伴うエントロピー変化}$$

$$+ \Delta G_{\mathrm{sol}} \sum_{i,j} S_i V_j e^{(-r_{ij}^2/2\sigma^2)} \quad \text{脱水和エネルギー}$$

インストール

autodock4.2.3_win32.exe
- デフォルトでは、C:¥Windows¥system32¥がインストール先として指定されますが、インストール時には、C:¥docking¥にインストールします（フォルダを選択する場合には、C:¥docking¥system32¥となりますが、手打ちでC:¥docking¥と書き直します）。インストールが終了するとC:¥docking¥下にautodock4.exe、autogrid4.exe、cygwin1.dllが作成されます。

python-2.5.2.msi
- 指示に従ってインストールします。インストール先のフォルダは、C:¥docking¥を選択します。

MGLTools-1.5.4-Setup.exe
- MGLToolsのインストールについても、インストール先をC:¥docking¥とします。インストール途中で右のウィンドウが表示されましたら、Non commercialを選択し、下のContinueをクリックします。起動用のショートカットは、C:¥USERS¥guest〇〇（ログインID）¥デスクトップ¥の中に作成されています。
- AutoDockToolショートカットアイコン のプロパティで、作業フォルダをC:¥dockingに変更しておきます。

Streptavidinとは　ウィキペディアより

ストレプトアビジン（Streptavidin）はストレプトマイセスの一種Streptomyces avidinii により作られるタンパク質であり、性質はアビジンとよく似ている。研究・検査用に利用されている。アビジンと同様に、ビオチンを非常に強く結合する特性がある。解離定数（Kd）は約10-15 mol/Lで、非共有結合の中では最も強い部類に属す。分子量53,000ダルトン（アビジンより少し小さい）の4量体を形成し、各モノマーがビオチン1分子を結合する。

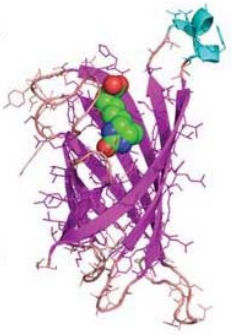
一方、アビジンとは次のような違いがある。変性により解離するが、ストレプトアビジンの方が変性に強い。アビジンにはついている糖鎖がなく、水溶性が低い。また等電点は弱酸性または中性（アビジンは塩基性）である。これらにより非特異的結合が少ないという利点があり、アビジンよりも多く利用されている。

なお以上の欠点を改良する目的で、糖鎖を除去したアビジン（NeutrAvidin）も使われている。

利用

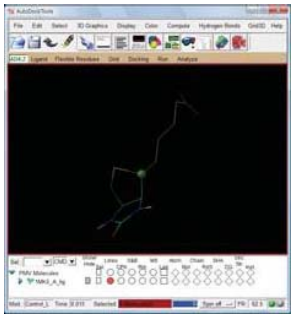
特定分子の検出、固定化、単離などに広く用いられる。例えば、DNA断片や抗体をビオチンで標識しておくと、これらはそれぞれ相補的なDNA・RNA配列、および対応する抗原分子を標的として結合する性質がある。また、ストレプトアビジンに色素または色素を生成する酵素を結合しておき、これを既に標的と結合したDNAまたは抗体と反応させることで、標的の可視的な検出あるいは定量ができる。これはブロッティング（ノーザン・サザン・ウェスタンの各方法）やELISAに応用される。

ストレプトアビジンを固定化しておけば、ビオチン標識した分子を固定化できる。これは標的分子の分離（アフィニティクロマトグラフィー等）にも応用される。



リガンドの準備

1. ADT:Ligand/Input/Open…
   - ファイルの種類をall files
   - 1MK5_A_lig.pdbを選択。ダイアログは[OK]
2. ADT:Ligand/TorsionTree/Detect Root…
3. ADT:Ligad/TorsionTree/Choose Torsions…[Done]
4. Ligand/Output/Save as PDBQT…
   - 1MK5_A_lig.pdbqtで保存
5. DashbordのShow/Hideを[off]



グリッドの作成

1. ADT:Grid/Macromolecule/Open
   - ファイルの種類:all files
   - 1MK5_A_rep.pdbを読み込む
   - ポップアップ:[OK]
   - ファイル保存ウィンドウ:
   - 1MK5_A_rep.pdbqt
2. ADT:Grid/Set Map Types/Choose Ligand…
   - Select a molecule: 1MK5_A_lig [Select Ligand]
3. ADT:Grid/Grid Box…
   - Ceter/Center on Ligand
   - 1MK5_A_ligのShow/Hide [on]で位置の確認
   - File/close saving current
4. ADT:Grid/Output/Save GPF
   - 1MK5_A_rep.gpf [保存]

グリッドの作成(続き)

- コマンドプロンプトでC:¥dockingに移動後、以下のコマンドを実行

```
C:¥docking>
C:¥docking>autogrid4.exe -p 1MK5_A_rep.gpf -l
  1MK5_A_rep.glg [Enter]
```

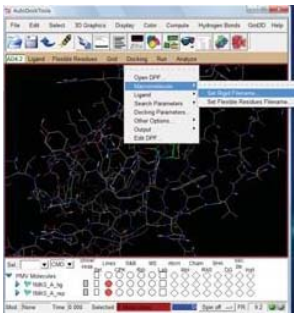
※数分かかります。

```
autogrid4.exe: Successful Completion
```

※正常終了

ドッキングパラメータファイルの作成

1. ADT:Docking/Macromolecule/Set Rigid Filename
   - 1MK5_A_rep.pdbqtを読み込む
2. ADT:Docking/Ligand/Choose/
   - 1MK5_A_lig [Select Ligand], [Close]
3. ADT:Docking/Search Parameters…/Genetic Algorithm…
   - Maximum Number of evals: short (実習のため) [Accept]
4. ADT:Docking/Docking Parameters…(変更の場合)
5. ADT:Docking/Output/Lamarckian GA…
   - 1MK5_A_lig.dpf (保存)



ドッキングの実行

- コマンドプロンプトでC:¥dockingに移動後、以下のコマンドを実行

```
C:¥docking>
C:¥docking>autodock4.exe -p 1MK5_A_lig.dpf -l
  1MK5_A_lig.dlg [Enter]
```

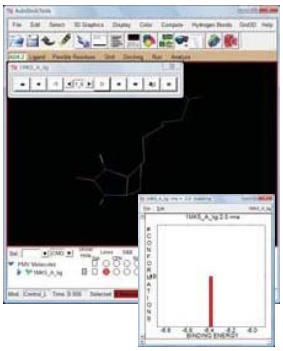
※数分かかります。

```
autodock4: Successful Completion on …
```

※正常終了

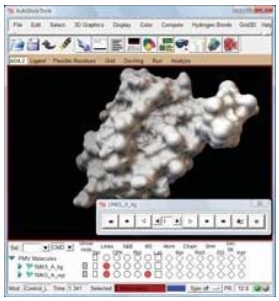
結果の可視化:データの読み込み

1. 分子をEditで削除するか、ADTを再起動
2. ADT:Analyze/Dockings/Open…
   - 1MK5_A_lig.dlgを読み込む
3. ADT:Analyze/Conformations/Load…
   - 10個のドッキングポーズを確認
4. Analyze/Clusterings/Show…
   - ポーズの結果を2.0 rmsでクラスタリングされた結果がヒストグラムで表示
   - ヒストグラムをクリックするとconformation playerが起動され、クラスターメンバーが表示



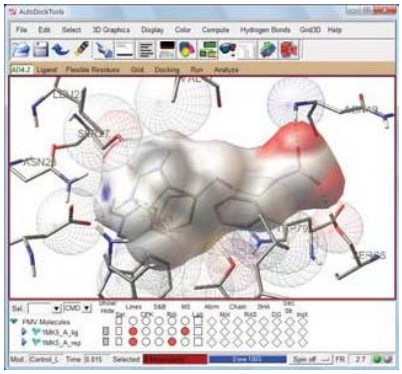
結果の可視化:相互作用の読み込み

1. ADT:Analyze/Macromolecule/Open…
   - 受容体構造が読み込まれる
2. ADT:Dashbord/MS
   - 1MK5_A_repに対応する○を[on/off]
   - 表面モデルが表示される
3. ADT:Dashbord/Rib
   - 1MK5_A_repに対応する○を[on/off]
   - リボンモデルが表示される
4. ADT:Analyze/Dockings/Show Interactions
   - 選択中のポーズについて水素結合相互作用等が可視化される
   - Conformation playでポーズを変更するとリアルタイムに解析



マウス操作

結果の可視化:相互作用の解析(続き)



表示形式設定

水素結合情報

© 2012 Takatsugu Hirokawa

関連リンク集（bitools.html）

Useful tools and database for protein bioinformatics

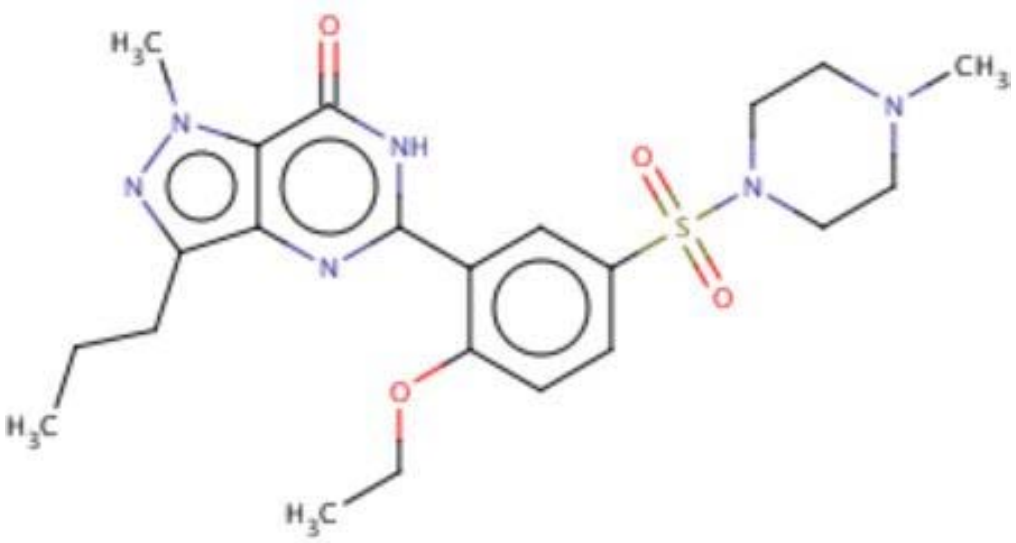
- GenomeNet
- NCBI FTP
- KEGG
- Sequence Analysis
  - BLAST
  - Primer3
  - ClustalX(exetutable only) and sequences
- Localization Prediction
  - SignalP and query sequence
  - PSORT and query sequence
  - InterPro and query sequence
  - TMHMM and query sequence
- Structure Prediction
  - Protein Data Bank
  - RasMol
  - SWISS-MODEL and query sequence
  - SWISS-MODEL results
  - Verify3D
- Function Prediction
  - SWISS-MODEL output
  - Catalytic Site Atlas
  - CASTp
- Target Database
  - DrugBank

課　題

学籍番号　　氏名

課題1：以下の化合物の名称（学術名、医薬品名どちらでも可）をDrugBankデータベース検索から同定せよ（DrugBank/Search/ChemQuery/Structure Searchを利用）

課題2：化合物名称を使ってPDBでタンパク質との複合体構造を検索し、LigandExplorerを利用して、タンパク質と相互作用に必要な原子に○をつけよ



© 2012 Takatsugu Hirokawa