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

http://www.iu.a.u-tokyo.ac.jp/~kadota/ 門田　幸二のホームページ

門田　幸二のホームページ

- 名前
  門田　幸二(かどた　こうじ)
- 所属
  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ユニット
- 身分
  特任
- 研究
  バイ
- 所属
  日本
  日本
- 研究

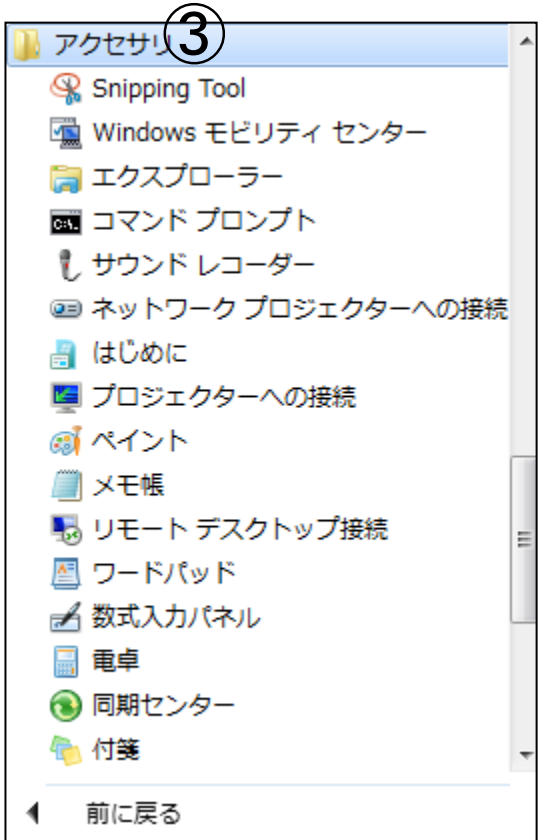
- リンク集
  - (Rで)マイクロアレイデータ解析(last modified: 2012.07.17)
  - (Rで)塩基配列解析(主に次世代シーケンサーのデータ)(last modified: 2012.07.20)
  - (マイクロアレイ)データ解析Tips(last modified: 2008.7.28)
  - 東大・院農・応生工・生物情報工学研究室
  - CBRC
  - Bioconductor
  - NCBI GEO
  - NCBI SRA
  - EMBOSS
  - BIOWEB(バイオ研究者支援サイト)
  - 新学術領域研究「複合適応形質進化の遺伝子基盤解明」ホームページ

「(Rで)塩基配列解析」と
「(Rで)マイクロアレイデータ解析」
の二つのウェブページを利用します

ご意見、ご質問はkadota@iu.a.u-tokyo.ac.jpまで。
Last modified: 2012.07.27

各種ソフトの場所

①



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）

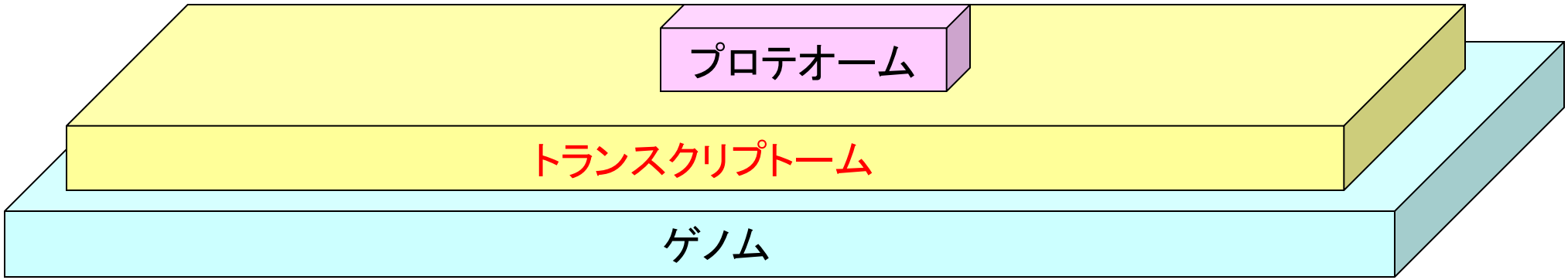
RNA　転写 (Transcription) + ome　→ トランスクリプトーム

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

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

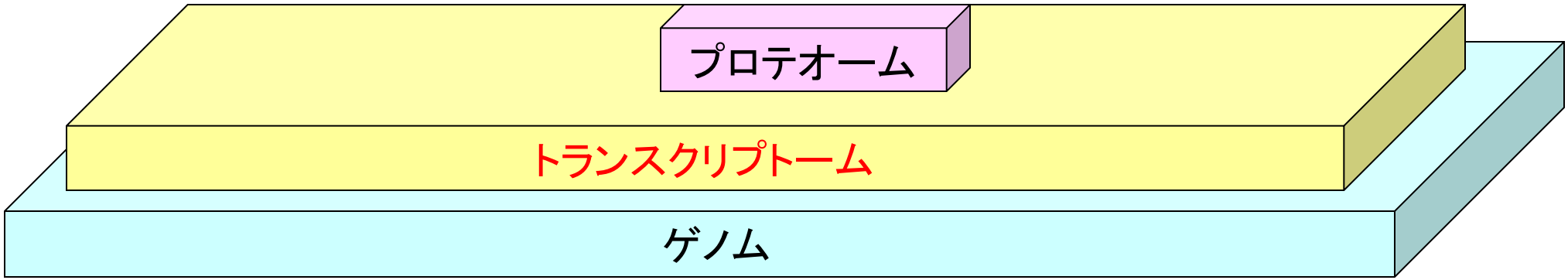


今後の実験技術

次世代シーケンサー（NGS）

GS FLX+　　5500xl SOLiD　　HiSeq 2000

二次元電気泳動法

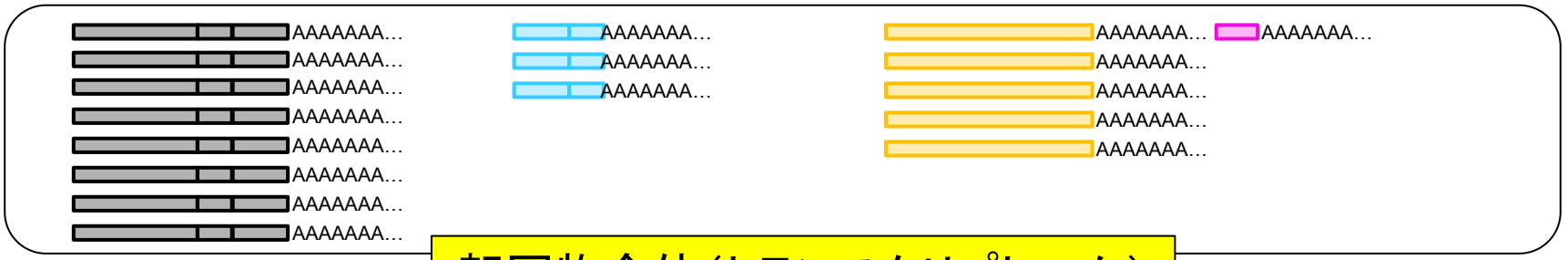


トランスクリプトームとは

- 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）

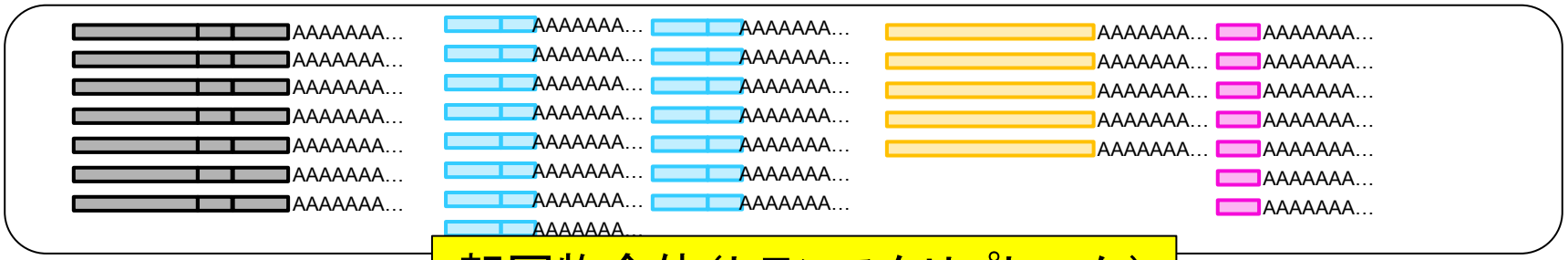
・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

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

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…）

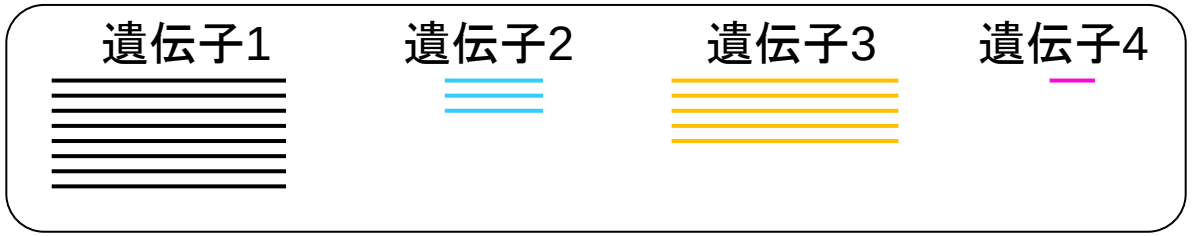


転写物全体（トランスクリプトーム）

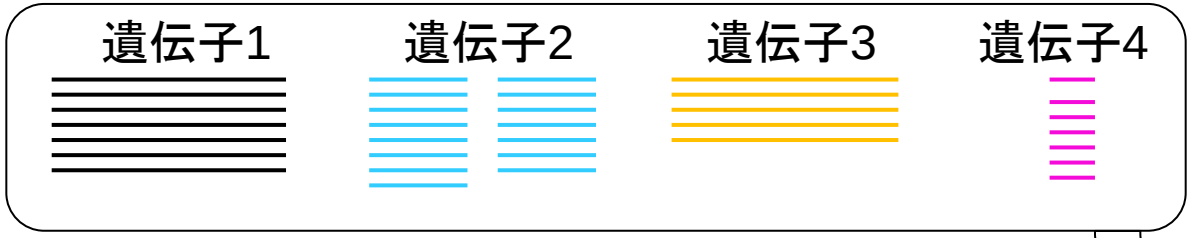
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

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



- 光刺激後（T2）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
…	…	…

- マイクロアレイ
- RNA-Seq（NGS）
- …

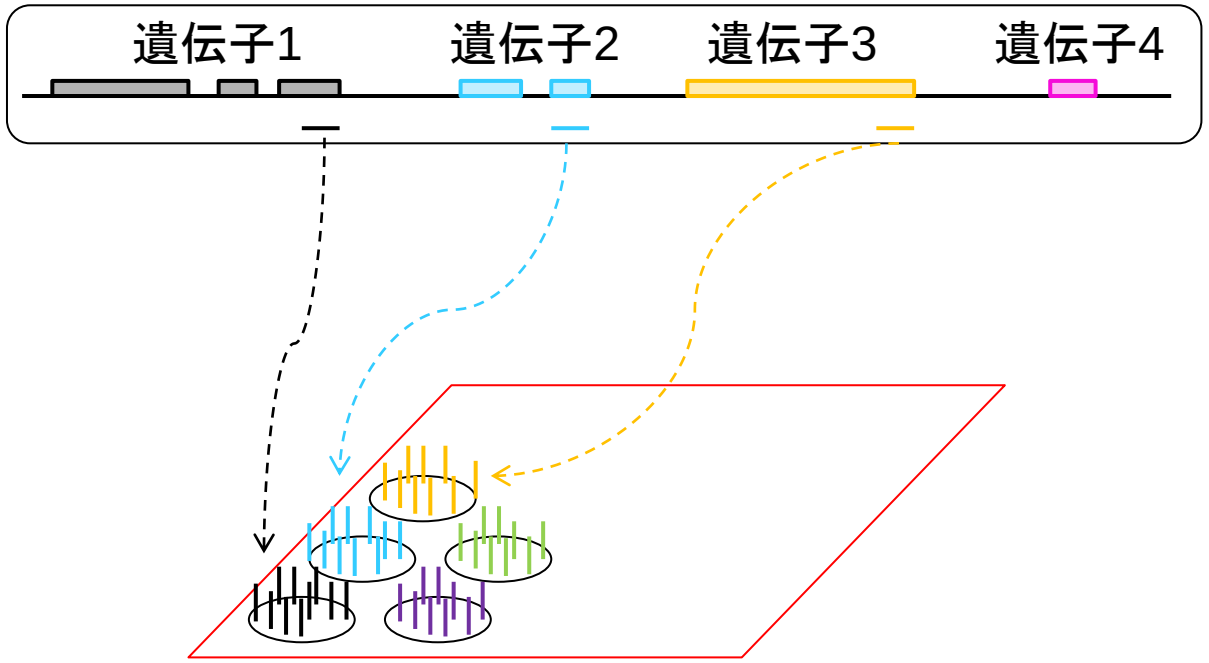
トランスクリプトームとは

- 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）

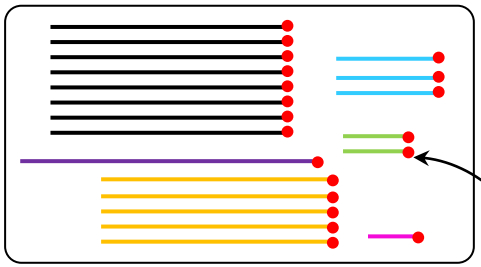
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

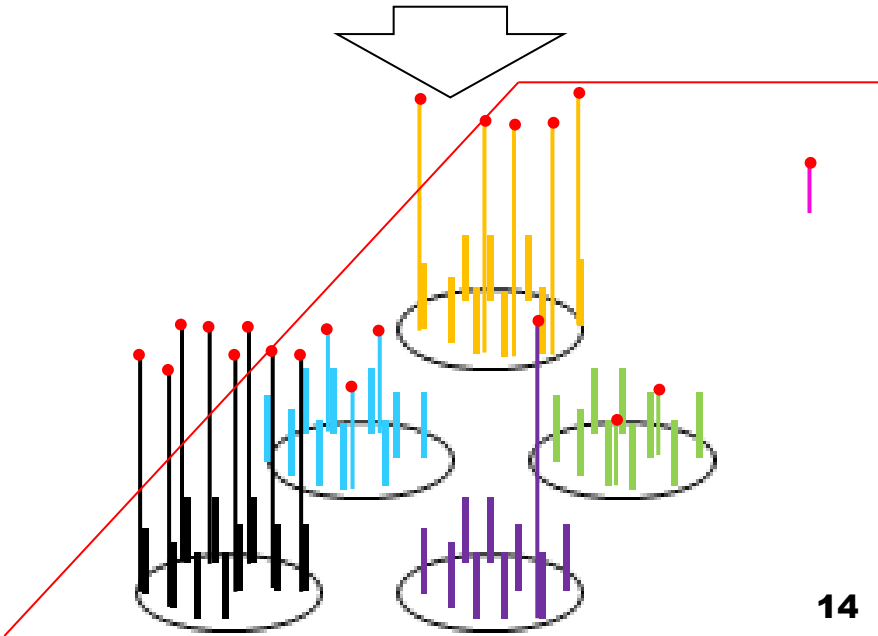
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



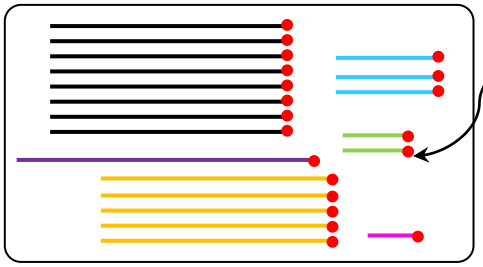
蛍光標識

ハイブリダイゼーション（二本鎖形成）

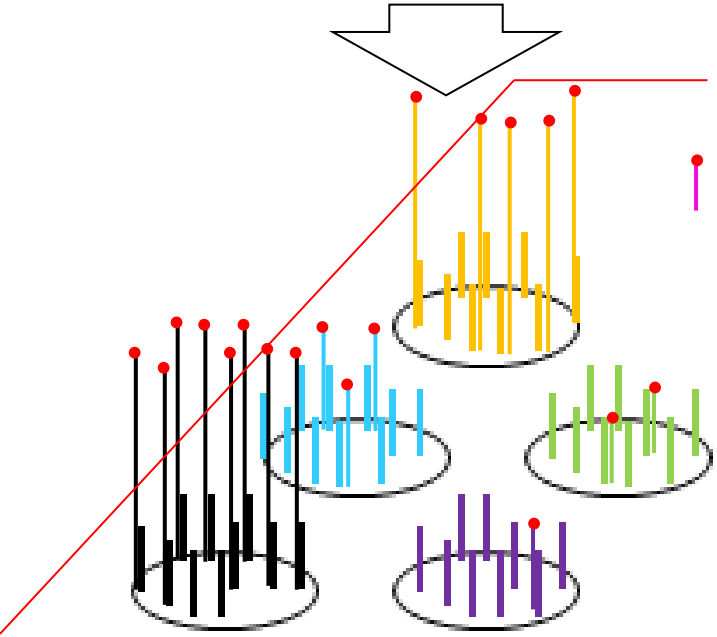


マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム

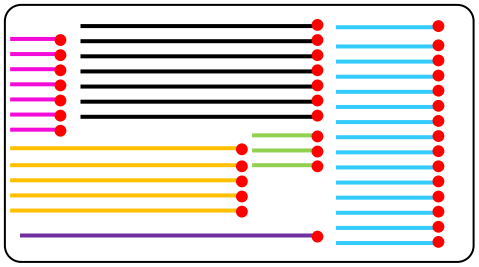


ハイブリダイゼーション
（二本鎖形成）



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



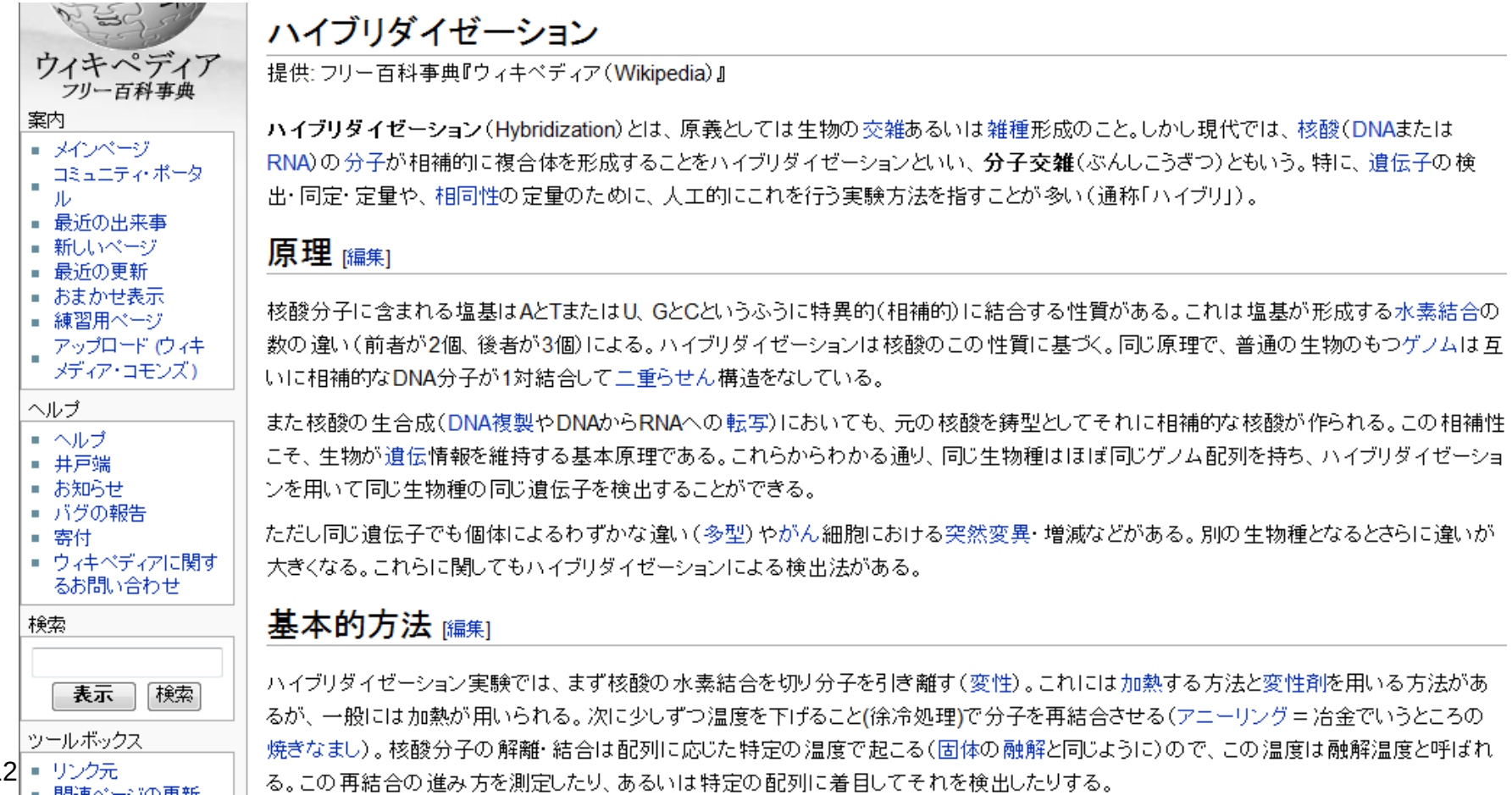
ハイブリダイゼーション
と
シグナル検出

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



ウィキペディア
フリー百科事典

案内
- メインページ
- コミュニティ・ポータル
- 最近の出来事
- 新しいページ
- 最近の更新
- おまかせ表示
- 練習用ページ
- アップロード (ウィキメディア・コモンズ)

ヘルプ
- ヘルプ
- 井戸端
- お知らせ
- バグの報告
- 寄付
- ウィキペディアに関するお問い合わせ

検索
表示 検索

ツールボックス
- リンク元

ハイブリダイゼーション

提供: フリー百科事典『ウィキペディア（Wikipedia）』

ハイブリダイゼーション（Hybridization）とは、原義としては生物の交雑あるいは雑種形成のこと。しかし現代では、核酸（DNAまたはRNA）の分子が相補的に複合体を形成することをハイブリダイゼーションといい、**分子交雑**（ぶんしこうざつ）ともいう。特に、遺伝子の検出・同定・定量や、相同性の定量のために、人工的にこれを行う実験方法を指すことが多い（通称「ハイブリ」）。

原理 [編集]

核酸分子に含まれる塩基はAとTまたはU、GとCというふうに特異的（相補的）に結合する性質がある。これは塩基が形成する水素結合の数の違い（前者が2個、後者が3個）による。ハイブリダイゼーションは核酸のこの性質に基づく。同じ原理で、普通の生物のもつゲノムは互いに相補的なDNA分子が1対結合して二重らせん構造をなしている。

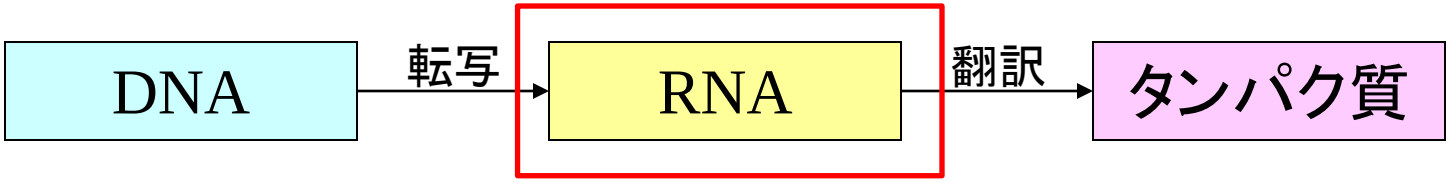
また核酸の生合成（DNA複製やDNAからRNAへの転写）においても、元の核酸を鋳型としてそれに相補的な核酸が作られる。この相補性こそ、生物が遺伝情報を維持する基本原理である。これらからわかる通り、同じ生物種はほぼ同じゲノム配列を持ち、ハイブリダイゼーションを用いて同じ生物種の同じ遺伝子を検出することができる。

ただし同じ遺伝子でも個体によるわずかな違い（多型）やがん細胞における突然変異・増減などがある。別の生物種となるとさらに違いが大きくなる。これらに関してもハイブリダイゼーションによる検出法がある。

基本的方法 [編集]

ハイブリダイゼーション実験では、まず核酸の水素結合を切り分子を引き離す（変性）。これには加熱する方法と変性剤を用いる方法があるが、一般には加熱が用いられる。次に少しずつ温度を下げること(徐冷処理)で分子を再結合させる（アニーリング＝冶金でいうところの焼きなまし）。核酸分子の解離・結合は配列に応じた特定の温度で起こる（固体の融解と同じように）ので、この温度は融解温度と呼ばれる。この再結合の進み方を測定したり、あるいは特定の配列に着目してそれを検出したりする。

トランスクリプトーム取得（RNA-Seq）

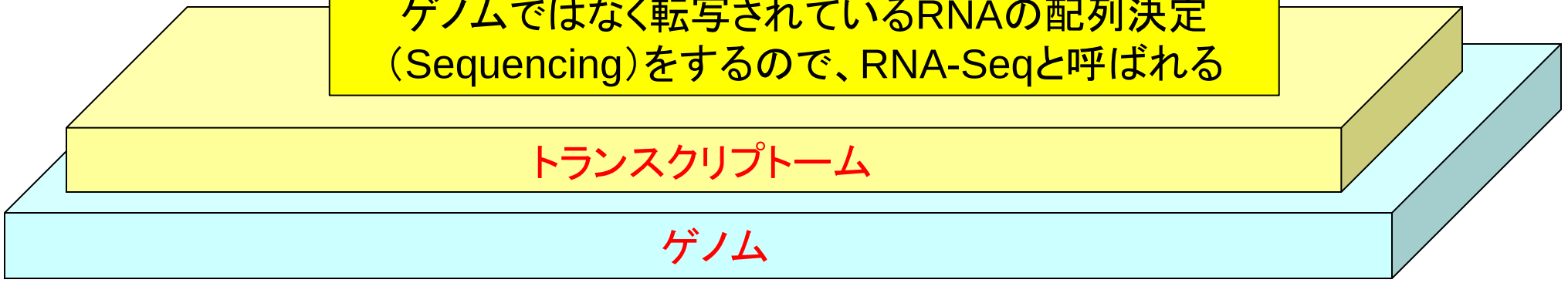


次世代シーケンサー（NGS）

GS FLX+　　5500xl SOLiD　　HiSeq 2000

二次元電気泳動法

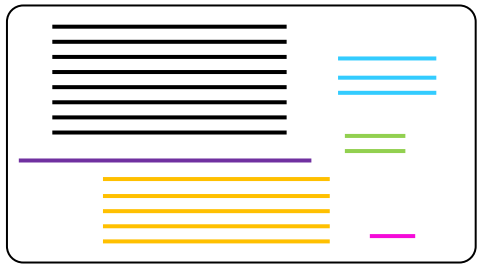
ゲノムではなく転写されているRNAの配列決定（Sequencing）をするので、RNA-Seqと呼ばれる



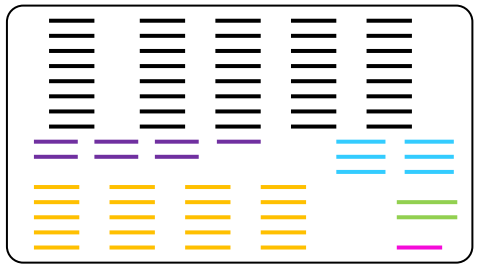
RNA-Seqデータ → 遺伝子発現行列

- 次世代シーケンサー（Illumina社の場合）

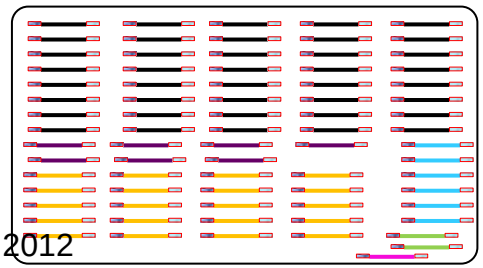
光刺激前（T1）の目のトランスクリプトーム



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

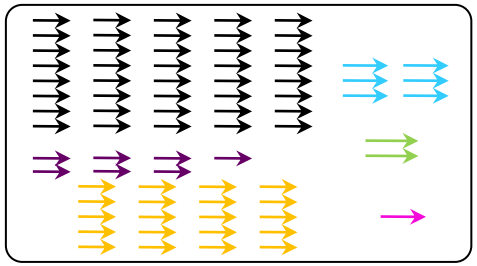


配列決定

・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

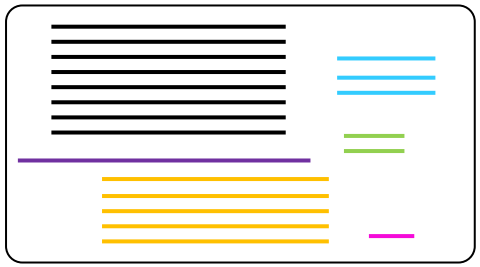
・シングルエンド法

シングルエンド法
の場合

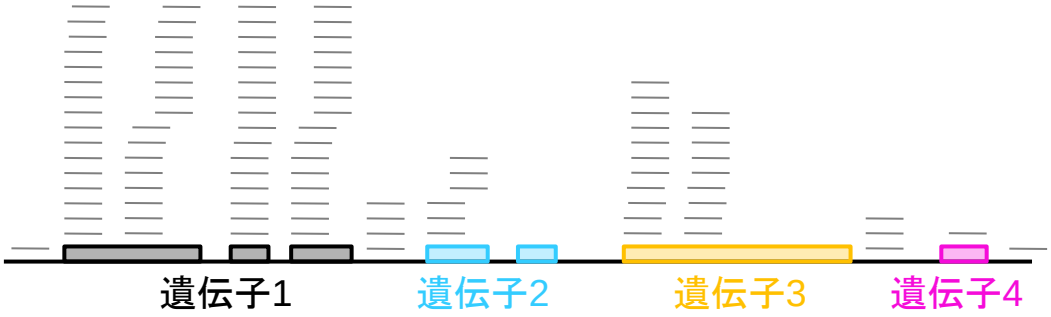


RNA-Seqデータ → 遺伝子発現行列

光刺激前（T1）の目のトランスクリプトーム

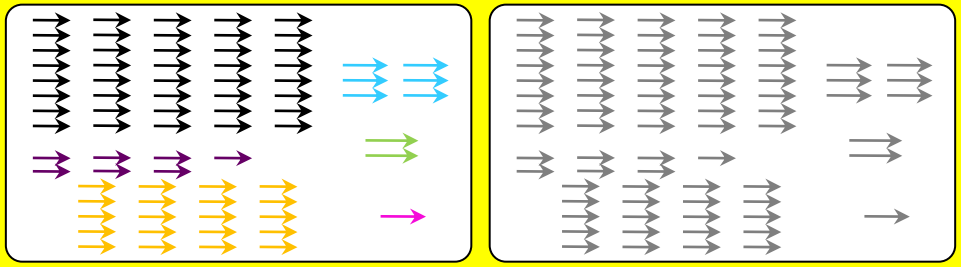


ゲノム配列にマッピング



－イメージ－
50-125塩基程度からなる配列が沢山ある

－実際－
数百万個の配列があり、どの遺伝子に対応するか不明



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
…
…

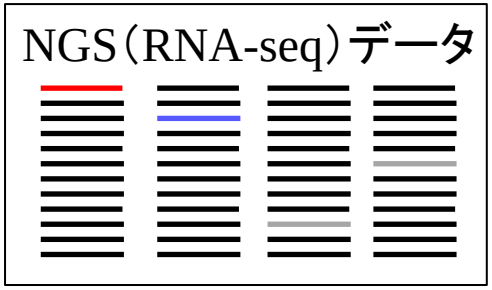
（短い）配列を読んだものという意味で（ショート）リードなどと呼ばれる

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」

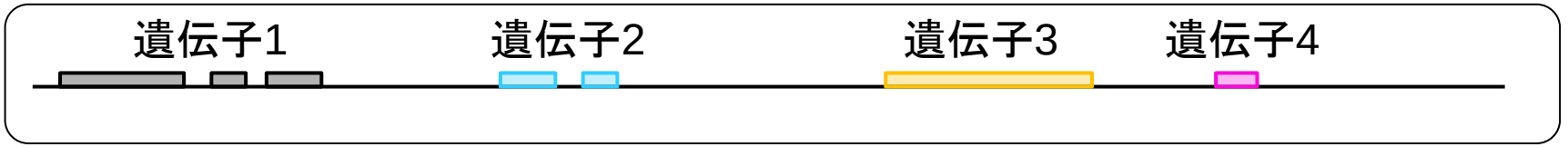
1番染色体　2番染色体　3番染色体 …



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



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

AAAAAAA…

>遺伝子2

AAAAAAA…

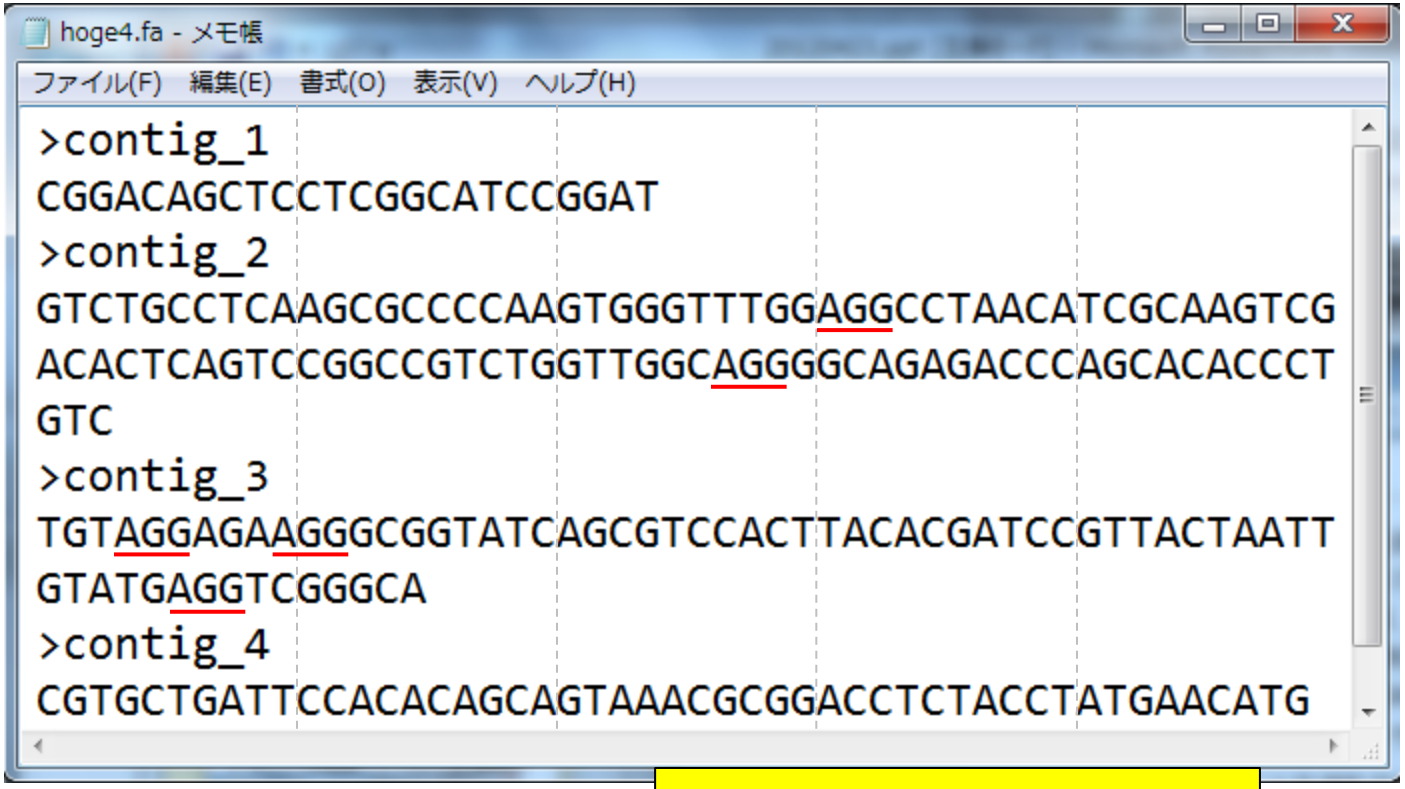
>遺伝子3

AAAAAAA…

…

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”



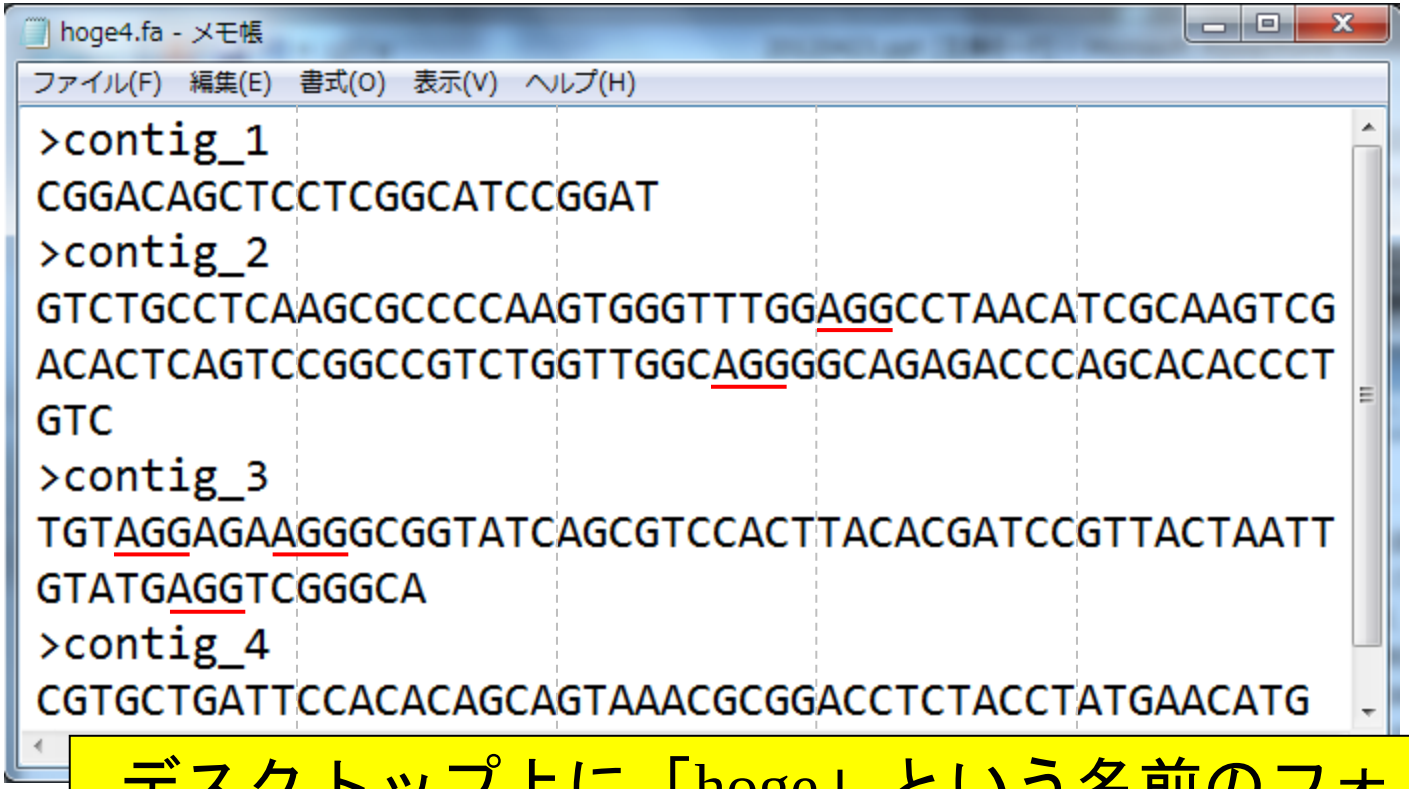
出力ファイル：hoge2.txt

	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



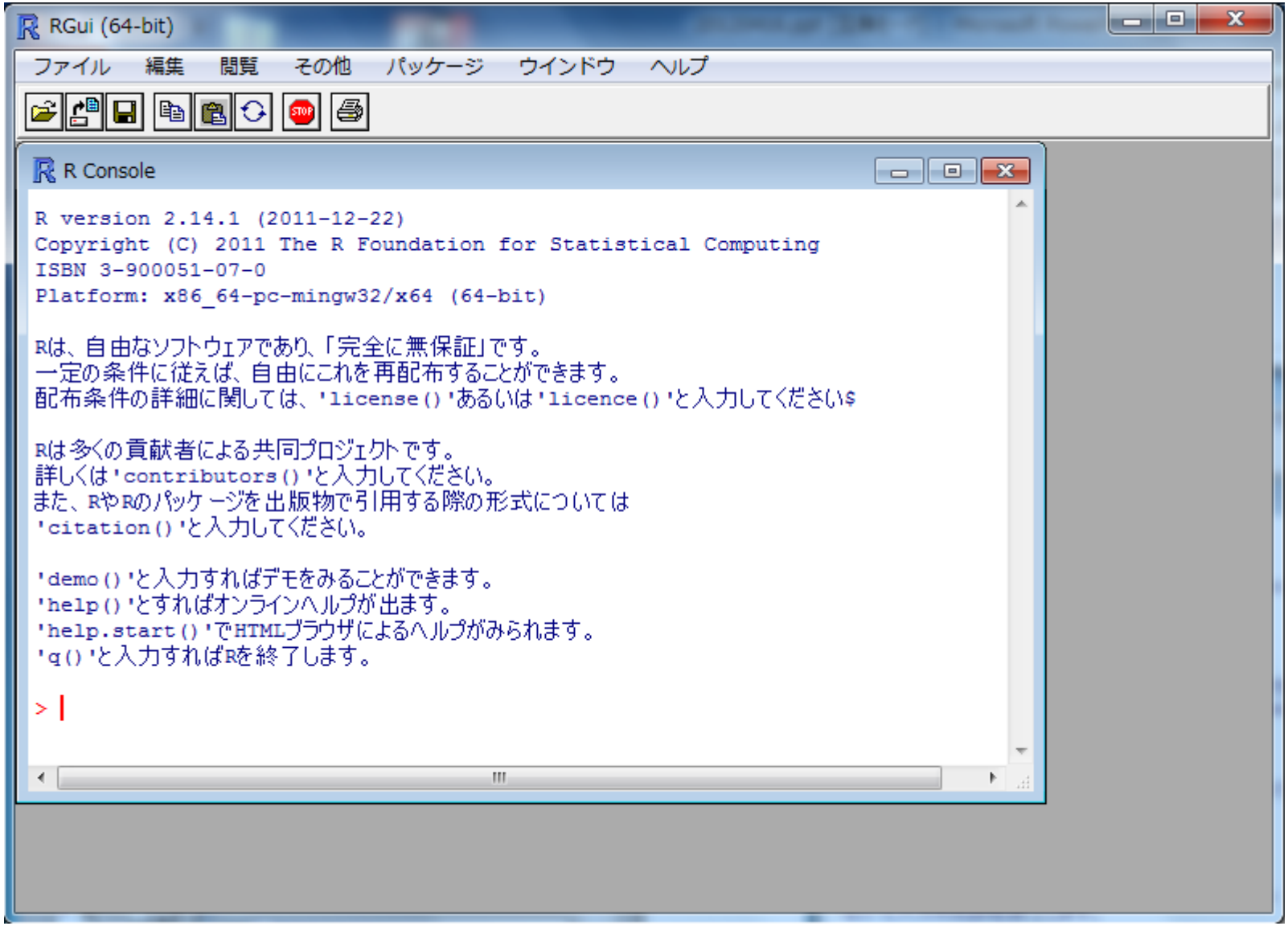
出力ファイル：hoge2.txt

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

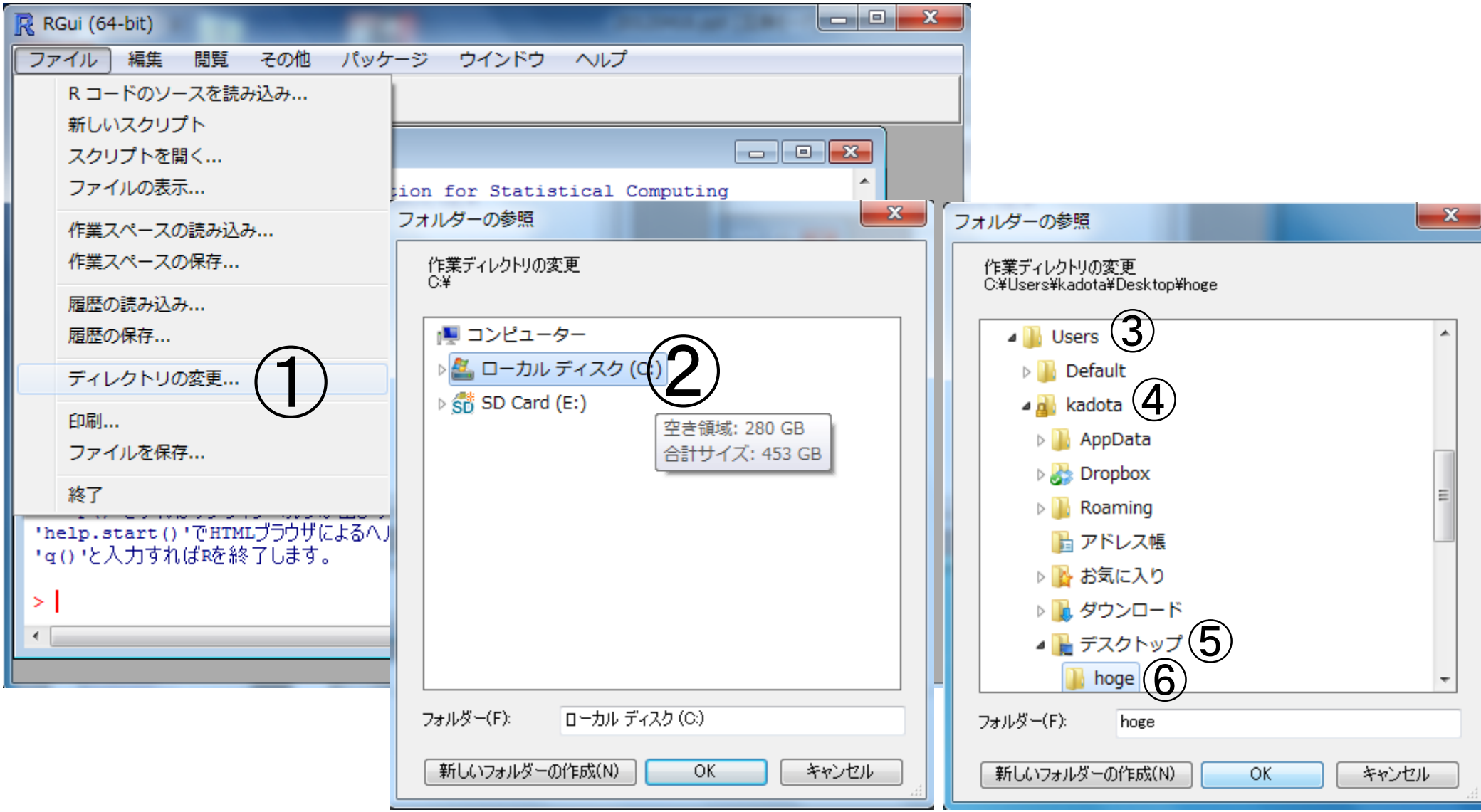
デスクトップ上に「hoge」という名前のフォルダがあり、フォルダ中に入力ファイル（**hoge4.fa**）が存在する、という前提

Rの起動

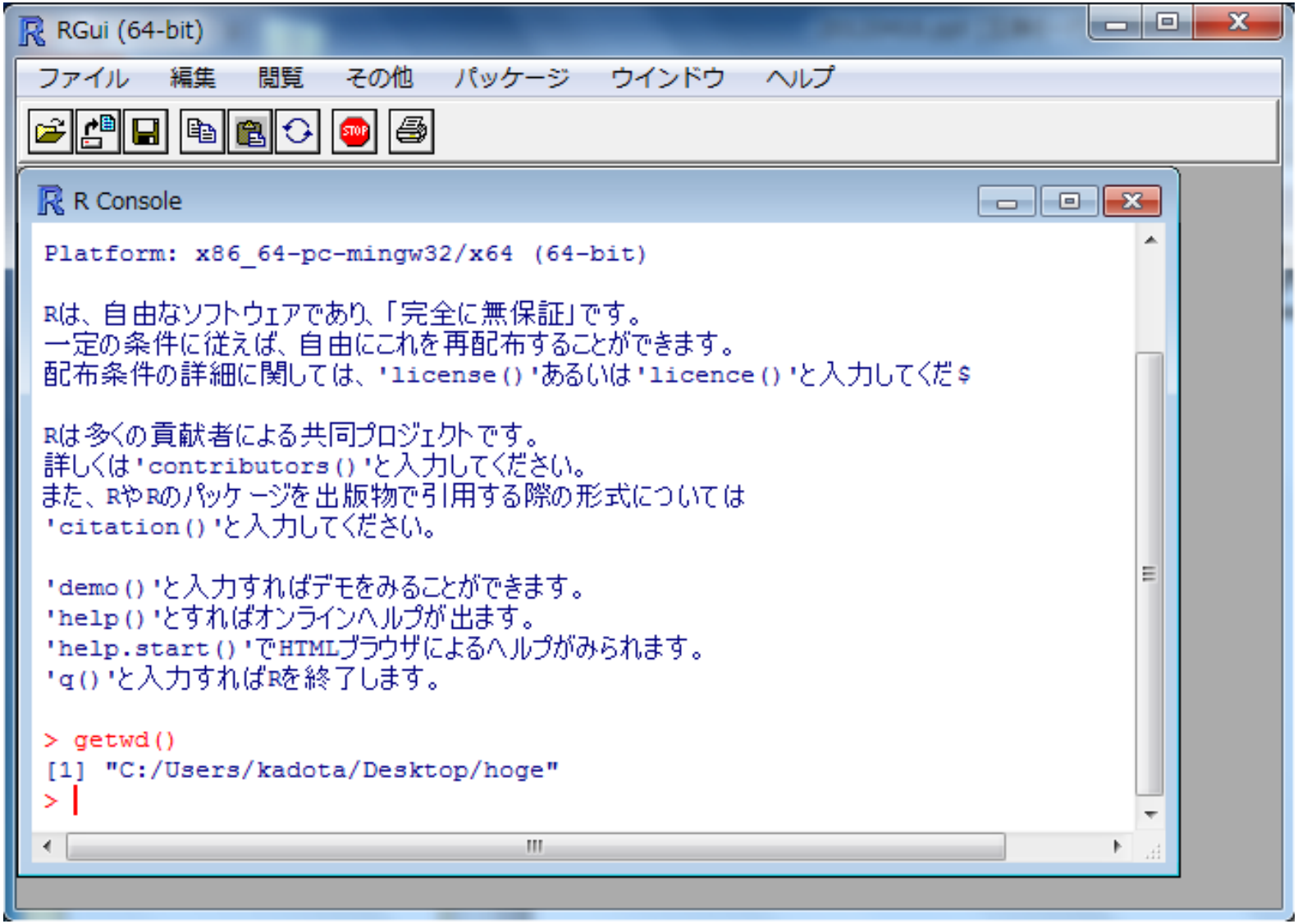

R x64 2.14.1


RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ
R Console
R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)
Rは、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()'あるいは'licence()'と入力してください$
Rは多くの貢献者による共同プロジェクトです。
詳しくは'contributors()'と入力してください。
また、RやRのパッケージを出版物で引用する際の形式については
'citation()'と入力してください。
'demo()'と入力すればデモをみることができます。
'help()'とすればオンラインヘルプが出ます。
'help.start()'でHTMLブラウザによるヘルプがみられます。
'q()'と入力すればRを終了します。
>

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

http://www.iu.a.u-tokyo.ac.jp/~kadota/ 門田　幸二のホームページ

門田　幸二のホームページ

- **名前**
  門田　幸二(かどた　こうじ)
- **所属**
  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ユニット
- **身分**
  特任
- **研究**
  バイ
- **所属**
  日本
  日本
- **研究**

- **リンク集**
  - (Rで)マイクロアレイデータ解析(last modified: 2012.07.17)
  - (Rで)塩基配列解析(主に次世代シーケンサーのデータ)(last modified: 2012.07.20)
  - (マイクロアレイ)データ解析Tips(last modified: 2008.7.28)
  - 東大・院農・応生工・生物情報工学研究室
  - CBRC
  - 放射線医学総合研究所
  - ライフサイエンス統合データベースプロジェク
  - ライフサイエンスQA
  - ライフサイエンス辞書オンラインサービス
  - WINGpro(データベースポータルサイト)
  - 統合TV
  - ゲノム解析ツールリンク集
  - Bioconductor
  - NCBI GEO
  - NCBI SRA
  - EMBOSS
  - BIOWEB(バイオ研究者支援サイト)
  - 新学術領域研究「複合適応形質進化の遺伝子基盤解明」ホームページ

ご意見、ご質問はkadota@iu.a.u-tokyo.ac.jpまで。
Last modified: 2012.07.27

「(Rで)塩基配列解析」のほうをクリック

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

(Rで)塩基配列解析（主に...

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2012/07/20)

What's new?
・RNA-seq周辺のノートPCを使用した集中講義農学
担当ではありませんがPC不使用の講義があります)
い受講希望者は私にコンタクトをとってください。また
て必要なパッケージをインストールしておいてください

・若干項目名を(あまりにも場違いだったものを)変更
NEW
・R2.15.1がリリースされていたのでこれに変更しました
・ここで書いてはいないものの、あちこち追加などして
・htmlのタグが「NAME="#XXX"」のようになっている
受けましたのでその周辺を修正しました。(2012/05/
・TbT論文中の正規化法TbTを実装したRのパッケー
(2012/05/09)

- はじめに (last modified 2012/03/29)
- Rのインストールと起動 (last modified 2012/07/06)
- サンプルデータ (last modified 2012/03/15)
- イントロダクション | NGS | 各種覚書 (last modified
- イントロダクション | NGS | 様々なプラットフォーム
- イントロダクション | NGS | リファレンス配列取得(
- イントロダクション | NGS | リファレンス配列取得後
- イントロダクション | NGS | リファレンス配列取得後
- イントロダクション | NGS | アノテーション情報取得
- イントロダクション | NGS | アノテーション情報取得

- 前処理 | 発現レベルの定量化(RPKM by DEGseq) (last modified 2011/01/06)
- 前処理 | 発現レベルの定量化(RPKM by ERANGE) (last modified 2010/11/26)
- 前処理 | 発現レベルの定量化(FPKM by Cufflinks) (last modified 2010/12/08)
- 前処理 | 発現レベルの定量化(FVKM by NEUMA) (last modified 2010/11/29)
- 前処理 | 発現レベルの定量化(Scripture) (last modified 2010/11/26)
- 前処理 | 発現レベルの定量化(NAC by ALEXA-seq) (last modified 2010/12/08)
- 前処理 | 発現レベルの定量化(Trans-ABySS) (last modified 2010/12/08)
- 前処理 | raw counts --> RPM (Mortazavi_2008) (last modified 2012/06/27)
- 前処理 | raw counts --> RPKM (Mortazavi_2008) (last modified 2012/02/29)
- 前処理 | TbT正規化(Kadota_2012) (last modified 2012/06/27)
- 前処理 | TMM正規化(Robinson_2010) (last modified 2012/06/27)
- 前処理 | DESeqの正規化(Anders_2010) (last modified 2012/06/27)
- 前処理 | 75percentile正規化(第3四分位数を揃える) (last modified 2011/08/09)
- 前処理 | GAM正規化(Zheng_2011) (last modified 2012/06/21)
- 解析 | 一般 | アラインメント(ペアワイズ;基本編1) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペア ズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング modified 2012/04/13)
- 解析 | 一般 | GC含量 (GC conten last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2012/06/27)
- 解析 | 一般 | 上流配列解析 | Local Distribution of Short Sequences (LDSS)解析 (Yamamoto_2007) (last modified 2012/07/17) NEW
- 解析 | 一般 | 上流配列解析 | R
- 解析 | NGS(RNA-seq) | その他
- 解析 | NGS(RNA-seq) | その他
- 解析 | NGS(RNA-seq) | その他
- 解析 | NGS(RNA-seq) | その他 | 負の二項分布に従うシミュレーションデータを作成する(fixed dispersion) (last modified 2011/10/31)
- 解析 | NGS(RNA-seq) | その他 | 負の二項分布に従うシミュレーションデータを作成する(random dispersion) (last modified 2011/10/31)
- 解析 | NGS(RNA-seq) | その他 | 負の二項分布に従うシミュレーションデータを作成する(tagwise dispersion) (last modified 2012/07/05)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | について (last modified 2012/06/14)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | DSS (Wu_201X) (last modified 2012/07/05)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | NOISeq (Tarazona_2011) (last modified 2012/06/21)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | GPseq (Srivastava_2010) (last modified 2012/06/21)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | NBPSeq coupled with TbT normalization (Kadota_2012) (last modified 2012/04/18)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | NBPSeq (Di_2011) (last modified 2012/03/15)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | baySeq (Hardcastle_2010) (last modified 2011/12/20)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | DESeq coupled with TbT normalization (Kadota_2012) (last modified 2012/06/27)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | DESeq (Anders_2010) (last modified 2012/06/27)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Fisher's exact test (FET) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Likelihood ratio test (LRT) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Fold change (FC) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | MARS (Wang 2010) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | edgeR coupled with TbT normalization (Kadota_2012) (last modified 2012/04/18)

「パターンマッチング」のところをクリック

パターンマッチング

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2012/04/13) NEW
- 解析 | 一般 | GC含量 (GC content) (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2012/04/04) NEW
- 解

• 解析 | 一般 | パターンマッチング

読み込んだリファレンス配列(reference sequence or subject sequence)から(短い)配列パターンを探す場合に利用します。
ここでは、以下の5つの例題を行います:
1. D
Finge
2. 1. と同
3. 2. と同
4. multi-
5. multi-
6. multi-
7. multi-
(記述の

4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：

```
------    ここから    ------
in_f <- "hoge4.fa"                                          #読み込みたいFASTA形式のファイル名を指
out_f <- "hoge2.txt"                                        #出力ファイル名を指定してout_fに格納
param <- "AGG"                                              #調べたい配列パターンを指定してparamに

#必要なパッケージをロード
library(Biostrings)                                         #パッケージの読み込み

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

#本番
out <- vmatchPattern(pattern=param, subject=seq)            #paramで指定した配列と100%マッチの領域
hoge <- cbind(start(unlist(out)), end(unlist(out)))         #一致領域の(start, end)の位置情報をhog
colnames(hoge) <- c("start", "end")                         #行列hogeに列名を付加
rownames(hoge) <- names(unlist(out))                        #行列hogeに行名を付加
tmp <- cbind(rownames(hoge), hoge)                          #ファイルに出力したい情報を連結してtmp
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名

------    ここまで    ------
```

基本はコピペ

```
4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：
------    ここから    ------
in_f <- "hoge4.fa"                    #読み込みたいFASTA形式のファイル名を指定してin_fに格納
out_f <- "hoge2.txt"                  #出力ファイル名を指定してout_fに格納
param <- "AGG"                        #調べたい配列パターンを指定してparamに格納

#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

#本番
out <- vmatchPattern(pat
hoge <- cbind(start(unli
colnames(hoge) <- c("sta
rownames(hoge) <- names(
tmp <- cbind(rownames(ho
write.table(tmp, out_f,

------    ここまで    --
```

切り取り(T)
コピー(C)
貼り付け
すべて選択(A)
印刷(I)...
印刷プレビュー(N)...
Bing でマップ
Bing で翻訳
Google で検索
電子メール (Windows
すべてのアクセラレー

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ

R Console

```
> getwd()
[1] "C:/Users/kadota_DELL/Desktop/hoge"
>
```

コピー　Ctrl+C
ペースト　Ctrl+V
コマンドのみペースト
コピー&ペースト　Ctrl+X
ウインドウの消去　Ctrl+L
全て選択
バッファに出力　Ctrl+W
ウィンドウを常にトップに置く

①一連のコマンド群をコピーして

②R Console画面上でペースト

実行結果

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ

R Console

```
> library(Biostrings)                                          #パッケ
>
> #入力ファイルの読み込み
> seq <- read.DNAStringSet(in_f, format="fasta")               #in_f
>
> #本番
> out <- vmatchPattern(pattern=param, subject=seq)             #para
> hoge <- cbind(start(unlist(out)), end(unlist(out)))          #一致
> colnames(hoge) <- c("start", "end")                          #行列
> rownames(hoge) <- names(unlist(out))                         #行列
> tmp <- cbind(rownames(hoge), hoge)                           #ファイ
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの
>
>
```

hoge2.txt - M.

ファイル　ホー　挿入　ペー　数式　デー

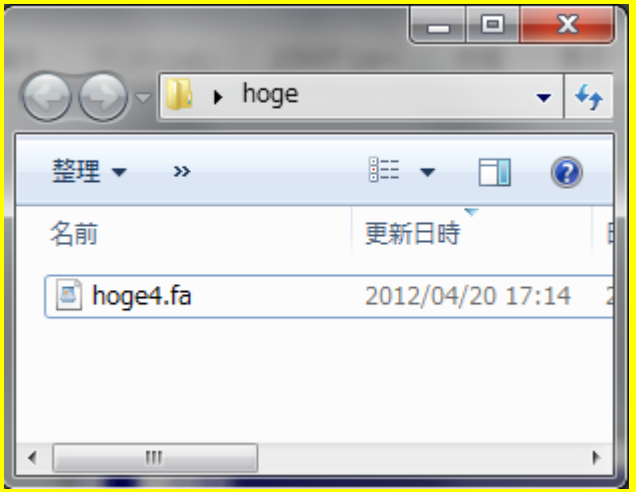
E12

	A	B	C	D
1		start	end	
2	contig_2	31	33	
3	contig_2	77	79	
4	contig_3	4	6	
5	contig_3	10	12	
6	contig_3	56	58	
7				

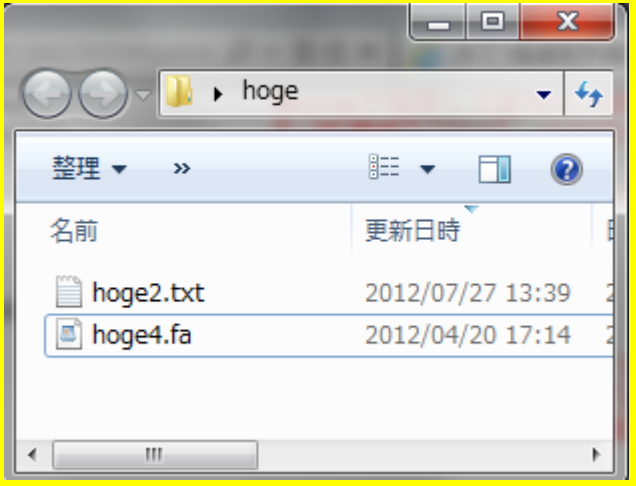
hoge2

コマンド　100%

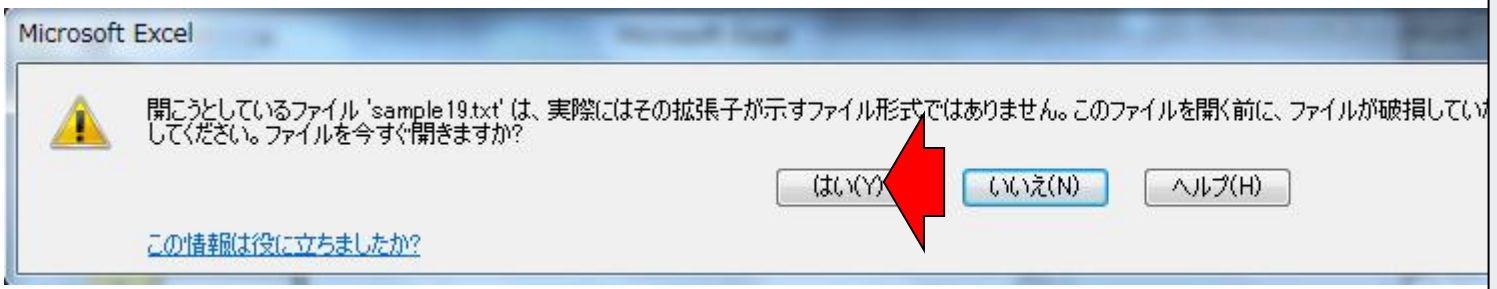
実行前のhogeフォルダ

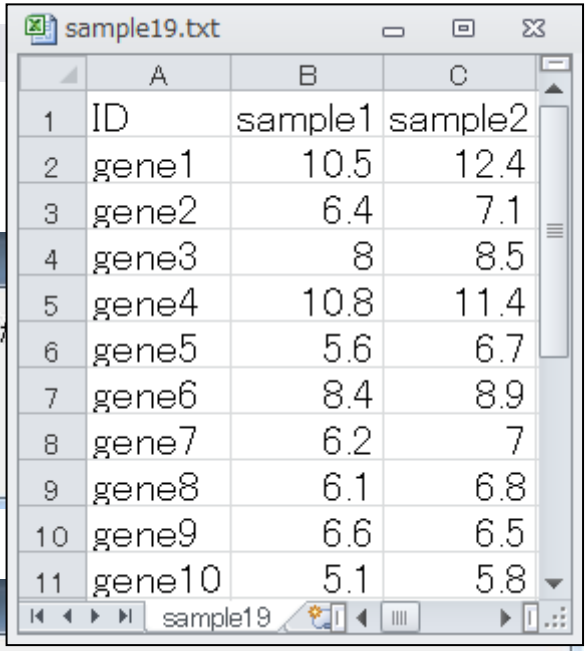


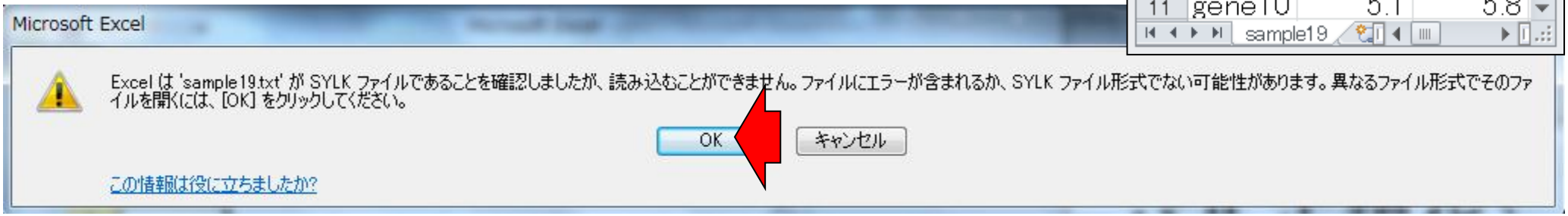
実行後のhogeフォルダ



エクセルで開くとき…

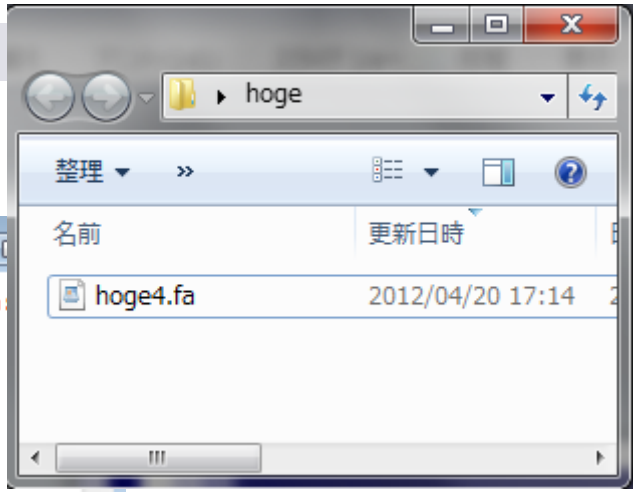






（ドラッグ＆ドロップで開こうとすると）エラーが出て一回目は開けないことがあるが、その後もう一度同じ作業を繰り返すと開けます…

ありがちなミス1

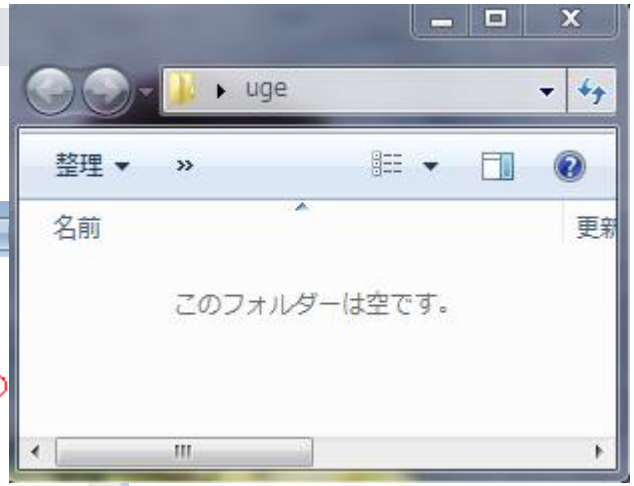


R Console

```
> in_f <- "hoge4.fa"                                          #読み込みたいFASTA形式の
> out_f <- "hoge2.txt"                                        #出力ファイル名を指定し$
> param <- "AGG"                                              #調べたい配列パターンを$
>
> #必要なパッケージをロード
> library(Biostrings)                                         #パッケージの読み込み
 要求されたパッケージ BiocGenerics をロード中です

 次のパッケージを付け加えます: ''BiocGenerics''

The following object(s) are masked from 'package:stats':

    xtabs

The following object(s) are masked from 'package:base':

    anyDuplicated, cbind, colnames, duplicated, eval, Filter, Find, get,
    intersect, lapply, Map, mapply, mget, order, paste, pmax, pmax.int, pmin,
    pmin.int, Position, rbind, Reduce, rep.int, r
    tapply, union, unique

 要求されたパッケージ IRanges をロード中です
>
> #入力ファイルの読み込み
> seq <- read.DNAStringSet(in_f, format="fasta")              #in_fで指定したファイル$
 以下にエラー .Call2("new_input_ExternalFilePtr", fp, PACKAGE = "Biostrings") :
  cannot open file 'hoge4.fa'
>
> #本番
> out <- vmatchPattern(pattern=param, subject=seq)            #paramで指定した配列と10$
 以下にエラー function (classes, fdef, mtable)  :
  unable to find an inherited method for function "vmatchPattern", for signature "function"
```

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



R Console

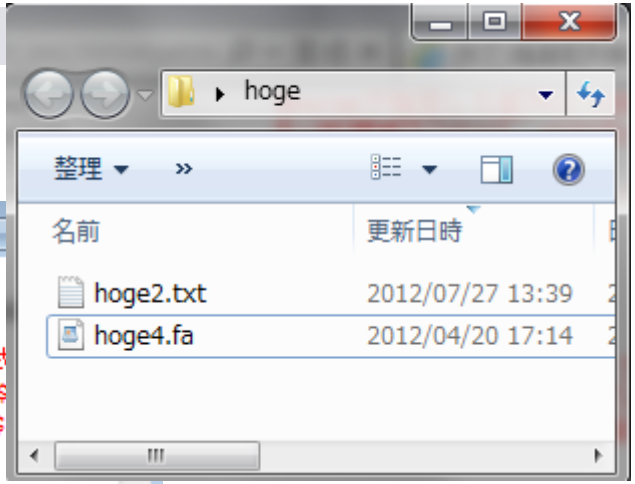
```
>
> getwd()
[1] "C:/Users/kadota_DELL/Desktop/uge"
> in_f <- "hoge4.fa"                                    #読み込みたいFASTA形式の
> out_f <- "hoge2.txt"                                  #出力ファイル名を指定し$
> param <- "AGG"                                        #調べたい配列パターンを$
>
> #必要なパッケージをロード
> library(Biostrings)                                   #パッケージの読み込み
>
> #入力ファイルの読み込み
> seq <- read.DNAStringSet(in_f, format="fasta")        #in_fで指定したファイル$
 以下にエラー .Call2("new_input_ExternalFilePtr", fp, PACKAGE = "Biostrings") :
  cannot open file 'hoge4.fa'
>
> #本番
> out <- vmatchPattern(pattern=param, subject=seq)      #paramで指定した配列と10$
 以下にエラー function (classes, fdef, mtable)  :
  unable to find an inherited method for function "vmatchPattern", for signature "function"
> hoge <- cbind(start(unlist(out)), end(unlist(out)))   #一致領域の(start, end)$
 以下にエラー start(unlist(out)) :
   引数 'x' の評価中にエラーが起きました (関数 'start' に対するメソッドの選択時):  以下にエ$
   引数 'x' の評価中にエラーが起きました (関数 'unlist' に対するメソッドの選択時):  エラー$


> colnames(hoge) <- c("start", "end")                   #行列hogeに列名を付加
 以下にエラー colnames(hoge) <- c("start", "end") :
   オブジェクト 'hoge' がありません
> rowname
 以下にエラ
   引数 '
```

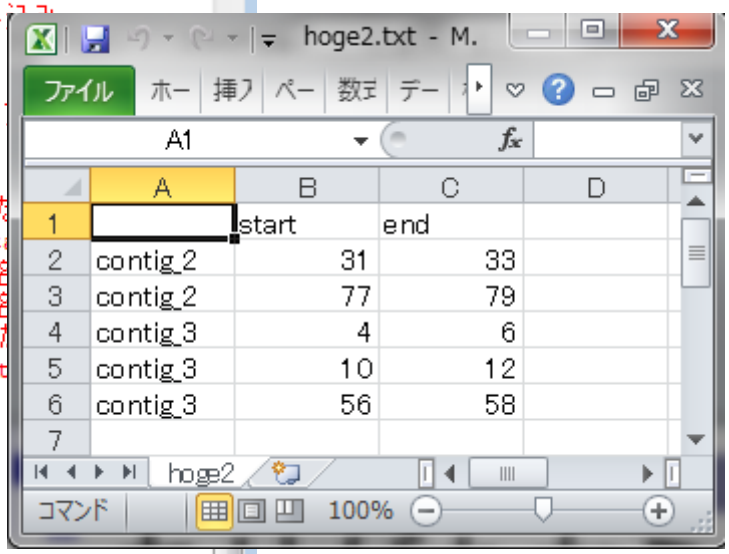
必要な入力ファイルが作業ディレクトリ中に存在しない…

ありがちなミス3

R Console

```
> getwd()
[1] "C:/Users/kadota_DELL/Desktop/hoge"
> in_f <- "hoge4.fa"                                    #読み込みたいFASTA形式
> out_f <- "hoge2.txt"                                  #出力ファイル名を指定し
> param <- "AGG"                                        #調べたい配列パターンを
>
> #必要なパッケージをロード
> library(Biostrings)                                   #パッケージの読み込
>
> #入力ファイルの読み込み
> seq <- read.DNAStringSet(in_f, format="fasta")        #in_fで指定した
>
> #本番
> out <- vmatchPattern(pattern=param, subject=seq)      #paramで指定し
> hoge <- cbind(start(unlist(out)), end(unlist(out)))   #一致領域の(st
> colnames(hoge) <- c("start", "end")                   #行列hogeに列名
> rownames(hoge) <- names(unlist(out))                  #行列hogeに行名
> tmp <- cbind(rownames(hoge), hoge)                    #ファイルに出力し
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身をout
 以下にエラー file(file, ifelse(append, "a", "w")) :
   コネクションを開くことができません
 追加情報:    警告メッセージ:
In file(file, ifelse(append, "a", "w")) :
   ファイル 'hoge2.txt' を開くことができません: Permission denied
>
>
```



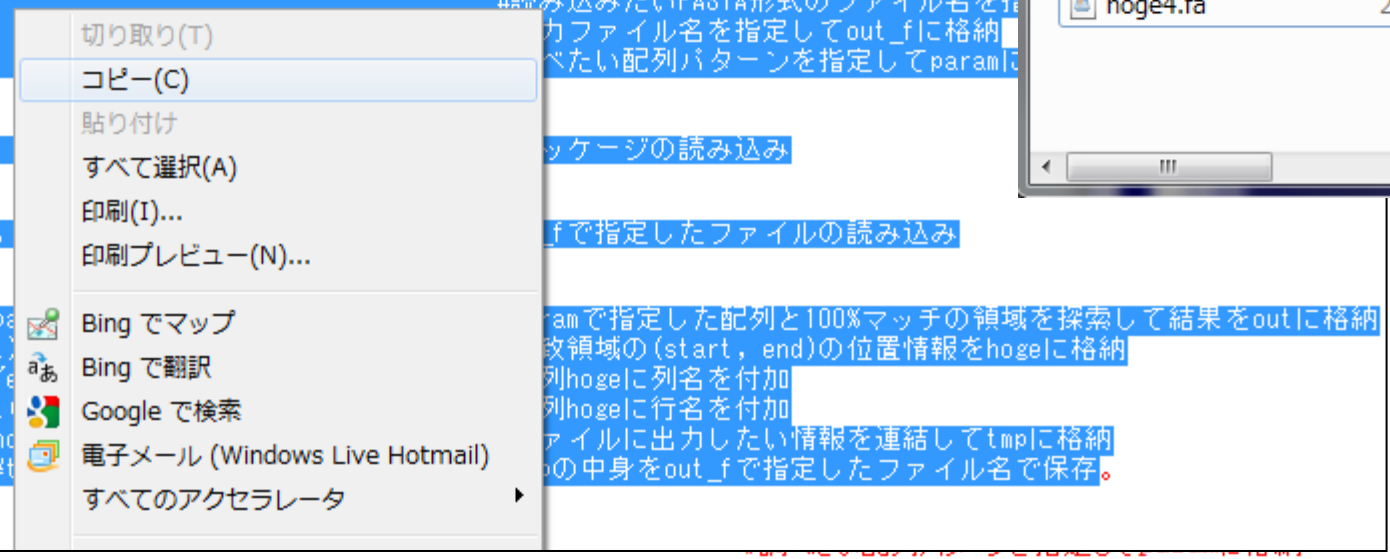


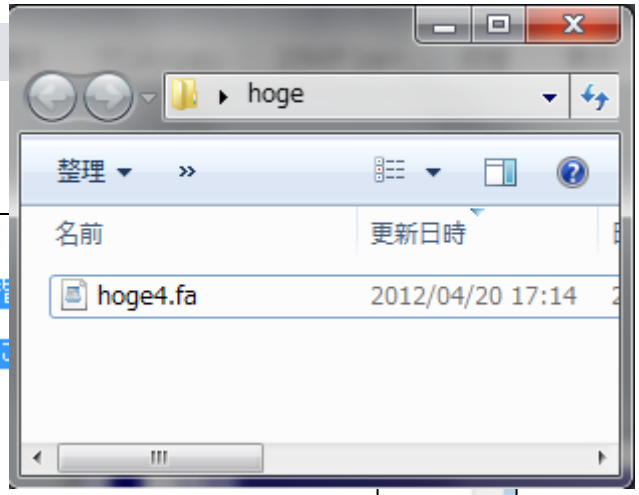
出力予定のファイル名と同じものを別のプログラムで開いているため最後のwrite.table関数のところでエラーが出る

ありがちなミス4

4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：

```
------    ここから    ------
in_f <- "hoge4.fa"                                    #読み込みたいFASTA形式のファイル名を指
out_f <- "hoge2.txt"                                  カファイル名を指定してout_fに格納
param <- "AGG"                                        べたい配列パターンを指定してparamに

#必要なパッケージをロード
library(Biostrings)                                   ッケージの読み込み

#入力ファイルの読み込み
seq <- read.DNAStringSet(in_f,                        fで指定したファイルの読み込み

#本番
out <- vmatchPattern(pattern=pa                       amで指定した配列と100%マッチの領域を探索して結果をoutに格納
hoge <- cbind(start(unlist(out)                       数領域の(start, end)の位置情報をhogeに格納
colnames(hoge) <- c("start", "e                       列hogeに列名を付加
rownames(hoge) <- names(unlist(                       列hogeに行名を付加
tmp <- cbind(rownames(hoge), ho                       ァイルに出力したい情報を連結してtmpに格納
write.table(tmp, out_f, sep="¥t                       の中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```





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）

```
$ロード
$("start", "end")                                     #行列hogeに列名を付加
$ames(unlist(out))                                    #行列hogeに行名を付加
$es(hoge), hoge)                                      #ファイルに出力したい情報を連結してtmpに格納
$t_f, sep="\t", append=F, quote=F, row.names=F)       #tmpの中身をout_fで指定したファイル名で保存
```

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2012/04/13) NEW
- 解析 | 一般 | GC含量 (GC contents) (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2012/04/04) NEW

- 解析 | 一般 | パターンマッチング

解析 | 一般 | パターンマッチング

読み込んだリファレンス配列(reference sequence or subject sequence)から(短い)配列パターンを探す場合に利用します。
ここでは、以下の6つの例題を行います:

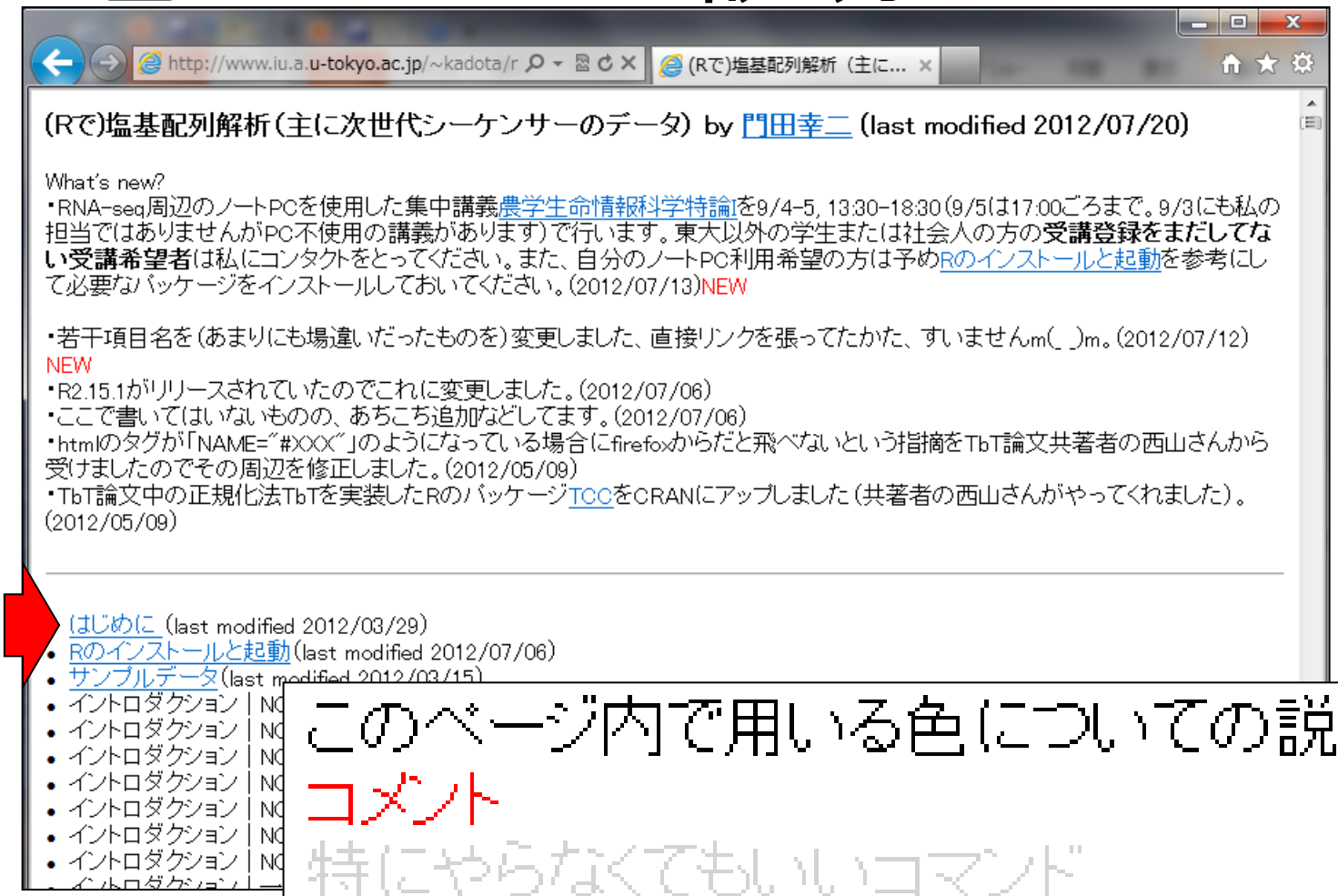
1. Dihyd
Finger N
2. 1. と同
3. 2. と同
4. multi-
5. multi-
6. multi-
7. multi-
(記述の

4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：

```
------    ここから    ------
in_f <- "hoge4.fa"                                  #読み込みたいFASTA形式のファイル名を指
out_f <- "hoge2.txt"                                #出力ファイル名を指定してout_fに格納
param <- "AGG"                                      #調べたい配列パターンを指定してparamに

#必要なパッケージをロード
library(Biostrings)                                 #パッケージの読み込み

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

#本番
out <- vmatchPattern(pattern=param, subject=seq)    #paramで指定した配列と100%マッチの領域
hoge <- cbind(start(unlist(out)), end(unlist(out)))  #一致領域の(start, end)の位置情報をhog
colnames(hoge) <- c("start", "end")                 #行列hogeに列名を付加
rownames(hoge) <- names(unlist(out))                #行列hogeに行名を付加
tmp <- cbind(rownames(hoge), hoge)                  #ファイルに出力したい情報を連結してtmp
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名

------    ここまで    ------
```

「----- 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

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2012/04/13) NEW
- 解析 | 一般 | GC含量 (GC contents) (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2012/04/04) NEW

- 解析 | 一般 | パターンマッチング

• 解析 | 一般 | パターンマッチング

読み込んだリファレンス配列(reference sequence or subject sequence)から(短い)配列パターンを探す場合に利用します。
ここでは、以下の6つの例題を行います:

1. Dihyd
Finger N
2. 1. と同
3. 2. と同
4. multi-
5. multi-
6. multi-
7. multi-
(記述の

4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：

```
------    ここから    ------
in_f <- "hoge4.fa"                                  #読み込みたいFASTA形式のファイル名を指
out_f <- "hoge2.txt"                                #出力ファイル名を指定してout_fに格納
param <- "AGG"                                      #調べたい配列パターンを指定してparamに

#必要なパッケージをロード
library(Biostrings)                                 #パッケージの読み込み

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

#本番
out <- vmatchPattern(pattern=param, subject=seq)    #paramで指定した配列と100%マッチの領域
hoge <- cbind(start(unlist(out)), end(unlist(out)))  #一致領域の(start, end)の位置情報をhog
colnames(hoge) <- c("start", "end")                 #行列hogeに列名を付加
rownames(hoge) <- names(unlist(out))                #行列hogeに行名を付加
tmp <- cbind(rownames(hoge), hoge)                  #ファイルに出力したい情報を連結してtmp
                                                    イル名
```

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？

4. multi-fastaファイルhoge4.faを入力として、"AGG"でキーワード探索を行う場合：

------　ここから　------

```
in_f <- "hoge4.fa"                                  #読み込みたいFASTA形式のファイル名を指定してin_fに格納
out_f <- "hoge2.txt"                                #出力ファイル名を指定してout_fに格納
param <- "AGG"

#必要なパッケージをロード
library(Biostrings)
```

切り取り(T)
コピー(C)
貼り付け
すべて選択(A)

①テンプレートのスクリプトをコピーして

無題 - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
in_f <- "hoge4.fa"                                  #読み込みたいFASTA形式のファイル名を指定してin_fに格納
out_f <- "hoge2.txt"                                #出力ファイル名を指定してout_fに格納
param <- "AGG"                                      #調べたい配列パターンを指定してparamに格納

#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_f, format="fasta")      #in_fで指定したファイルの読み込み

#本番
out <- vmatchPattern(pattern=param, subject=seq)    #paramで指定した配列と100%マッチの領域を探索して結果をoutに
```

②メモ帳などのテキストエディタにペーストして

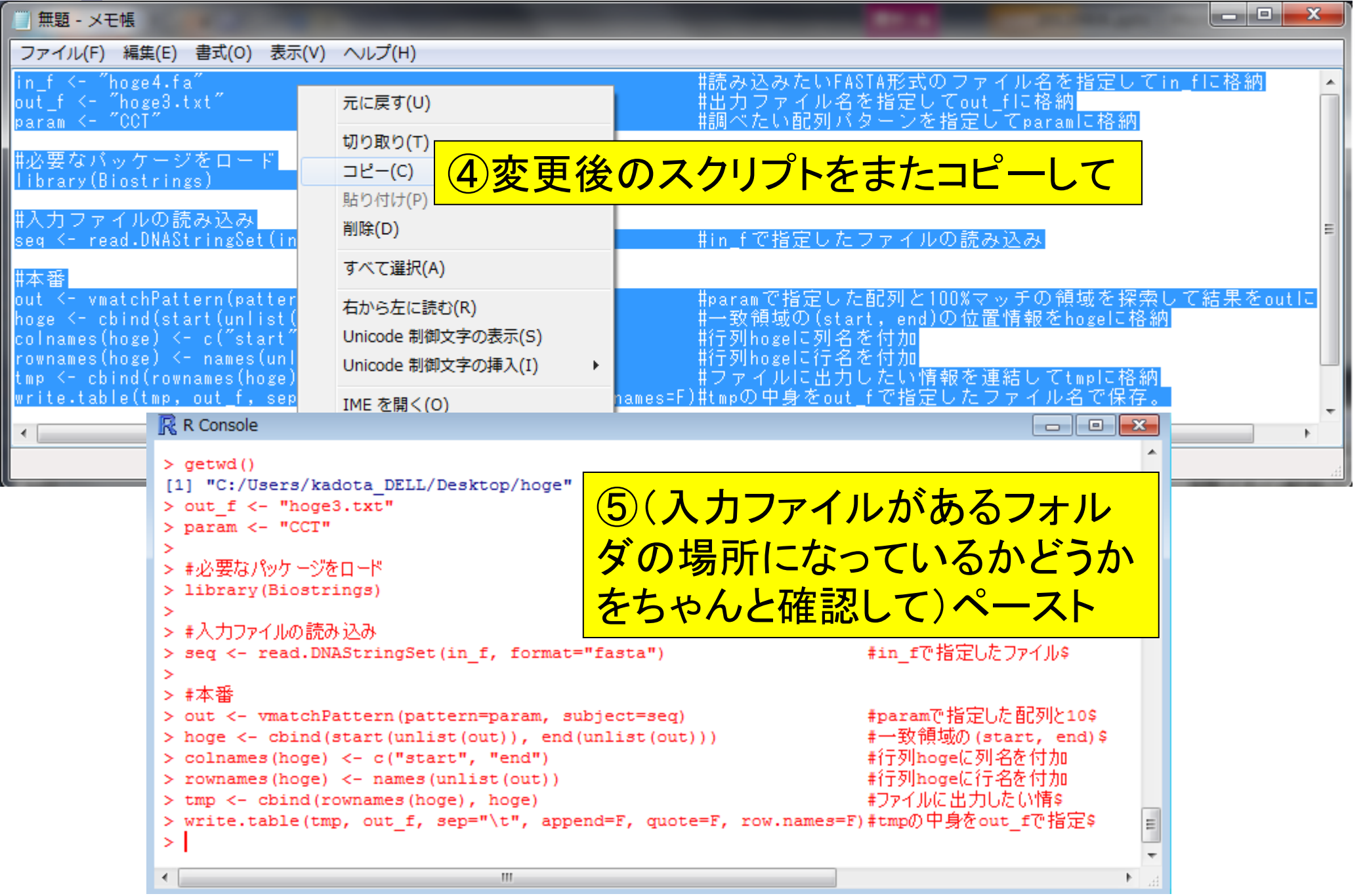
無題 - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
in_f <- "hoge4.fa"                                  #読み込みたいFASTA形式のファイル名を指定してin_fに格納
out_f <- "hoge3.txt"                                #出力ファイル名を指定してout_fに格納
param <- "CCT"                                      #調べたい配列パターンを指定してparamに格納

#必要なパッケージをロード
library(Biostrings)                                 #パッケージの読み込み

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

#本番
out <- vmatchPattern(pattern=param, s               配列と100%マッチの領域を探索して結果をoutに
hoge <- cbind(start(unlist(out)), enc               t, end)の位置情報をhogeに格納
colnames(hoge) <- c("start", "end")                 に付加
rownames(hoge) <- names(unlist(out))                #行列hogeに行名を付加
tmp <- cbind(rownames(hoge), hoge)                  #ファイルに出力したい情報を連結してtmpに格納
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。
```

③必要な箇所を変更して

2 行、16 列


無題 - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
in_f <- "hoge4.fa" #読み込みたいFASTA形式のファイル名を指定してin_fに格納
out_f <- "hoge3.txt" #出力ファイル名を指定してout_fに格納
param <- "CCT" #調べたい配列パターンを指定してparamに格納
#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 #in_fで指定したファイルの読み込み
#本番
out <- vmatchPattern(patter #paramで指定した配列と100%マッチの領域を探索して結果をoutに
hoge <- cbind(start(unlist(#一致領域の(start, end)の位置情報をhogeに格納
colnames(hoge) <- c("start" #行列hogeに列名を付加
rownames(hoge) <- names(unl #行列hogeに行名を付加
tmp <- cbind(rownames(hoge) #ファイルに出力したい情報を連結してtmpに格納
write.table(tmp, out_f, sep names=F)#tmpの中身をout_fで指定したファイル名で保存。
元に戻す(U)
切り取り(T)
コピー(C)
貼り付け(P)
削除(D)
すべて選択(A)
右から左に読む(R)
Unicode 制御文字の表示(S)
Unicode 制御文字の挿入(I)
IME を開く(O)
④変更後のスクリプトをまたコピーして
R Console
> getwd()
[1] "C:/Users/kadota_DELL/Desktop/hoge"
> out_f <- "hoge3.txt"
> param <- "CCT"
>
> #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_f, format="fasta") #in_fで指定したファイル$
>
> #本番
> out <- vmatchPattern(pattern=param, subject=seq) #paramで指定した配列と10$
> hoge <- cbind(start(unlist(out)), end(unlist(out))) #一致領域の(start, end)$
> colnames(hoge) <- c("start", "end") #行列hogeに列名を付加
> rownames(hoge) <- names(unlist(out)) #行列hogeに行名を付加
> tmp <- cbind(rownames(hoge), hoge) #ファイルに出力したい情$
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定$
>
⑤(入力ファイルがあるフォルダの場所になっているかどうかをちゃんと確認して)ペースト

実行結果

hoge3.txt M...

ファイル ホー 挿入 ペー 数式 デー

A1

	A	B	C	D
1		start	end	
2	contig_1	10	12	
3	contig_2	6	8	
4	contig_2	34	36	
5	contig_2	98	100	
6	contig_4	32	34	
7	contig_4	38	40	
8				

hoge3

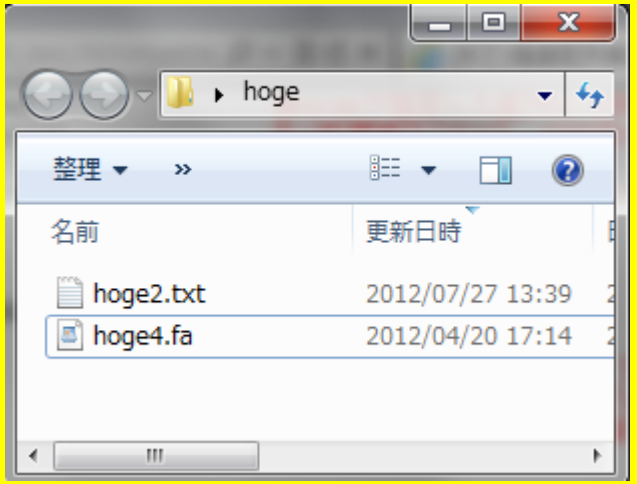
コマンド 100%

hoge4.fa - メモ帳

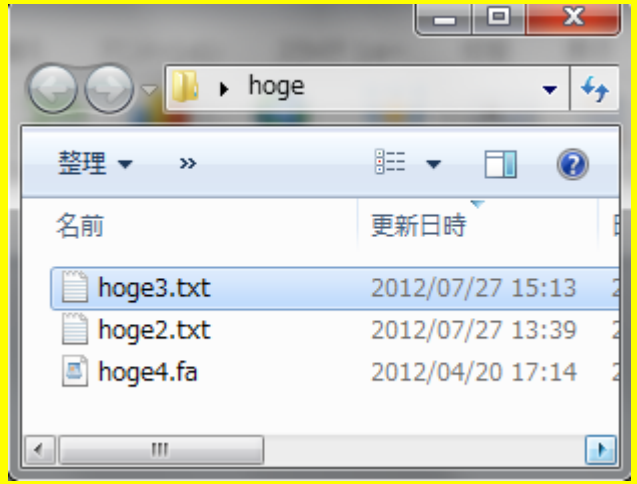
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

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

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

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2012/04/13) NEW
- 解析 | 一般 | GC含量 (GC content) (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2012/04/04) NEW
- 解

data_reads.txt

```
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA
```

• 解析 | 一般 | パターンマッチング

読み込んだリファレンス配列(reference sequence or subject sequence)から(短い)配 …利用します
ここで …ドからなるdata_reads.txtで
1. Dihy
Finger
2. 1. と
3. 2. と
4. mul
5. mul
6. mul
7. mul
(記述

6. multi-fastaファイルhoge4.faをリファレンス配列（マップされ …ドからなるdata_reads.txtで

```
------    ここから    ------
in_f1 <- "hoge4.fa"                                  #読み込みたいFASTA形式ファイルを指定して
in_f2 <- "data_reads.txt"                            #読み込みたいFASTA形式ファイルを指定して
out_f <- "hoge4.txt"                                 #出力ファイル名を指定してout_fに格納

#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1
    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hoge2
    hoge3 <- rep(as.character(reads[i]), length(hoge2))  #hoge2の要素数分だけ、マップする側の配列
    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    ------
```

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

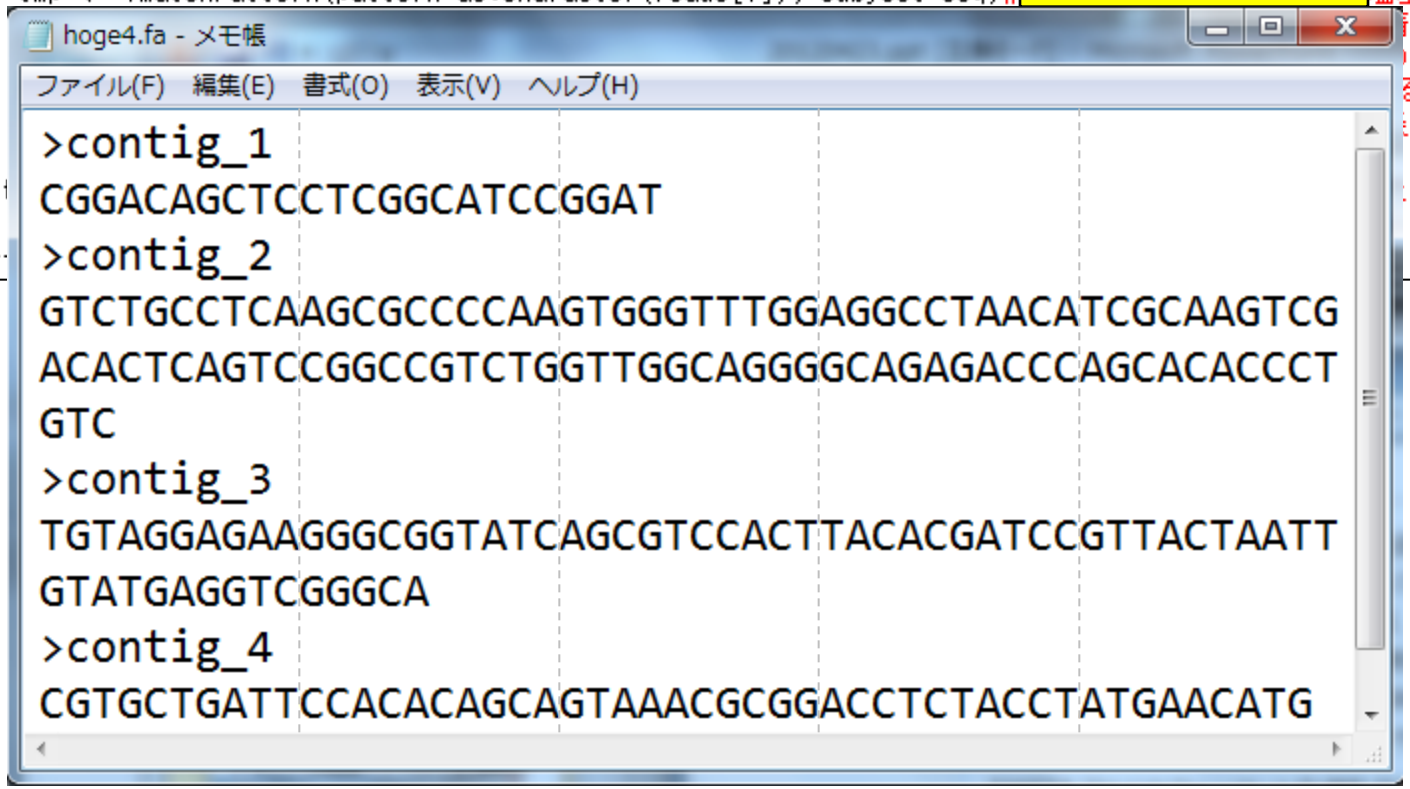
```
6. multi-fastaファイルhoge4.faをリファレンス配列（マップされる側）として、4リードからなるdata_reads.txtで
------    ここから    ------
in_f1 <- "hoge4.fa"                                    #読み
in_f2 <- "data_reads.txt"                              #読み
out_f <- "hoge4.txt"                                   #出力

#必要なパッケージをロード
library(Biostrings)                                    #パッ

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

#本番
out <- c("in_f2", "in_f1", "start", "end")             #最終
for(i in 1:length(reads)){                             #リー
    tmp <- vmatchPattern(pattern=as.character(reads[i]), subject=seq)#
}
writ
```

data_reads.txt

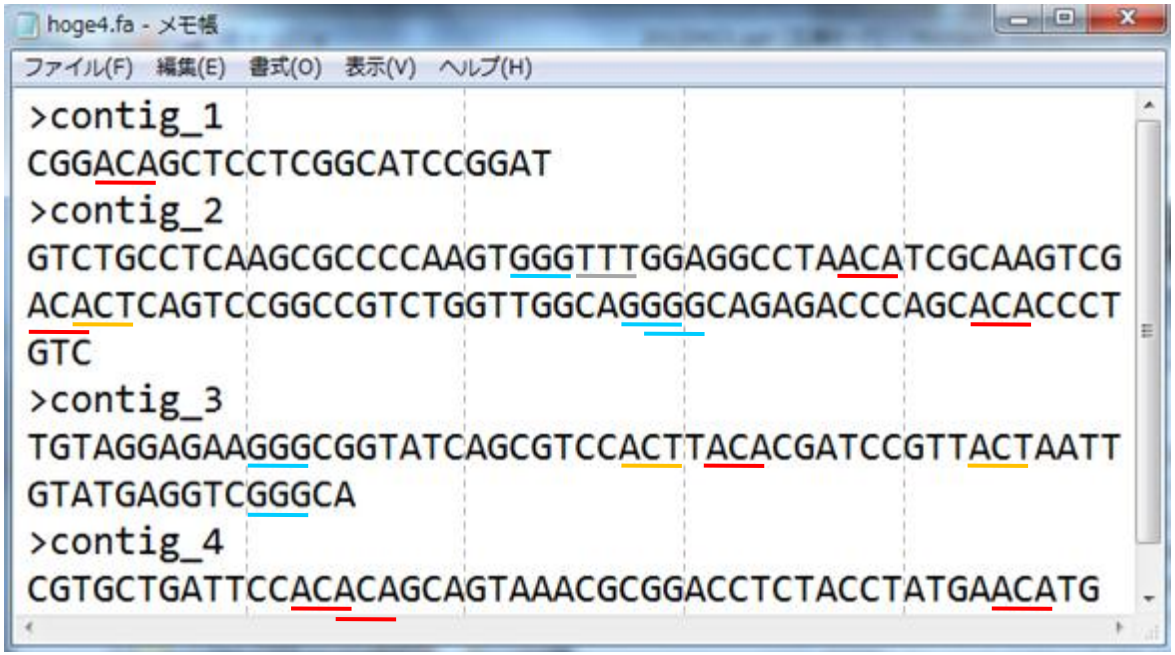
```
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA
```

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



出力結果ファイルと発現量の関係

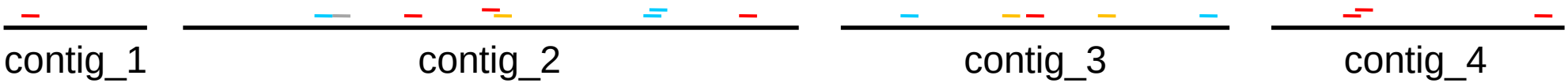


data_reads.txt

```
>seq1
TTT
>seq2
GGG
>seq3
ACT
>seq4
ACA
```

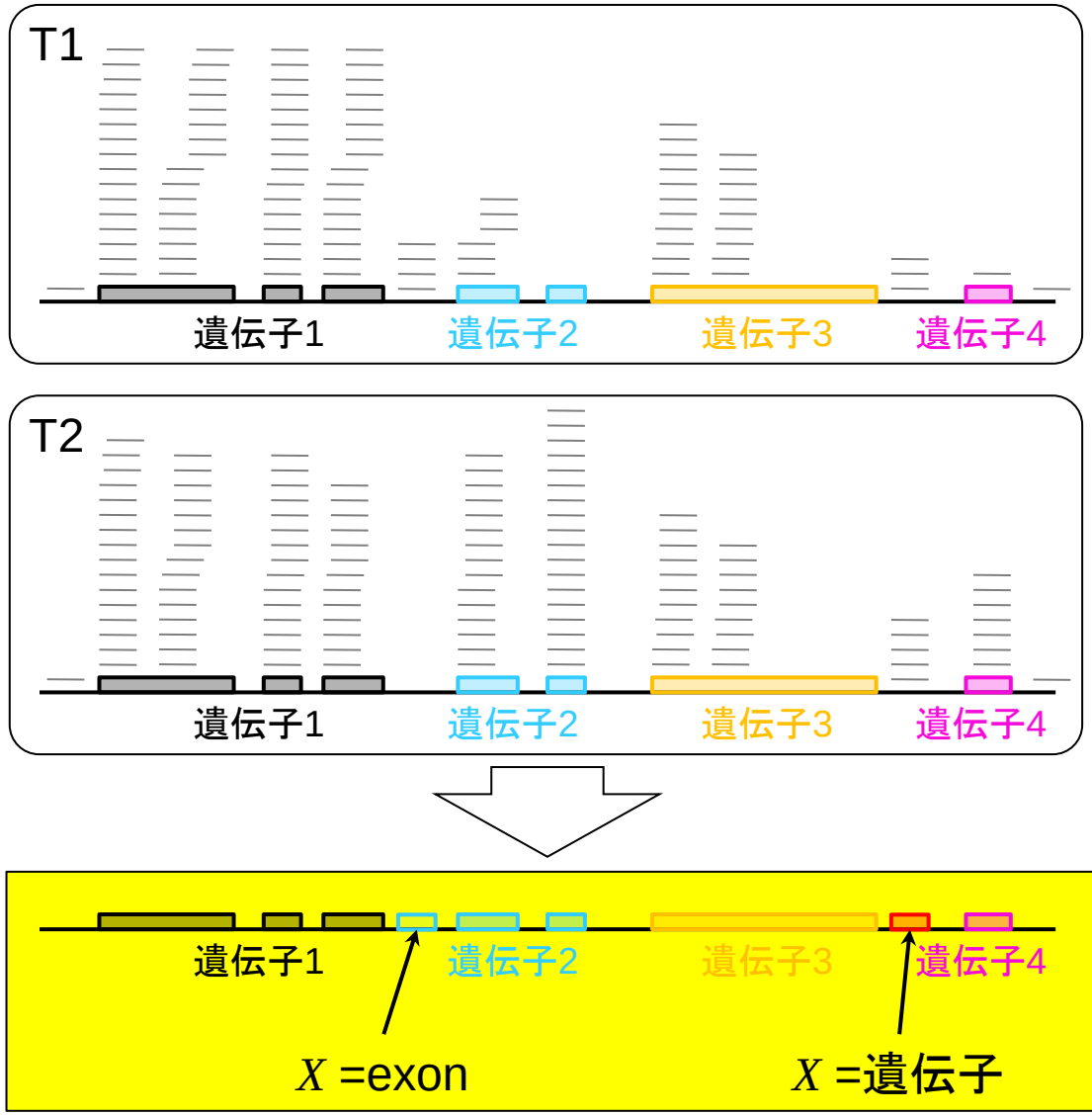
出力ファイル：hoge4.txt

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



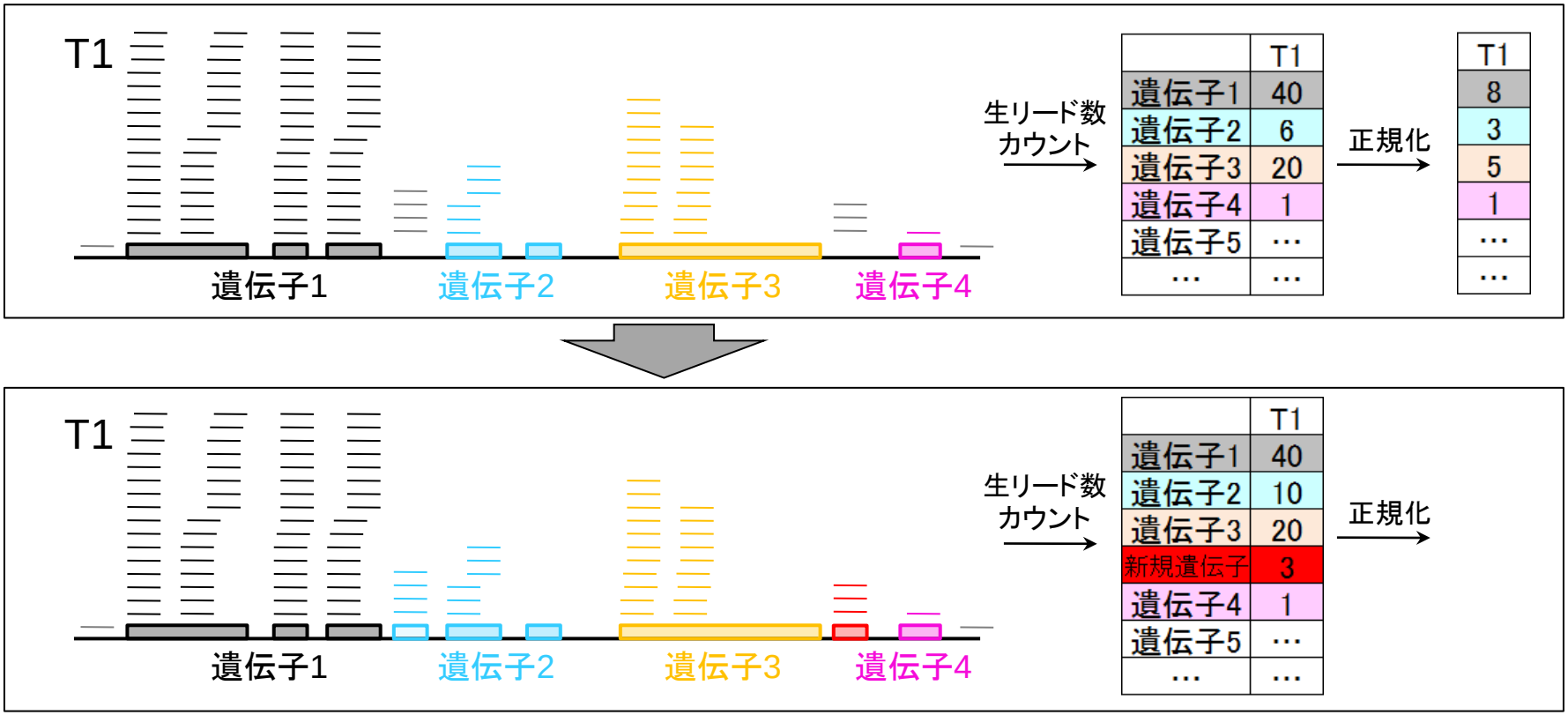
RNA-Seqの長所

- 新規Xの同定
  - X =exon, 遺伝子, …



RNA-Seqの長所

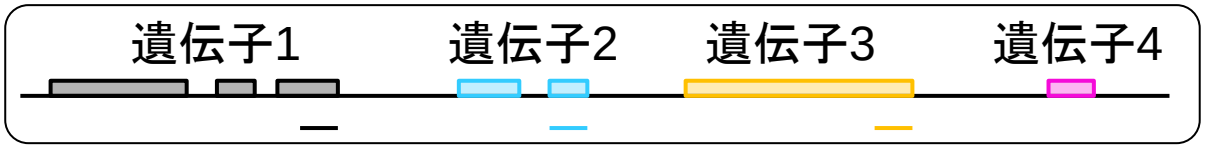
- 新規Xの同定



・“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

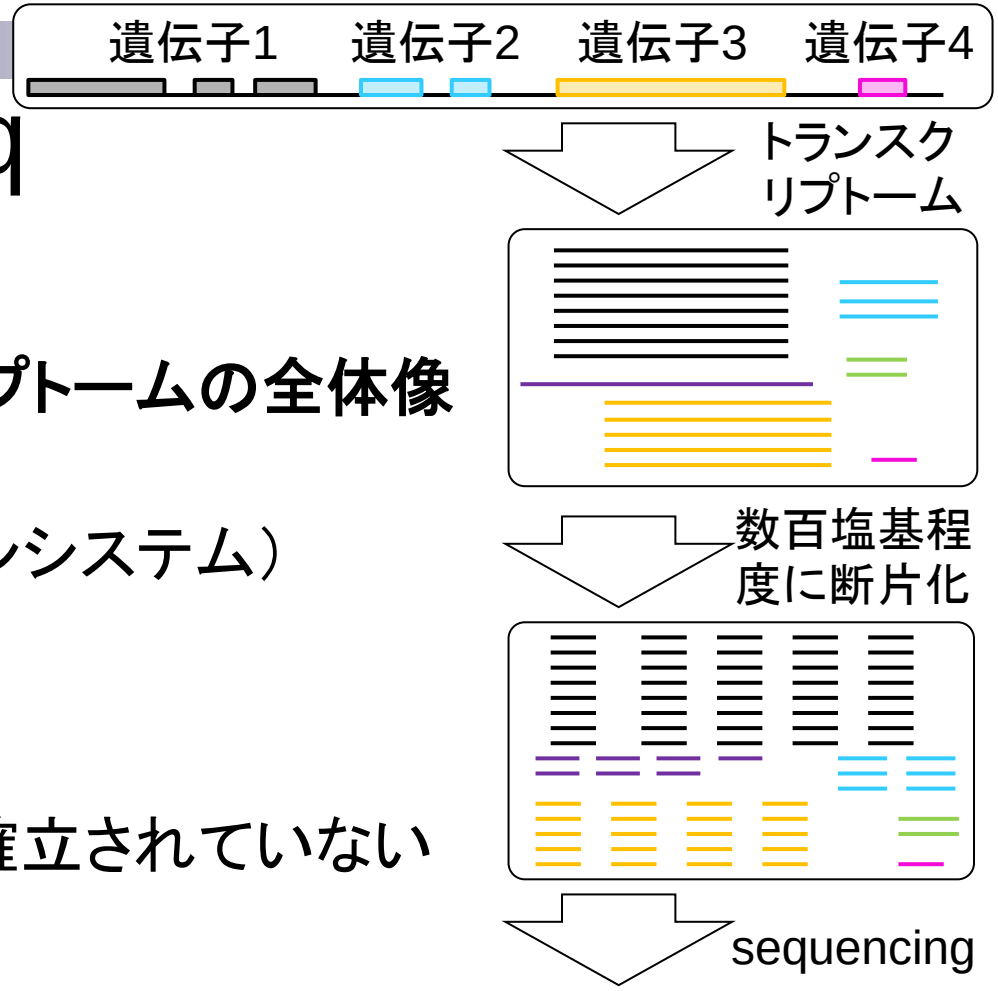
長所・短所：マイクロアレイ

- 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…）

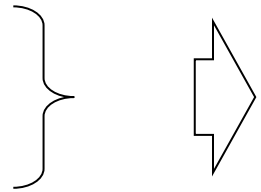
マイクロアレイの臨床応用例

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

既に臨床診断に利用されている

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

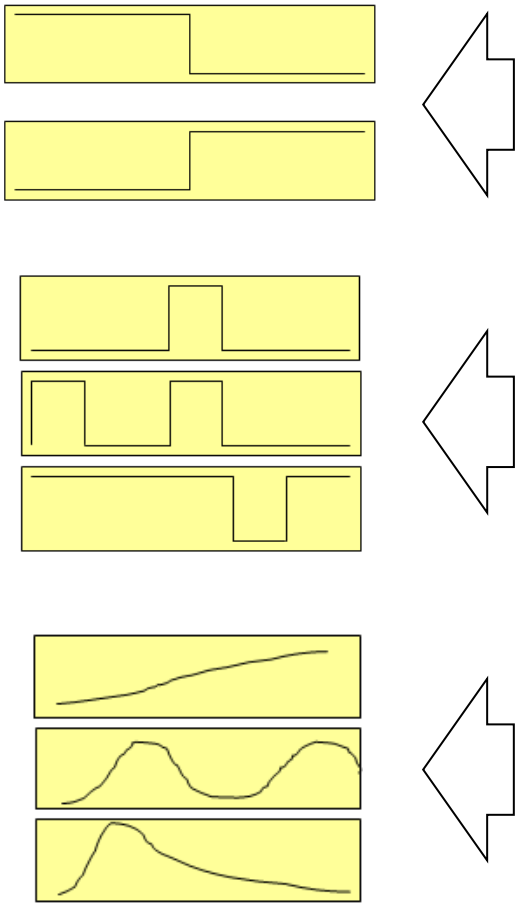
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

- 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

発現変動遺伝子同定



遺伝子発現行列

二群間比較用

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

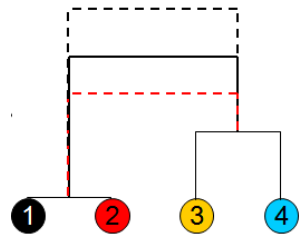
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

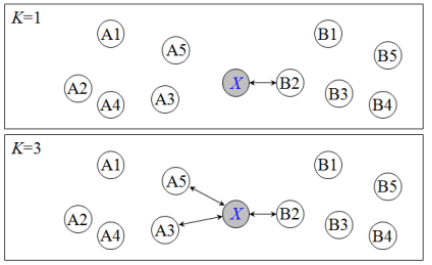
クラスタリング



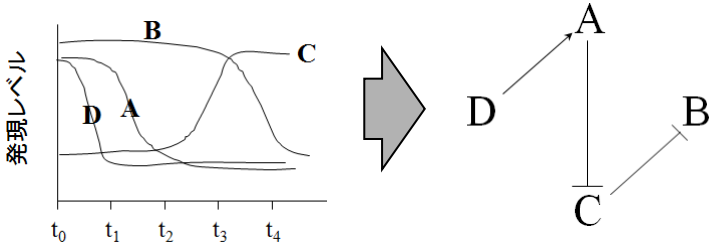
機能解析

- Gene Ontology（GO）
- パスウェイ解析

分類（診断）

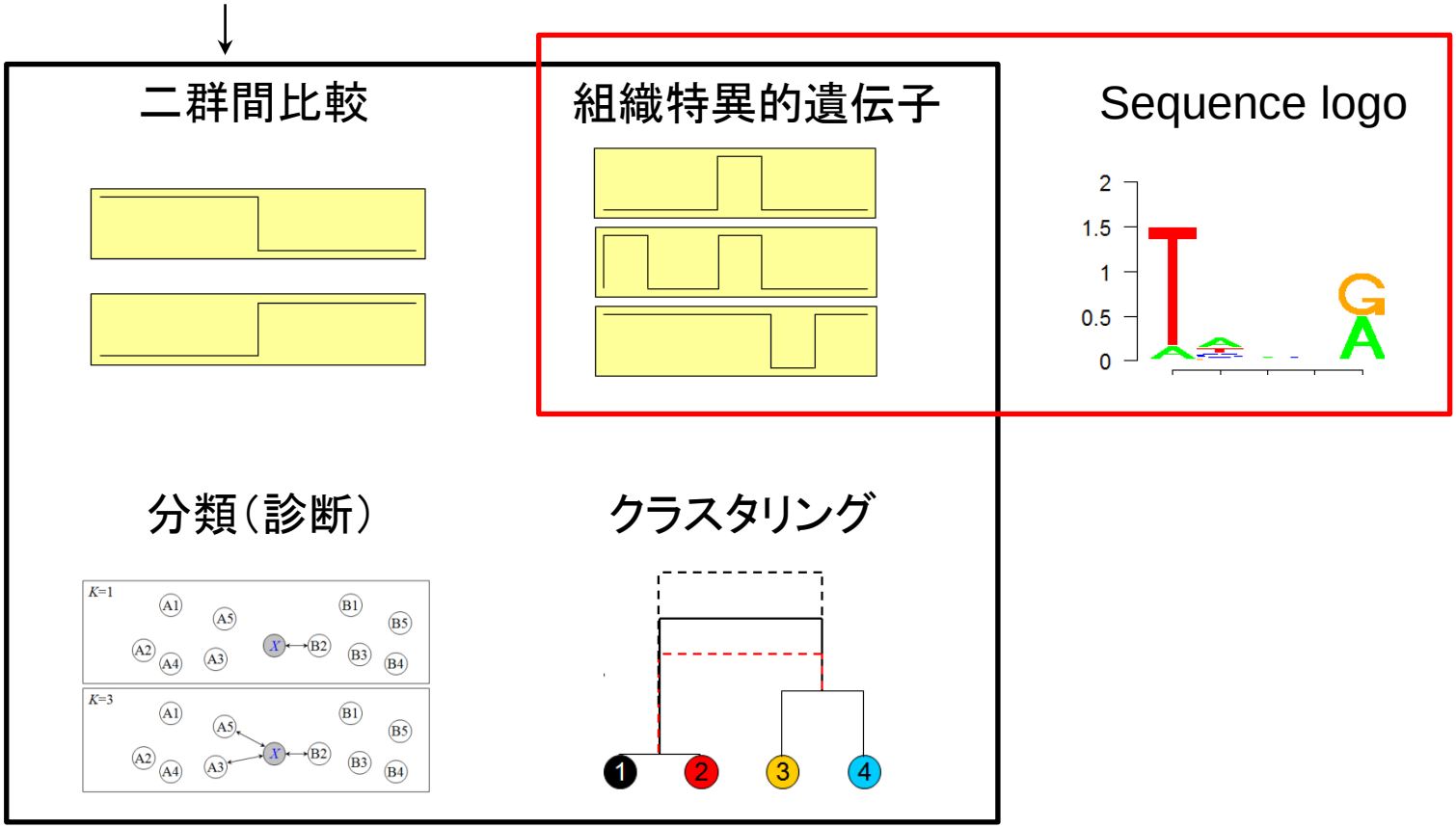


遺伝子ネットワーク推定



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

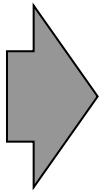


マイクロアレイデータの正規化

- 「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

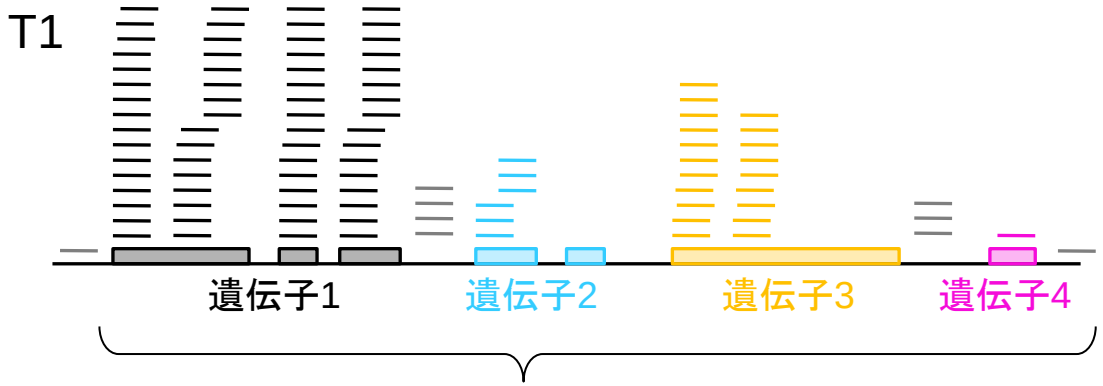


	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
総和	100.0	100.0

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

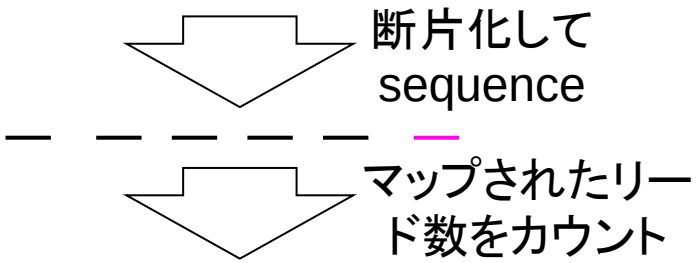
RPM正規化

	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**

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



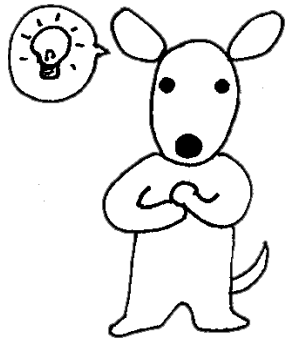
mRNA	リード数
AAAAAAA...	5
AAAAAAA...	1

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

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



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

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

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2012/04/26, since 2005)

What's new?
・R2.14.2がリリースされていたのでこれに変更しました。(2012/03/13)NEW
・最新の
(2011/0
・GSA (
・Hook (
・Agilent
・作図 |
・このペ
あります

- はじめ
- Rのイ
- Rの昔
- 使用
- サン

- 正規化 | Affymetrix GeneChip | GCRMA (Wu 2004) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | PLIER (Affymetrix 2004) (last modified 2010/9/22)
- 正規化 | Affymetrix GeneChip | PDNN (Zhang 2003) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | VSN (Huber 2002) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | MAS, MBEI, RMA (package: affy) (last modified 2009/8/6)
- 前処理について(last modified 2009/7/9)
- 前処理 | スケーリング | サンプル間のシグナル強度の平均値をそろえる (last modified 2012/04/25)
- 前処理 | スケーリング | サンプル間のシグナル強度の中央値をそろえる (last modified 2009/6/5)
- 前処理 | スケーリング | 各サンプルのシグナル強度の平均を0、標準偏差を1にする (last modified 20
- 前処理
- 前処理
- 前処理
- 前処理
- 前処理
- 前処理
- 前処理
- 前処理
- 前処理

sample19.txt

	A	B	C
1	ID	sample1	sample2
2	gene1	10.5	12.4
3	gene2	6.4	7.1
4	gene3	8	8.5
5	gene4	10.8	11.4
6	gene5	5.6	6.7
7	gene6	8.4	8.9
8	gene7	6.2	7
9	gene8	6.1	6.8
10	gene9	6.6	6.5
11	gene10	5.1	5.8

sample19

・前処理 | スケーリング | サンプル間のシグナル強度の平均値をそろえる

よく論文中で各サンプル(各列)の発現データの平均をxにそろえて...などという記述か
ます。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移

1. 10 genes ×2 samplesのデータファイル (sample19.txt) の場合：

```
------    ここから    ------
in_f <- "sample19.txt"                                          #入力ファイル名を指定
out_f <- "hoge.txt"                                             #出力ファイル名を指定
param <- 10                                                     #各サンプルの正規化後の平均値を指定

data_tmp <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="") #ファイルの読み込み
tmp_mean <- apply(data_tmp, 2, mean, na.rm=TRUE)                #各サンプル（列）の平均シグナル強度を計算した結果をtm
data <- sweep(data_tmp, 2, param/tmp_mean, "*")                 #各列中の全てのシグナル値にparam/tmp_meanを掛け、その
data                                                            #結果を表示
apply(data, 2, mean)                                            #各列のmeanを表示させ、正常に動作しているか確認
tmp <- cbind(rownames(data), data)                              #遺伝子IDの列を行列dataの左端に挿入し、結果をtmpに格納
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

「(Rで)マイクロアレイデータ解析」のほうです

実行結果

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ

R Console

```
> tmp_mean <- apply(data_tmp, 2, mean, na.rm=TRUE)                #
> data <- sweep(data_tmp, 2, param/tmp_mean, "*")                  #
> data                                                             #
          sample1   sample2
gene1  14.246947 15.289766
gene2   8.683853  8.754624
gene3  10.854817 10.480888
gene4  14.654003 14.056720
gene5   7.598372  8.261406
gene6  11.397558 10.974106
gene7   8.412483  8.631319
gene8   8.276798  8.384710
gene9   8.955224  8.014797
gene10  6.919946  7.151665
> apply(data, 2, mean)                                             #
sample1 sample2
     10      10
> tmp <- cbind(rownames(data))                                     #
> write.table(tmp, out_f, sep                                  F)#
>
>
```

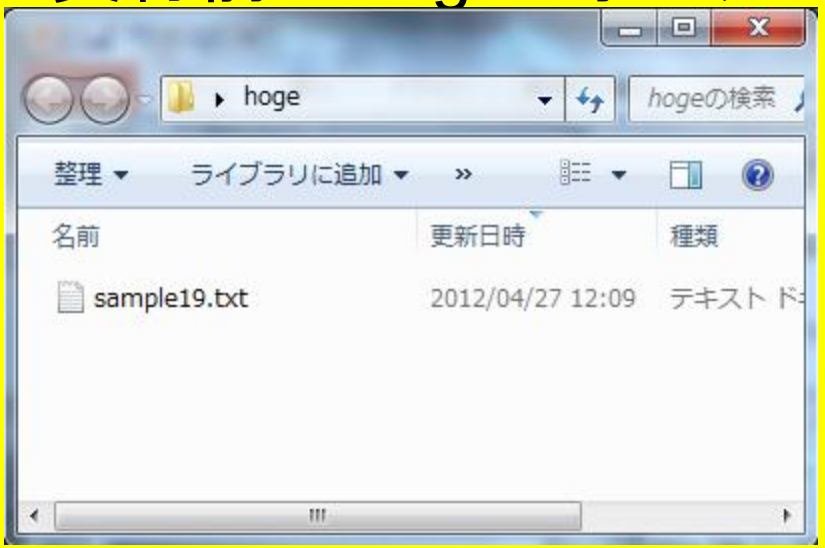
hoge.txt

	A	B	C
1	rownam	sample1	sample2
2	gene1	14.25	15.29
3	gene2	8.68	8.75
4	gene3	10.85	10.48
5	gene4	14.65	14.06
6	gene5	7.60	8.26
7	gene6	11.40	10.97
8	gene7	8.41	8.63
9	gene8	8.28	8.38
10	gene9	8.96	8.01
11	gene10	6.92	7.15

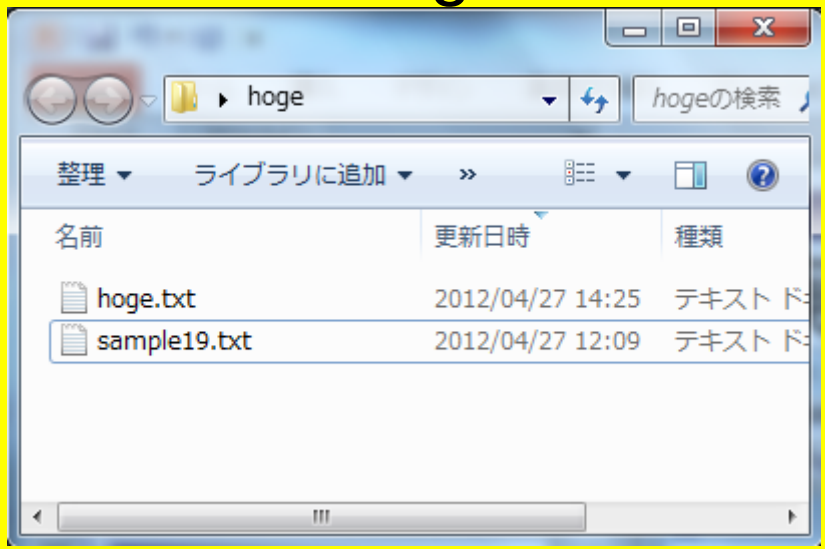
hoge

スライド54と対応

実行前のhogeフォルダ

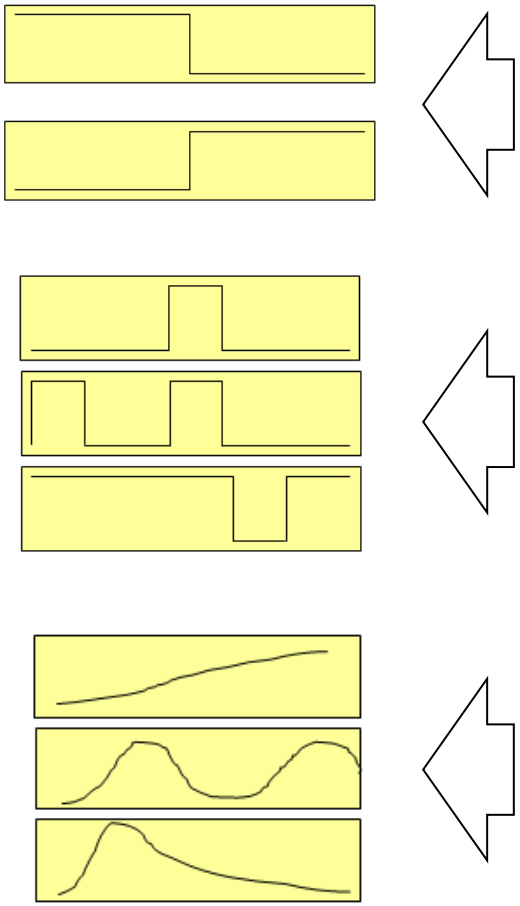


実行後のhogeフォルダ



データ解析の流れ

発現変動遺伝子同定



遺伝子発現行列

二群間比較用

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

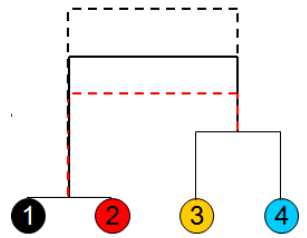
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

クラスタリング

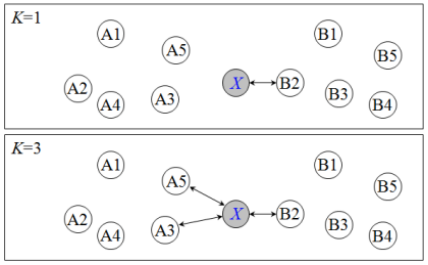


データの背後にある本質的な特徴を把握

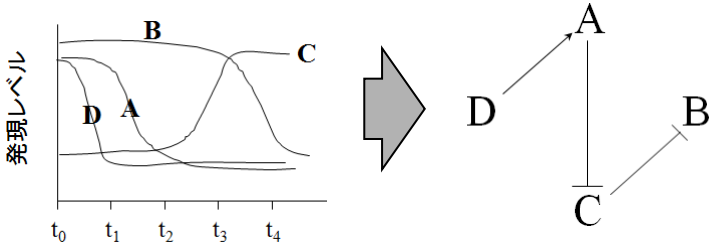
機能解析

- Gene Ontology（GO）
- パスウェイ解析

分類（診断）

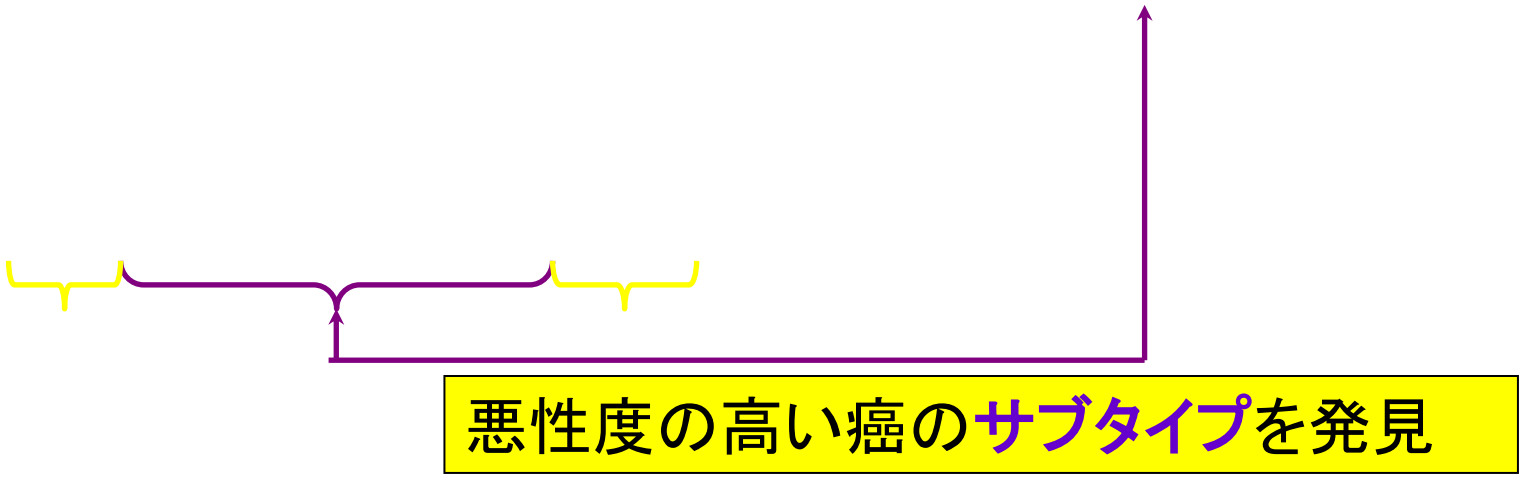


遺伝子ネットワーク推定



（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



クラスタリング（教師なし学習）

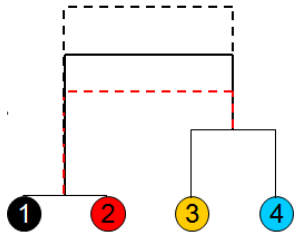
- 決めておくべき二つの基準（事柄）
  - 距離（類似度）の定義
    - ユークリッド距離、マンハッタン距離など
  - クラスターをまとめる（併合する）方法
    - クラスター間の距離を定義する方法、とほぼ同じ
    - 最短距離法、平均連結法、ウォード法など

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

- xとyの発現パターンが酷似 → $r \approx 1$
- xとyの発現パターンがばらばら → $r \approx 0$
- xとyの発現パターンがほぼ正反対 → $r \approx -1$

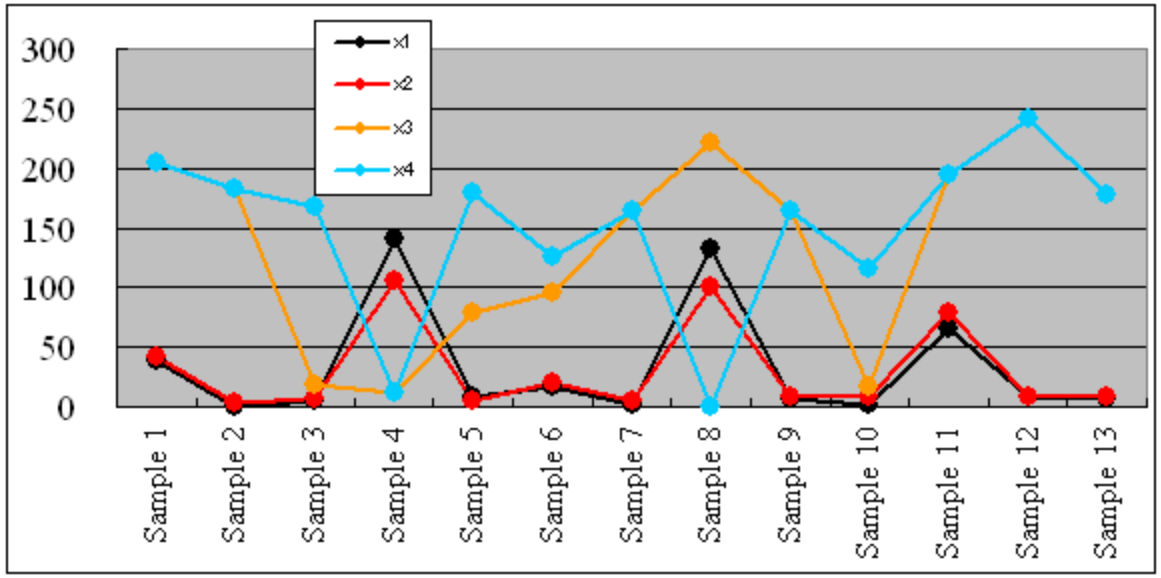
$$\text{距離 } D(x,y) = 1-r \quad (0 \le D \le 2)$$

- $r=1 \rightarrow D=1-1=0$
- $r=0 \rightarrow D=1-0=1$
- $r=-1 \rightarrow D=1-(-1)=2$

「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 $D = \max(| x_1 - y_1 |,... | x_i - y_i |,..., | x_n - y_n |)$
  - 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）

	A	B	C	D	E	F
1		*x*	*y*		|xi − yi|	|xi − yi|/|xi + yi|
2	gene1	10.5	12.4		1.9	0.0830
3	gene2	6.4	7.1		0.7	0.0519
4	gene3	8	8.5		0.5	0.0303
5	gene4	10.8	11.4		0.6	0.0270
6	gene5	5.6	6.7		1.1	0.0894
7	gene6	8.4	8.9		0.5	0.0289
8	gene7	6.2	7		0.8	0.0606
9	gene8	6.1	6.8		0.7	0.0543
10	gene9	6.6	6.5		0.1	0.0076
11	gene10	5.1	5.8		0.7	0.0642

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

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

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

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

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

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

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2012/04/26, since 2005)

What's new?

- 前処理 | フィルタリング | 遺伝子のフィルタリング4 (CVが小さいものを削除) (last modified 2009/8/10)
- 前処理 | 同じ遺伝子名を持つものをまとめる (last modified 2012/04/13) NEW
- 解析 | 二つの遺伝子リストファイルから共通遺伝子部分のみを抽出 (last modified 2012/04/13)
- 解析 | 二つのベクトル間の距離を定義する様々な方法を知る (last modified 2012/05/16) NEW
- 解析 | 遺伝子ごとの平均発現量など (last modified 2009/8/6)

解析 | 二つのベクトル間の距離を定義する様々な方法を知る

二つのベクトル間の距離を定義する方法は多数存在します。ここでは10 genes ×2 samplesのデータ (sample19.txt)を読み込んで二つのサンプル間の距離をいくつかの方法で算出します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し、以下をコピ

```
1. 10 genes ×
------  ここ
in_f <- "samp

#ファイルの読
data <- read.t

#本番
dist(t(data),
dist(t(data),
dist(t(data),
dist(t(data),
1 - cor(data,

dist(t(data),
dist(t(data),
1 - cor(data,

------  ここ
```

sample19.txt

	A	B	C
1	ID	sample1	sample2
2	gene1	10.5	12.4
3	gene2	6.4	7.1
4	gene3	8	8.5
5	gene4	10.8	11.4
6	gene5	5.6	6.7
7	gene6	8.4	8.9
8	gene7	6.2	7
9	gene8	6.1	6.8
10	gene9	6.6	6.5
11	gene10	5.1	5.8

他にどんな距離を利用可能か調べたい場合は…

R Console

```
> in_f <- "sample19.txt"
>
> #ファイルの読み込み
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote=""
>
> #本番
> dist(t(data), method="euclidean")                    #ユークリッド距離
         sample1
sample2 2.792848
> dist(t(data), method="manhattan")                    #マンハッタン距離
        sample1
sample2     7.6
> dist(t(data), method="maximum")                      #最大距離
        sample1
sample2     1.9
> dist(t(data), method="canberra")                     #キャンベラ距離
          sample1
sample2 0.4972074
```

「?関数名」で詳細な使用法を学ぶ

R Console

```
> ?dist
starting httpd hel
>
> dist(t(data))
          sample1
sample2 2.792848
>
```

Description

This function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.

Usage

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

as.dist(m, diag = FALSE, upper
## Default S3 method:
as.dist(m, diag = FALSE, upper

## S3 method for class 'dist'
print(x, diag = NULL, upper = NULL,
      digits = getOption("digits"), justify = "none",
      right = TRUE, ...)

## S3 method for class 'dist'
as.matrix(x, ...)
```

ユークリッド距離でよければ、「method=“xxx”」のところを記述しなくてもいいようだ

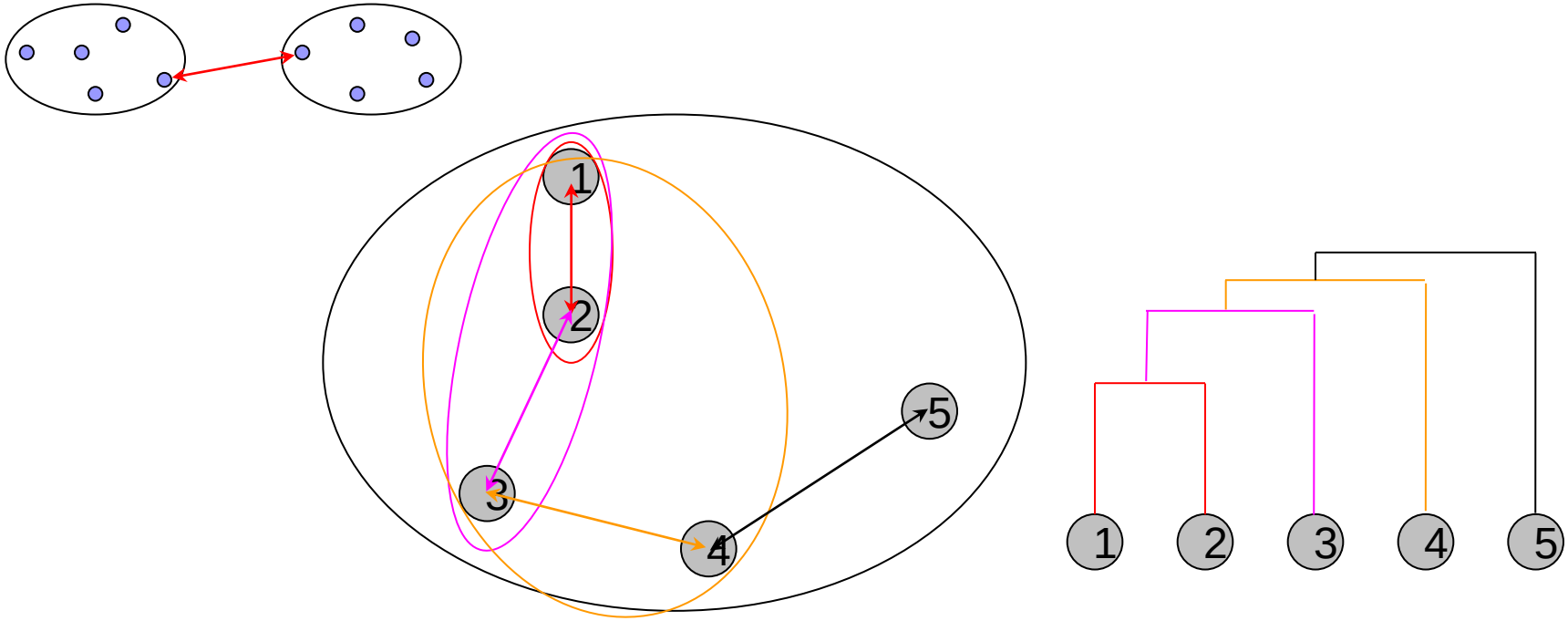
Arguments

x	a numeric matrix, data frame or "dist" object.
method	the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.

“binary”や”minkowski”というものも指定できるようだが「1-相関係数」を指定することはできないようだ…orz

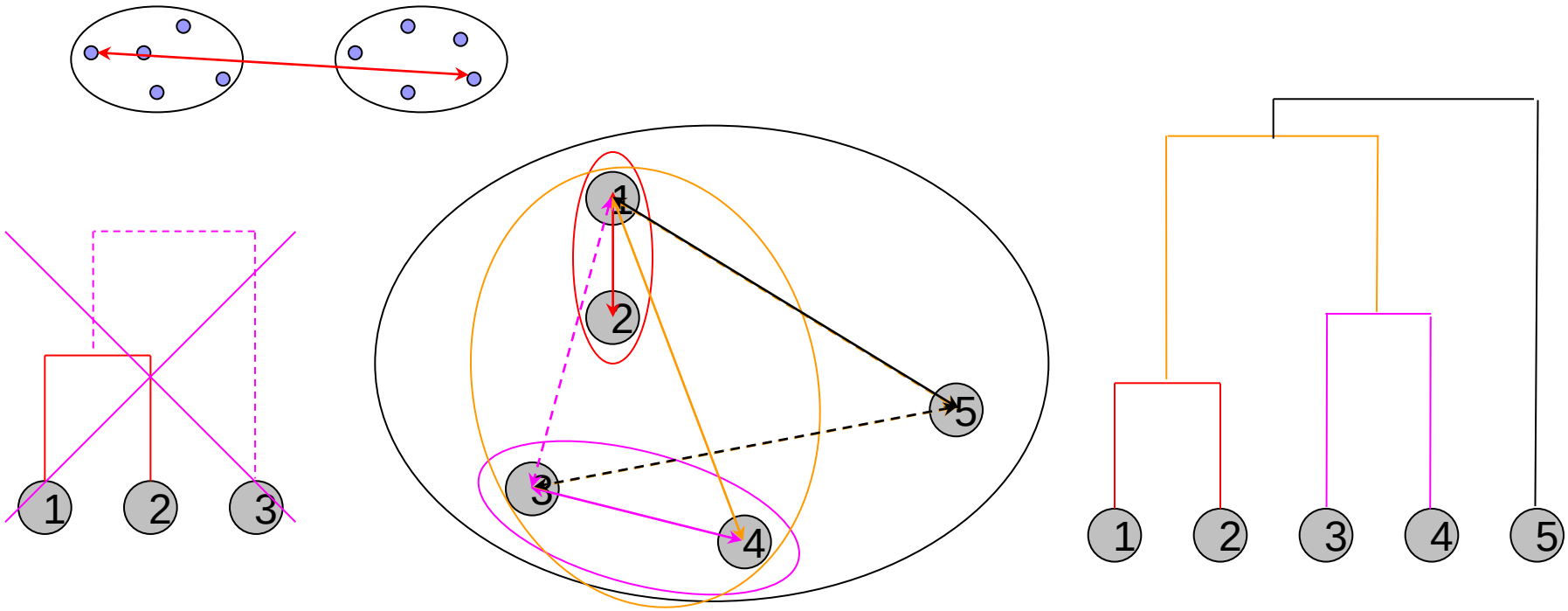
クラスターを併合する方法

- 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

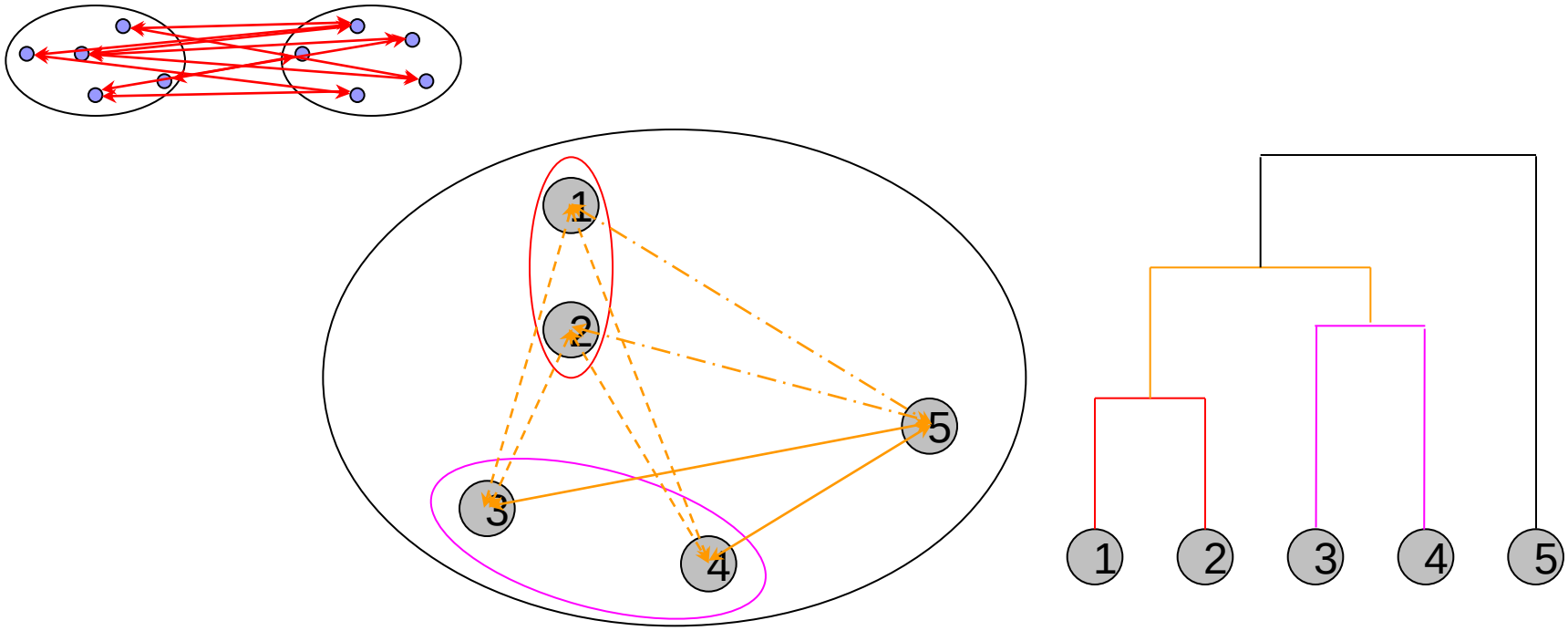
- 最長距離法（完全連結法；complete-linkage ）
  - クラスターをmerge（併合、合併）するときの基準として、最遠距離を用いる



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

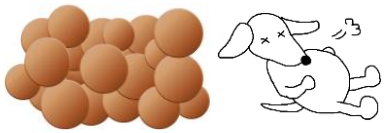


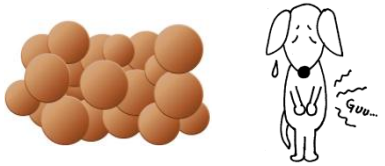
用いる方法によって得られるデンドログラム（樹形図）が異なる

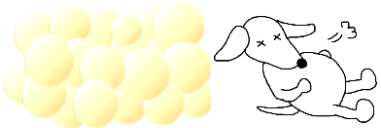
他のクラスター併合手段

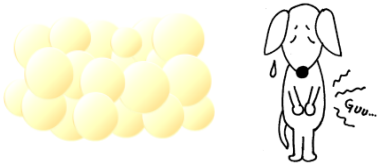
- 重心法（Centroid）：重心間距離を利用
- ウォード法：群内平方和の増加量が最小となるクラスターと併合
- メディアン（Median）法：群間距離の中央値を利用
- McQuitty法…
- 可変（flexible）法…

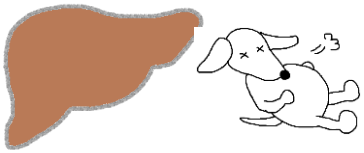
実データをクラスタリングしてみよう

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**」

	A	B	C	D	E
1		褐色脂肪_満腹1	褐色脂肪_満腹2	褐色脂肪_満腹3	…
2	1367452_at	10.52	10.23	10.35	
3	1367453_at	9.66	10.05	9.90	
4	1367454_at	9.65	9.40	9.44	
5	1367455_at	10.76	11.10	10.82	
6	1367456_at	11.71	11.60	11.59	
7	…				

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2012/04/26, since 2005)

- 解析｜機能解析｜Gene Set Analysis (GSA) (Efron_2007?!) (last modified 2010/8/30)
- 解析｜機能解析｜Gene Ontlogy解析｜topGO (Alexa_2006) (last modified 2011/12/14)
- 解析｜クラスタリング｜階層的｜について (last modified 2009/8/12)
- 解析｜クラスタリング｜階層的｜pvclust (Suzuki_2006) (last modified 2010/8/5)
- 解析｜クラスタリング｜階層的｜hclust (last modified 2011/12/20)
- 解析｜クラスタリング｜階層的｜hclust後の詳細な解析 (last modified 2009/8/7)

- **解析｜クラスタリング｜階層的｜hclust**

1. サンプル間クラスタリング(類似度：「1-Pearson相関係数」、方法：平均連結法(average))でR Graphics画面上に表示したい場合：

```
------    ここから    ------
in_f <- "sample3.txt"                                   #入力ファイル名(発現データファイル)を指定
param2 <- "average"                                     #方法(method)を指定

data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.
out <-
plot(

------
```

「data_rma_2.txt」ファイルを入力データとして与え、
「Fig1.png」ファイルに出力したいときは？

2. サンプル間クラスタリング(類似度：「1-Pearson相関係数」、方法：平均連結法(average))でpng形式のファイルで図の大きさを指定

```
------    ここから    ------
in_f <- "sample3.txt"                                   #入力ファイル名(発現データファイル)を指定
param2 <- "average"                                     #方法(method)を指定
out_f <- "hoge.png"                                     #出力ファイル名(クラスタリング結果ファイル)を指定
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

data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.dist <- as.dist(1 - cor(data, method="pearson"))   #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method=param2)                 #階層的クラスタリングを実行し、結果をoutに格納
png(out_f, pointsize=param5, width=param3, height=param4)#出力ファイルの各種パラメータを指定
plot(out)                                               #樹形図（デンドログラム）の表示
dev.off()                                               #おまじない

------    ここまで    ------
```

2. サンプル間クラスタリング(類似度：「1-Pearson相関係数」、方法：平均連結法(average))でpng形式のファイルで図の大きさを指定し

------　ここから　------

```
in_f <- "sample3.txt"                    #入力ファイル名(発現データファイル)を指定
param2 <- "average"                      #方法(method)を指定
out_f <- "hoge.png"                      #出力ファイル名(クラスタリング結果ファイル)を指定
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
```

切り取り(T)
コピー(C)
貼り付け

①テンプレートのスクリプトをコピーして

無題 - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
in_f <- "sample3.txt"                    #入力ファイル名(発現データファイル)を指定
param2 <- "average"                      #方法(method)を指定
out_f <- "hoge.png"                      #出力ファイル名(クラスタリング結果ファイル)を指定
param3 <- 500
param4 <- 400
param5 <- 14

data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.dist <- as.dist(1 - cor(data, method="pearson"))  #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method=param2)  #階層的クラスタリングを実行し、結果をoutに格納
png(out_f, pointsize=param5, width=param3, height=param4)  #出力ファイルの各種パラメータを指定
plot(out)                                #樹形図（デンドログラム）の表示
dev.off()                                #おまじない
```

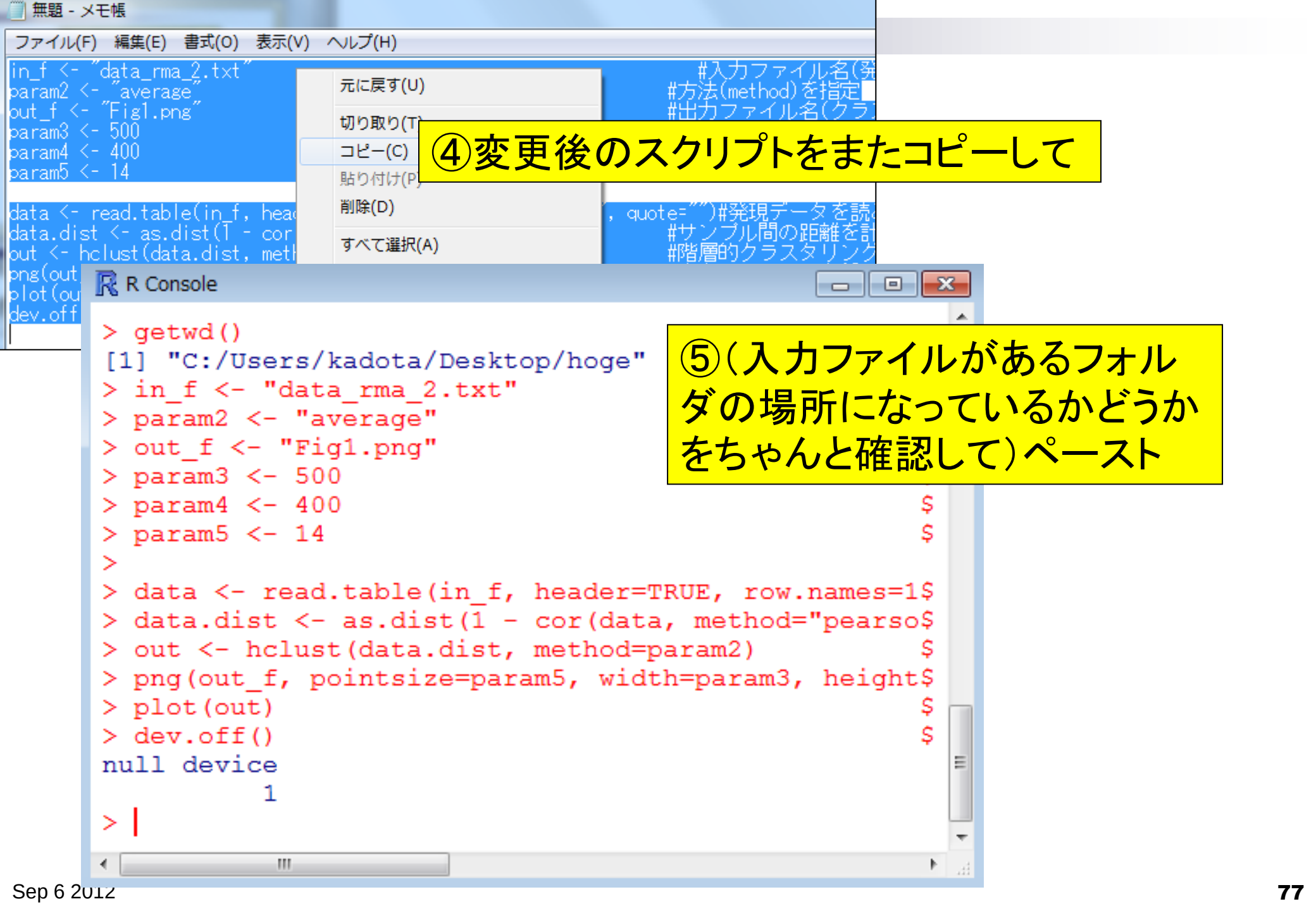
②メモ帳などのテキストエディタにペーストして

無題 - メモ帳

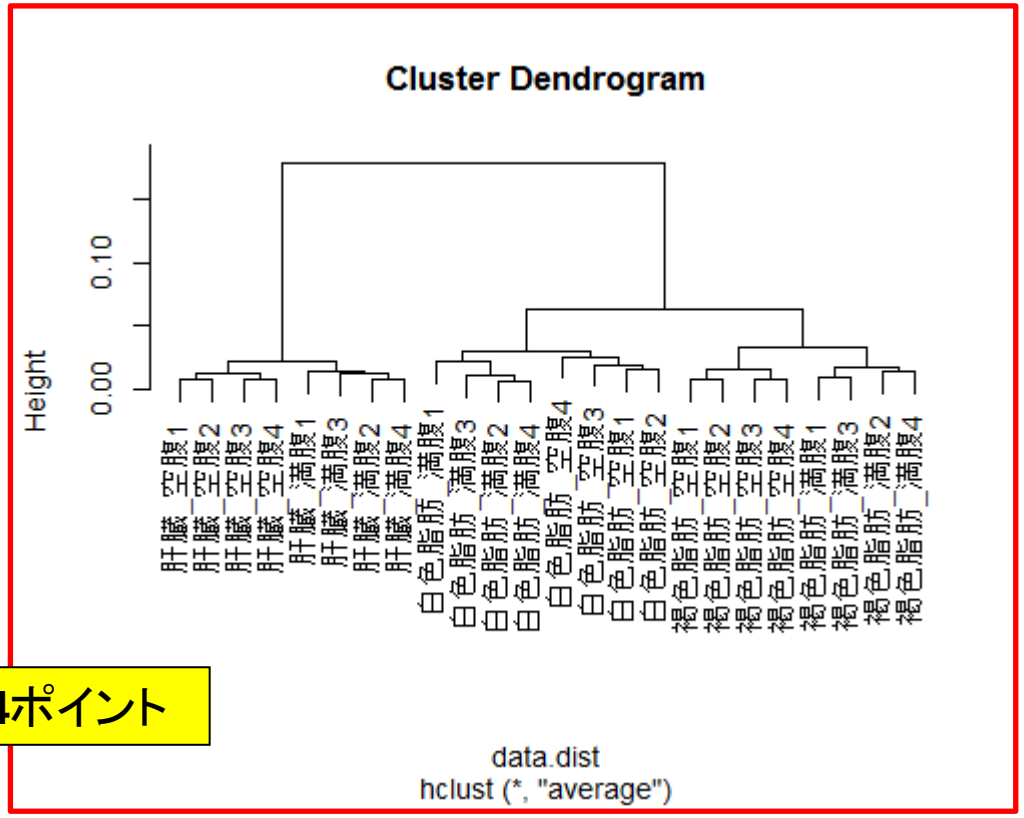
ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
in_f <- "data_rma_2.txt"                 #入力ファイル名(発現データファイル)を指定
param2 <- "average"                      #方法(method)を指定
out_f <- "Fig1.png"                      #出力ファイル名(クラスタリング結果ファイル)を指定
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

data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.dist <- as.dist(1 - cor(data, method="pearson"))  #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method=param2)  #階層的クラスタリングを実行し、結果をoutに格納
png(out_f, pointsize=param5, width=param3, height=param4)  #出力ファイルの各種パラメータを指定
plot(out)                                #樹形図（デンドログラム）の表示
dev.off()                                #おまじない
```

③必要な箇所を変更して


無題 - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
in_f <- "data_rma_2.txt"
param2 <- "average"
out_f <- "Fig1.png"
param3 <- 500
param4 <- 400
param5 <- 14
元に戻す(U)
切り取り(T)
コピー(C)
貼り付け(P)
削除(D)
すべて選択(A)
④変更後のスクリプトをまたコピーして
R Console
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> in_f <- "data_rma_2.txt"
> param2 <- "average"
> out_f <- "Fig1.png"
> param3 <- 500
> param4 <- 400
> param5 <- 14
>
> data <- read.table(in_f, header=TRUE, row.names=1
> data.dist <- as.dist(1 - cor(data, method="pearso
> out <- hclust(data.dist, method=param2)
> png(out_f, pointsize=param5, width=param3, height
> plot(out)
> dev.off()
null device
1
>
⑤(入力ファイルがあるフォルダの場所になっているかどうかをちゃんと確認して)ペースト

出力ファイル（Fig1.png）の解説



縦幅が400ピクセル

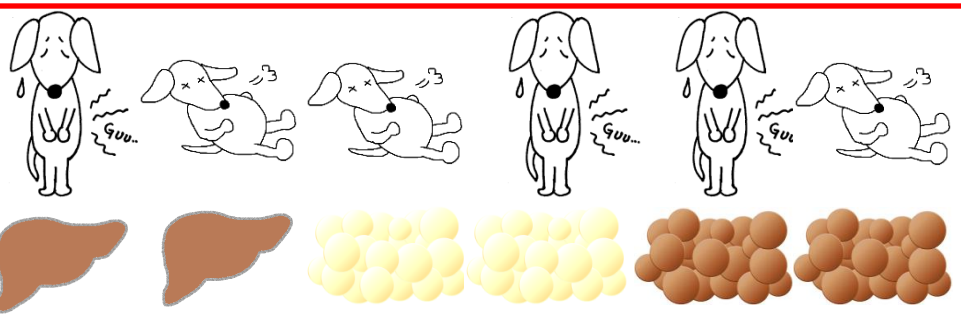
文字の大きさが14ポイント

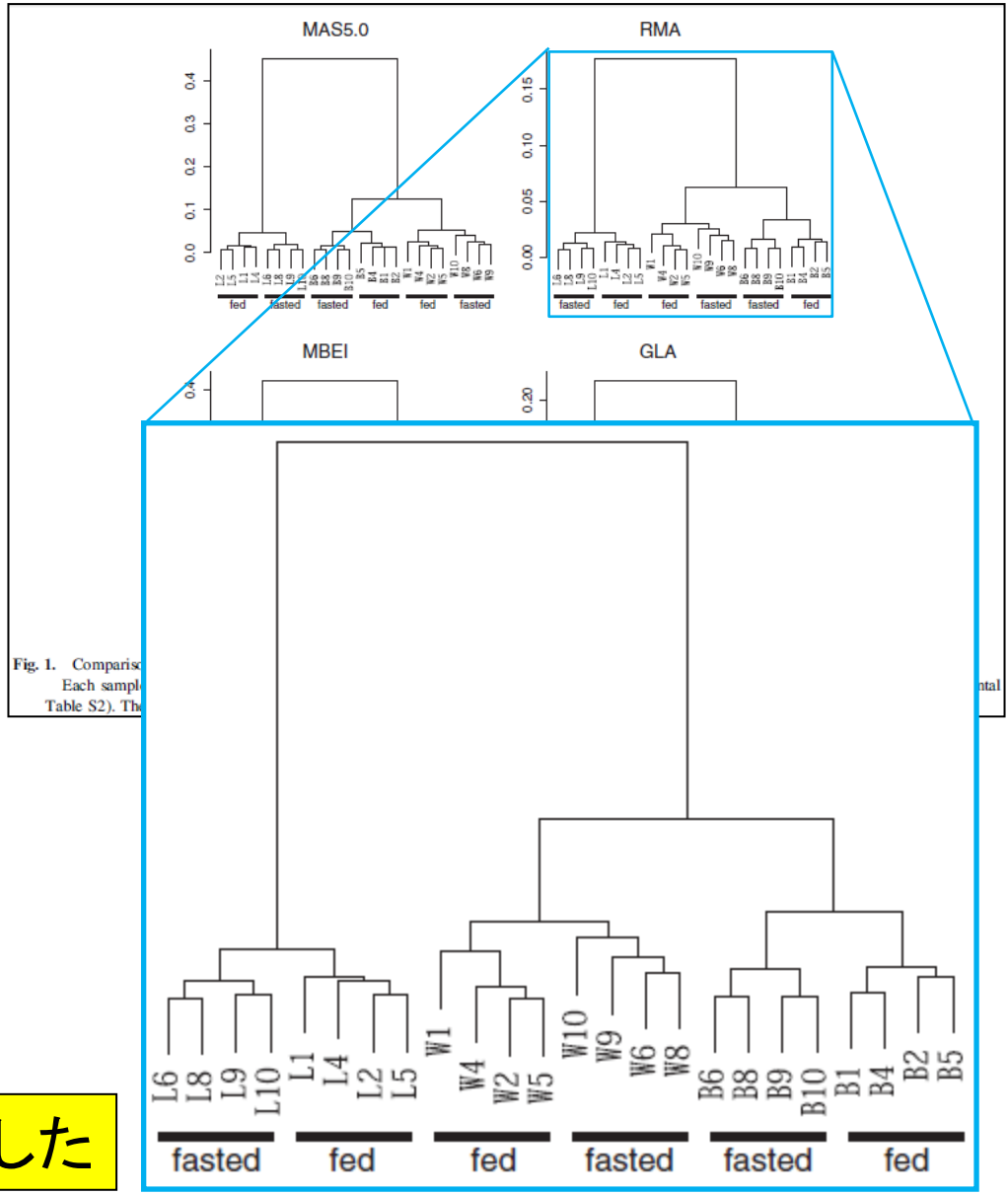
横幅が500ピクセル

```
param3 <- 500
param4 <- 400
param5 <- 14
```

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

原著論文と比較



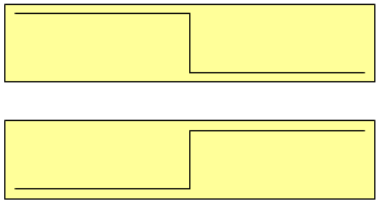


ばっちりFig. 1（の一部）を再現できました

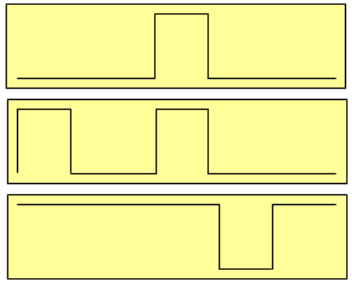
講義内容

発現変動遺伝子同定

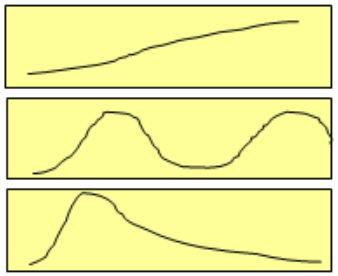
1. 二群間比較



2. 様々な組織（条件）



3. 時系列データ



遺伝子発現行列

二群間比較用

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

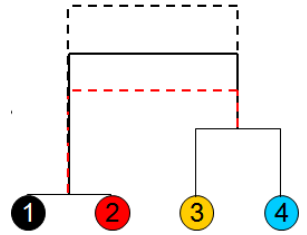
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

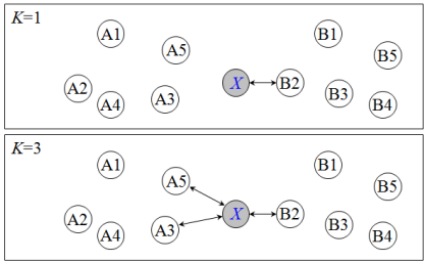
クラスタリング



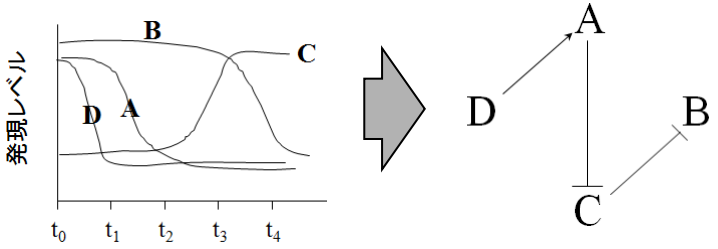
機能解析

・Gene Ontology（GO）
・パスウェイ解析

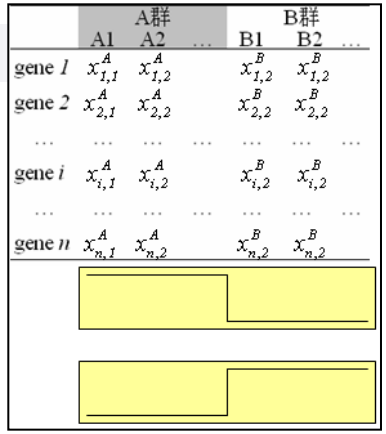
分類（診断）



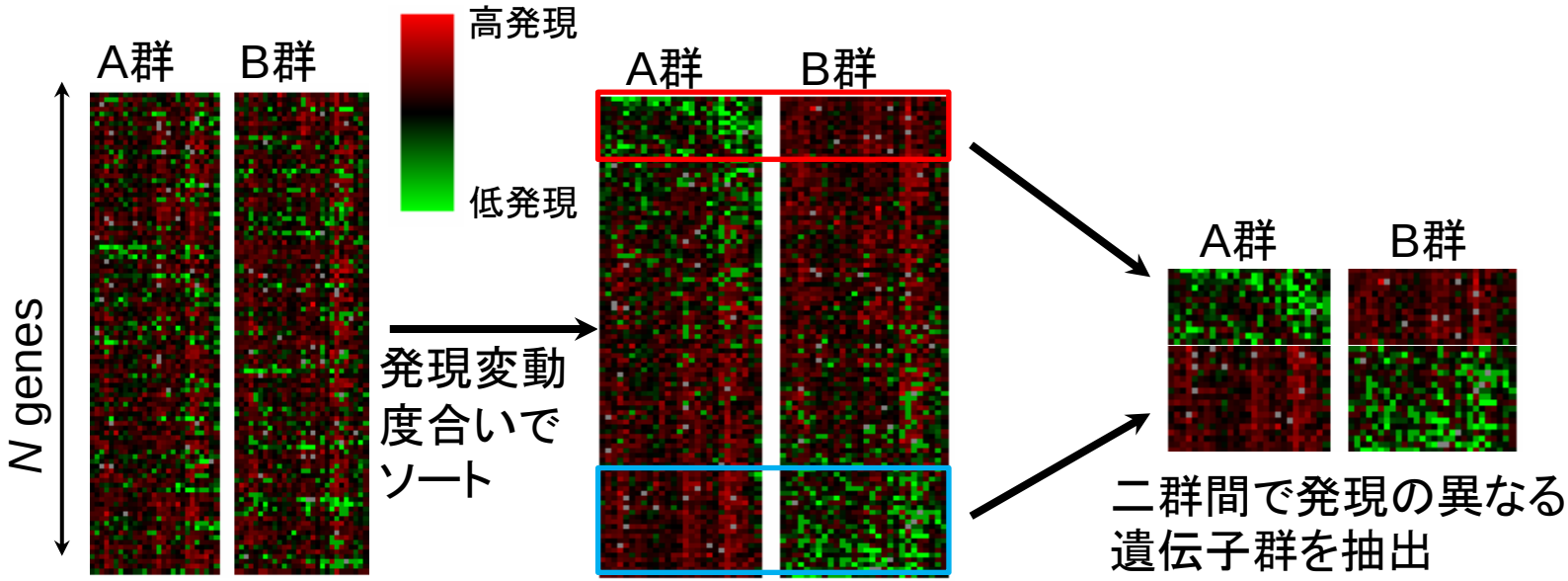
遺伝子ネットワーク推定



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）

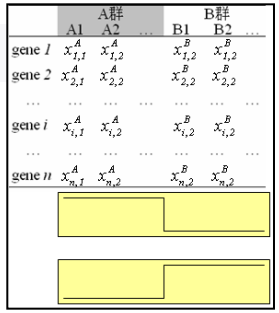


解析例(二群間比較)

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

急性リンパ性白血病　急性骨髄性白血病

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

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

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

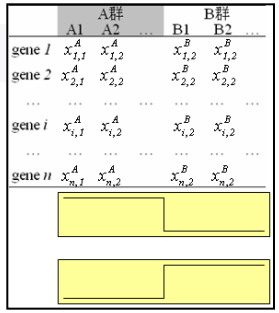
$$r_{gene3} = \frac{-6.41}{14.88 \times 0.52} = -0.825$$

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

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2012/05/18, since 2005)

What's new?

- R2.14
- 最新(
(2011/
- GSA
- Hook
- Agilen
- 作図
- この
あり



- 解析｜発現変動遺伝子｜二群間｜対応なし｜Student's t-test (last modified 2009/7/28)
- 解析｜発現変動遺伝子｜二群間｜対応なし｜Welch t-test (last modified 2009/7/28)
- 解析｜発現変動遺伝子｜二群間｜対応なし｜Mann-Whitney U-test (last modified 2009/7/28)
- 解析｜発現変動遺伝子｜二群間｜対応なし｜パターンマッチング法 (last modified 2011/10/13)
- 解析｜発現変動遺伝子｜二群間｜対応あり｜について (last modified 2009/11/11)
- 解析｜発現変動遺伝子｜二群間｜対応あり｜SAM (Tusher 2001) (last modified 2009/7/28)
- 解析｜発現変動遺伝子｜二群間｜対応あり｜SAM (Tusher 2001)にWADのようなシグナル強度による重みをかけたい (
- 解析
- 解析
- 解析
- 解析
- 解析
- 解析

• 解析｜発現変動遺伝子｜二群間｜対応なし｜パターンマッチング法

パターンマッチング法を用いて、二群間での発現変動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15中のsample16.txtファイル（遺伝子発現データ）とsample16_cl.txtファイル（クラスラベルデータ）を置いてあるディレクトリに移動し、以下をコピペ

1. クラスラベル情報ファイル（sample16_cl.txt）を読み込んでテンプレートパターン情報を得る場合：

```
------    ここから    ------
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data.clとの
tmp <- cbind(rownames(data), data, r)                      #入力データの右側に相関係数rのベクトルを結合した結果をt
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

パターンマッチング法を用いて、二群間

「ファイル」－「ディレクトリの変更」で解析
とsample16_cl.txtファイル（クラスラベル

1．クラスラベル情報ファイル（sample16

```
------    ここから    ------
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data.clと
tmp <- cbind(rownames(data), data, r)                      #入力データの右側に相関係数rのベクトルを結合した結果
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

sample16.txt

	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

sample16

sample16_cl.txt

	A	B
1	A1	1
2	A2	1
3	A3	1
4	A4	1
5	A5	1
6	A6	1
7	B1	0
8	B2	0
9	B3	0
10	B4	0
11	B5	0

sample1

R Console

```
> 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$
> out_f <- "hoge.txt"                                      #出力ファイル名を指定
>
> data <- read.table(in_f1,
> hoge <- read.table(in_f2,
> data.cl <- hoge[,2]
> r <- apply(data, 1, cor, y
> tmp <- cbind(rownames(data
> write.table(tmp, out_f, se
>
```

hoge.txt 出力ファイル

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	rownames(data)	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	r
2	gene1	87	79	91	82	90	84	12	21	19	13	17	0.9936
3	gene2	56	122	106	47	84	98	7	44	2	11	18	0.8418
4	gene3	15	28	33	9	27	41	48	46	52	50	49	-0.825

hoge

• 解析 | 発現変動遺伝子 | 二群間 | 対応なし | パタ

パターンマッチング法を用いて、二群間での発現変動遺伝

「ファイル」－「ディレクトリの変更」で解析したいサンプルマ
とsample16_cl.txtファイル（クラスラベルデータ）を置いてあ

1. クラスラベル情報ファイル（sample16_cl.txt）を読み込ん

```
------    ここから    ------
in_f1 <- "sample16.txt"
in_f2 <- "sample16_cl.txt"
out_f <- "hoge.txt"

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.
data.cl <- ho
r <- apply(da
tmp <- cbind(
write.table(t

------    こ
```

#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
#出力ファイル名を指定

ンdata.clとの
合した結果をt
存。

sample16.txt

	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

sample16

R Console

```
>
>
> data <- read.table(in_f1, header=TRUE, row.na$
> dim(data)
[1]  3 11
> 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
>
```

オブジェクトdataには遺伝子発現データが入っている

1. クラスラベル情報ファイル（sample16_cl.txt）を読み込んでテンプレートパターン情報を得る場合：

```
------    ここから    ------
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
da                                                                     #ターンベクトルdata.clを作成
r                                                                      #についてテンプレートパターンdata.clと
tm                                                                     #に相関係数rのベクトルを結合した結果
wr                                                                     #fで指定したファイル名で保存。

--
```

sample16_cl.txt

	A	B
1	A1	1
2	A2	1
3	A3	1
4	A4	1
5	A5	1
6	A6	1
7	B1	0
8	B2	0
9	B3	0
10	B4	0
11	B5	0

sample1

R Console

```
> hoge <- read.table(in_f2, sep="\t", q$
> dim(hoge)
[1] 11  2
> hoge
    V1 V2
1   A1  1
2   A2  1
3   A3  1
4   A4  1
5   A5  1
6   A6  1
7   B1  0
8   B2  0
9   B3  0
10 B4  0
11 B5  0
>
```

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…

1. クラスラベル情報ファイル（sample16_cl.txt）を読み込んでテンプレートパターン情報を得る場合：

```
------    ここから    ------
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
```

トパターンdata.clと
トルを結合した結果
ル名で保存。

R Console

```
> data.cl <- hoge[,2]                                    $
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
> length(data.cl)
[1] 11
>
```

sample16_cl.txt

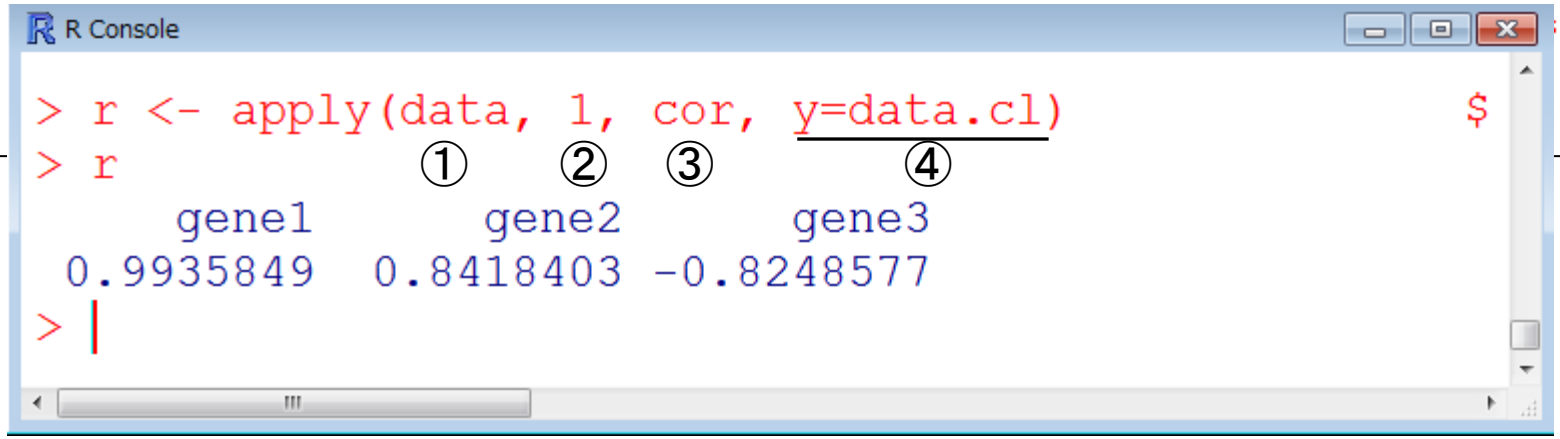
	A	B
1	A1	1
2	A2	1
3	A3	1
4	A4	1
5	A5	1
6	A6	1
7	B1	0
8	B2	0
9	B3	0
10	B4	0
11	B5	0

sample1

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

1. クラスラベル情報ファイル（sample16_cl.txt）を読み込んでテンプレートパターン情報を得る場合：

```
------    ここから    ------
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data.clと
                                                           数rのベクトルを結合した結果を
                                                           たファイル名で保存。
```



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

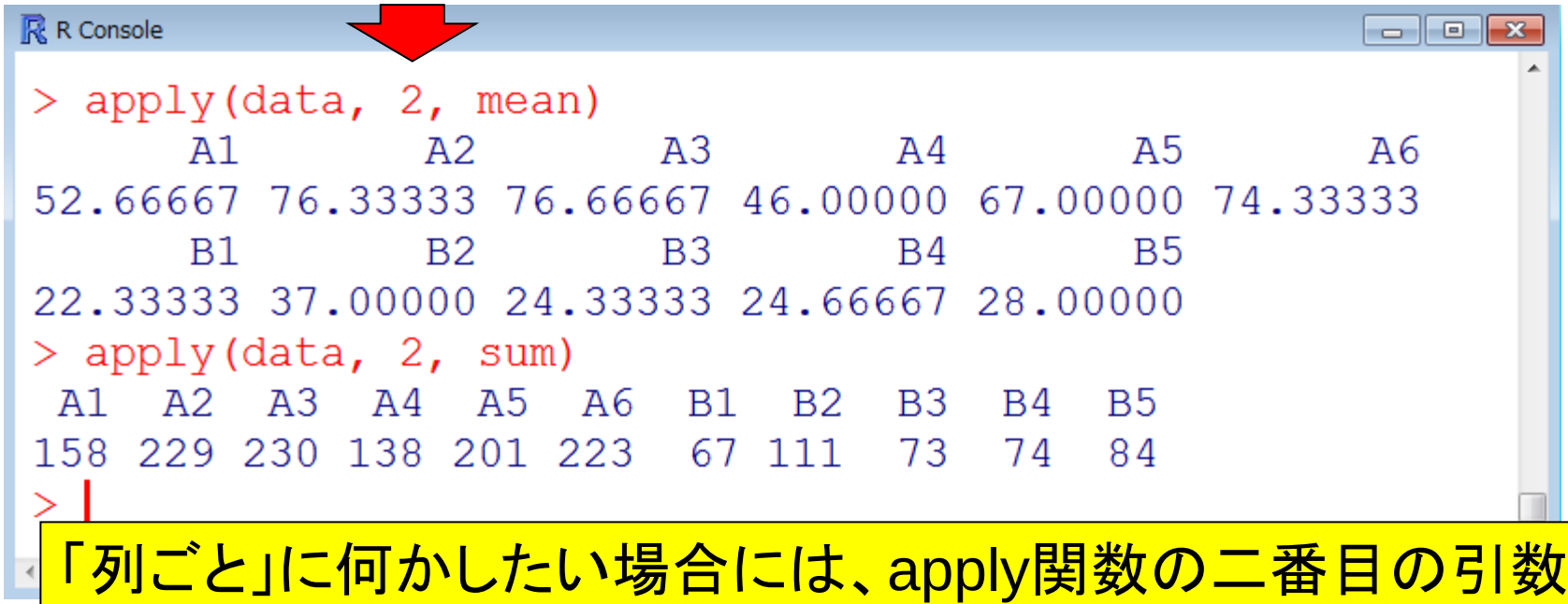
R Console

```
> 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
> apply(data, 1, mean)
   gene1    gene2    gene3
54.09091 54.09091 36.18182
> apply(data, 1, sum)
gene1 gene2 gene3
  595   595   398
>
```

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5		平均	和
2	gene1	87	79	91	82	90	84	12	21	19	13	17		54.09	595
3	gene2	56	122	106	47	84	98	7	44	2	11	18		54.09	595
4	gene3	15	28	33	9	27	41	48	46	52	50	49		36.18	398

apply関数



「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

R Console

```
> ?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

- 第24節の「apply()ファミリー」のところ

R Console

```
> ?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

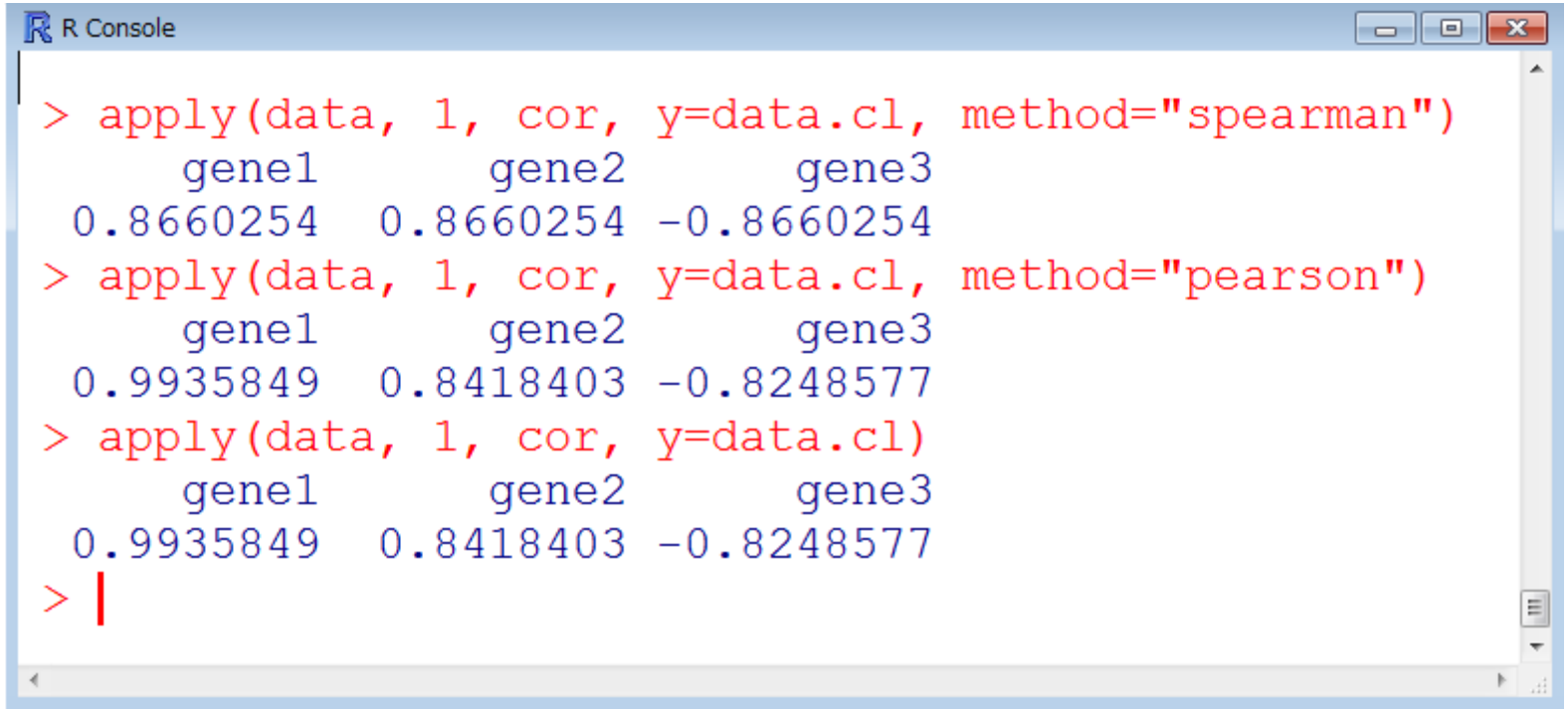
```
var(x, y = NULL, na.rm = FALSE, use)

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
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
wr                                               #保存
```

R Console

```
> 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

sample16.txt

	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

sample16

①

②

```
in_f1 <- "sample16.txt"
in_f2 <- "sample16_cl.txt"
out_f <- "hoge.txt"

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="")
data.cl <- hoge[,2]
r <- apply(data, 1, cor, y=data.cl)
tmp <- cbind(rownames(data), data, r)
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)
```

#出力ファイル名を指定
#入力ファイル2を読み込んでhogeに格納
#テンプレートパターンベクトルdata.clを作成

①読み込むファイルの最初の行は（数値データではなく）ヘッダー部分です
②最初の列は（数値データではなく）行の名前に相当する部分です

という処理を読み込み時に行なっているからrownames関数やcolnames関数で該当情報を抽出可能

R Console

```
> 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
> rownames(data)
[1] "gene1" "gene2" "gene3"
> colnames(data)
 [1] "A1" "A2" "A3" "A4" "A5" "A6" "B1" "B2" "B3" "B4" "B5"
>
```

cbind関数（やrbind関数）使い倒し

R Console

```
> tmp <- cbind(r, rownames(data), data)
> tmp
              r rownames(data) A1  A2  A3 A4 A5 A6 B1 B2 B3 B4 B5
gene1  0.9935849          gene1 87  79  91 82 90 84 12 21 19 13 17
gene2  0.8418403          gene2 56 122 106 47 84 98  7 44  2 11 18
gene3 -0.8248577          gene3 15  28  33  9 27 41 48 46 52 50 49
> 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
> tmp <- rbind(data, apply(data, 2, sum))
> tmp
   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
4     158 229 230 138 201 223 67 111 73 74 84
> 
```

cbind関数は列（coloum）方向で結合
rbind関数は行（row）方向で結合

パターンマッチング法を用いて、二群間

「ファイル」－「ディレクトリの変更」で解析
とsample16_cl.txtファイル（クラスラベル

1．クラスラベル情報ファイル（sample16

```
------    ここから    ------
in_f1 <- "sample16.txt"
in_f2 <- "sample16_cl.txt"
out_f <- "hoge.txt"

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data.clと
tmp <- cbind(rownames(data), data, r)                                  #入力データの右側に相関係数rのベクトルを結合した結果
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    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
#出力ファイル名を指定

sample16.txt

	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

sample16

sample16_cl.txt

	A	B
1	A1	1
2	A2	1
3	A3	1
4	A4	1
5	A5	1
6	A6	1
7	B1	0
8	B2	0
9	B3	0
10	B4	0
11	B5	0

sample1

テンプレートパターンとの相関係数（の絶対値）が高い遺伝子ほど、より二群間で発現変動している

R Console

```
> in_f1 <- "sample1
> in_f2 <- "sample1
> out_f <- "hoge.txt"
>
> data <- read.table(in_f1,
> hoge <- read.table(in_f2,
> data.cl <- hoge[,2]
> r <- apply(data, 1, cor, y
> tmp <- cbind(rownames(data
> write.table(tmp, out_f, se
>
```

#出力ファイル名を指定

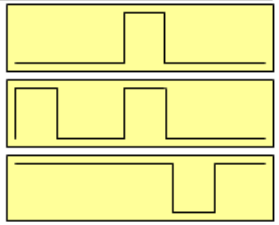
hoge.txt 出力ファイル

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	rownames(data)	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	r
2	gene1	87	79	91	82	90	84	12	21	19	13	17	0.9936
3	gene2	56	122	106	47	84	98	7	44	2	11	18	0.8418
4	gene3	15	28	33	9	27	41	48	46	52	50	49	−0.825

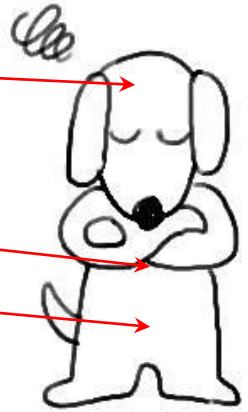
hoge

2. 様々な組織(条件)

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

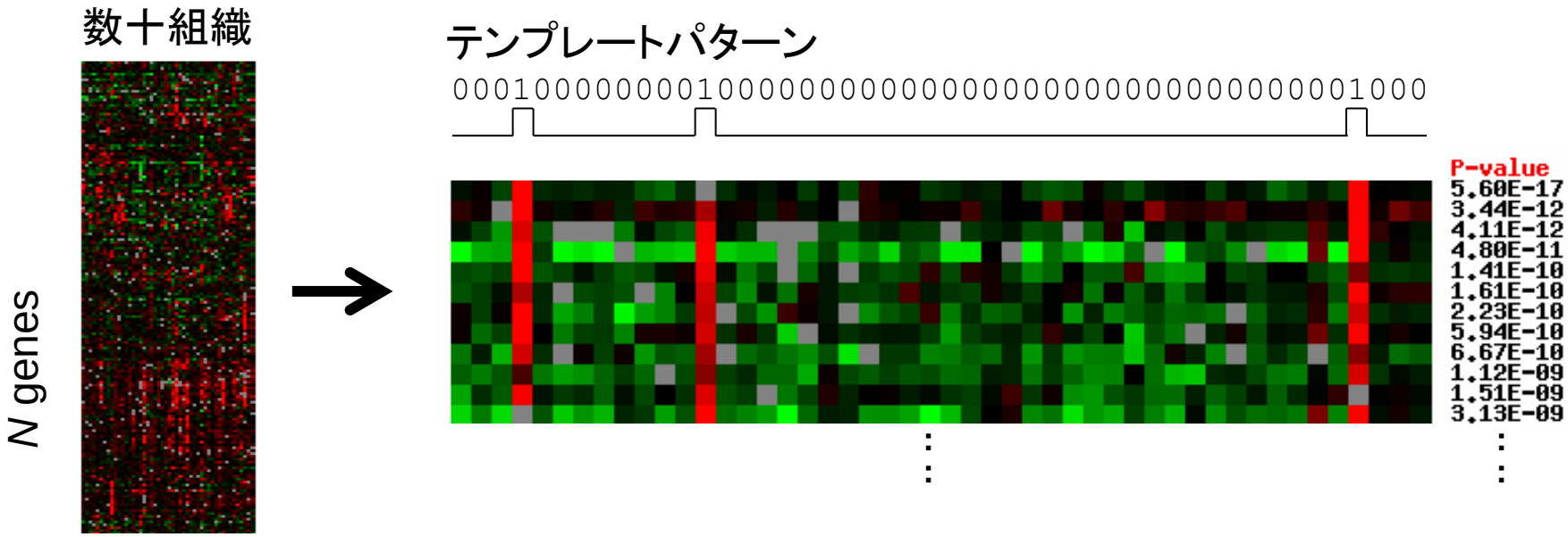


- 組織(条件)特異的(選択的)発現遺伝子
  - 特定の(複数)組織のみで高(or 低)発現している遺伝子
    - 「脳」特異的遺伝子
    - 「心臓」特異的
    - 「腸」特異的
    - …
    - **栄養条件**特異的



2. 様々な組織（条件）

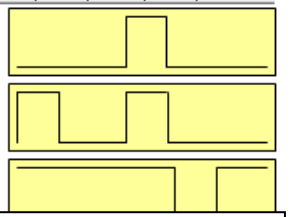
- パターンマッチング法



(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2012/05/18, since 2005)

What's new?
・R2.14.2がリリースされていたのでこれに変更しました。(2012/03/13)NEW

	S1	S2	S3	S4	…
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜について(last modified 2011/6/6)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜ROKU(Kadota_2006) (last modified 2009/07/30)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Sprent's non-parametric method (Ge_2005) (last modified 2009/
- 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)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Schug's Q statistic (Schug_2005) (last modified 2009/07/31)
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜Ueda's AIC-based method (Kadota_2003) (last modified 2009/07
- 解析｜発現変動遺伝子｜組織特異的(選択的)パターン｜パターンマッチング法(テンプレートマッチング法) (last modified 2
- 解析｜発現変動遺伝子｜時系列データ｜Periodic genes｜Lomb-Scargle periodogram (Glynn_2006) (last modified 2006/7/
- 解析｜発現変動遺伝子｜時系列データ｜Periodic genes｜GeneCycle (Ahdesmaki_2005) (last modified 2009/8/3)

・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"                                    #入力ファイル名1(発現データ)を指定
in_f2 <- "sample15_cl.txt"                                 #入力ファイル名2（テンプレート情報）を指定
out_f <- "hoge1.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data.clとの相関係数を計算し

tmp <- cbind(rownames(data), data, r)                      #入力データの右側に相関係数rのベクトルを結合した結果をtmpに格納。
write.table(tmp, out_f, sep="¥t", append=F, quote=F, row.names=F)#tmpの中身をout_fで指定したファイル名で保存。

------    ここまで    ------
```

• 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(rowname
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

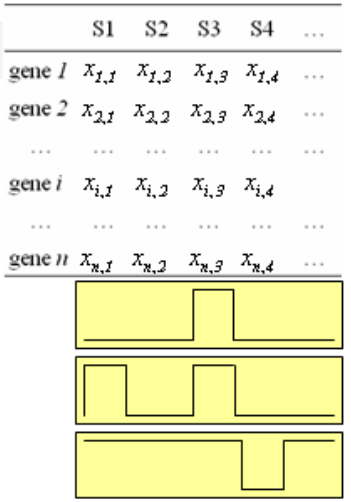
sample15_cl.txt

	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)

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:$0 \leqq H \leqq \log_2 N \rightarrow 0 \leqq H \leqq 3$

i

j	x_{ij}	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	

i

j	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	

i

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

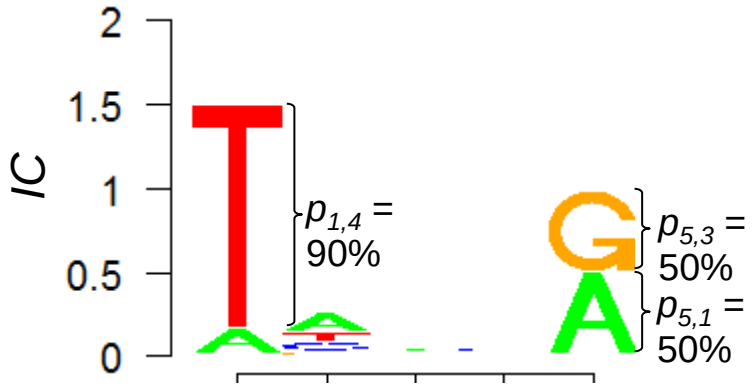
そうでないものは高い値

エントロピー（Sequence logos）

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

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

	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	…



IC	1.53	0.24	0.03	0.03	1.00	…

x_{ij}	*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	

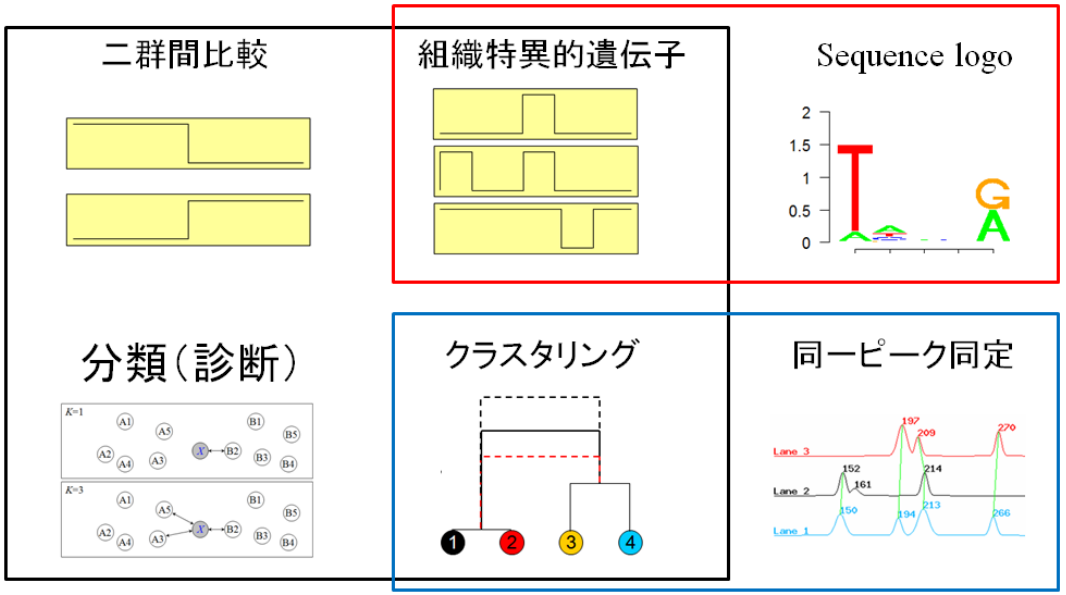
j

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*との**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。