ゲノム間の保存配列の解析

大阪大学微生物病研究所
遺伝情報実験センター
ゲノム情報解析分野

後藤　直久

2005年10月12日

目次

- 自己紹介・研究内容
- 保存配列の解析（１）
  - すべての生物のゲノムに保存されている配列の解析
    - ゲノムデータのダウンロード
    - BioRuby
- 保存配列の解析（２）
  - 転写開始点上流の保存配列の解析
    - sim4, BLAT, Spidey
    - BioRuby
    - モチーフ抽出ソフトウェア

研究内容

- ゲノム配列の配列解析
  - 配列から生命現象の解明を目指す
  - ゲノムから見た生物の進化
  - 多数の生物のゲノム配列を比較
    - 保存されている配列の解析
- 研究に必要なソフトウェアの開発
  - 配列解析ソフトウェアの開発
    - 保存配列検出ソフトウェア「CONSERV」
  - バイオインフォマティクス用ツールの開発
    - BioRuby

現在までに200種以上の生物のゲノム全配列が決定

ゲノム全配列

生物の生命活動に必要なすべての情報が含まれる

解明された事実はまだまだ少ない

ゲノム配列の解析 → 生命現象の解明

多数の種のゲノム全配列を比較解析

単一種のゲノム解析では得られない知見が得られる

→
- 種間で保存されている配列や遺伝子
- 特定の種に固有の配列や遺伝子

すべての生物のゲノムに
保存されている配列は何か？

多数の生物に保存されている配列

- 生命活動に必須の重要な機能
- 生命誕生初期から不変？

すべての生物のゲノムに保存されている配列は何か？

- 材料
  - ゲノムが決定済の全生物のゲノム全配列
    - 細菌約217種, 古細菌約22種, 真核生物約23種
- 方法
  - 配列を単純に比較
    - 適したソフトがなかったので新規開発した
    - BioRubyスクリプトも併用

ゲノム配列データの入手

- 今回の解析は、入手可能な全生物のゲノム配列の端から端まで全部をもれなく使う
- ウェブでブラウズできるだけではダメ
- データを一括ダウンロードできる必要がある
  - ゲノム全配列（塩基配列）
  - アノテーション情報
  - 全タンパク質のアミノ酸配列
- できる限り統合的なデータベースを利用
    - あちこちのサイトを巡るのは面倒
- 利用条件は緩やかなほうが望ましい

原核生物ゲノムのダウンロード

- NCBI (http://www.ncbi.nlm.nih.gov/)
  - 微生物ゲノムの一覧表
    http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
  - ftpによるファイルのダウンロードが可能
    - 2か所に微妙に異なる(大部分は同一)データが存在
  - ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/
    - GenBank（登録者のデータをそのまま掲載）
  - ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
    - RefSeq （NCBIが独自に手を加えたデータベース）

原核生物ゲノムのダウンロード

- NCBIのゲノムデータファイル
  - 種毎（真核生物の一部は染色体毎）に別ディレクトリに格納されている

- *****.fna　ゲノム配列
- *****.faa　タンパク質のアミノ酸配列
- *****.ffn　遺伝子の塩基配列（exonを繋いだもの）
- *****.gbk　GenBank形式のデータ

原核生物ゲノムのダウンロード

- EBI (European Bioinformatics Institute)
  - http://www.ebi.ac.uk/
- EMBL Genomes (http://www.ebi.ac.uk/genomes/)
  - 古細菌(Archaea)ゲノム一覧表
  - http://www.ebi.ac.uk/genomes/archaea.html
  - 細菌(Bacteria)ゲノム一覧表
  - http://www.ebi.ac.uk/genomes/bacteria.html
  - ftpでのデーター括ダウンロードも一応は可能
    - ただし全データがごちゃまぜなので少々ややこしい
  - ftp://ftp.ebi.ac.uk/pub/databases/embl/expanded_con/

原核生物ゲノムのダウンロード

- KEGG
  - http://www.genome.jp/kegg/
  - 統合的なゲノムデータベース
  - 代謝経路の図・データが充実

  - KEGG登録生物一覧表
    http://www.genome.jp/kegg/catalog/org_list.html
    - データのダウンロードが可能
    - ftp://ftp.genome.jp/pub/kegg/genomes/
    - 真核生物も一覧表にあるがゲノム全配列は無い？

その他の原核生物ゲノムデータベース

- GIB (Genome Information Broker)
- http://gib.genes.nig.ac.jp/
  - DNA Databank of Japan (DDBJ) (遺伝学研究所が運営)による微生物ゲノムデータベース
    - http://www.ddbj.nig.ac.jp/
  - GIBのデータの一括ダウンロードはできない(？)
    - しかし、ウェブから閲覧するには便利

その他の原核生物ゲノムデータベース

- Comprehensive Microbial Resources
- http://cmr.tigr.org/
  - The Institute of Genome Research (TIGR) アメリカの研究所) による微生物ゲノムデータベース
    - http://www.tigr.org/
    - データのバッチダウンロードが可能
    - データベースの全データの一括ダウンロードは無理？
    - ftpにはTIGRでシーケンスしたゲノムのデータのみが置いてある

真核生物ゲノム — 概要

- Genomes at the EBI の Eukaryotes が便利
  - http://www.ebi.ac.uk/genomes/eukaryota.html
    - 真核生物ゲノムの一覧表
- NCBI Genomic Biology
  - http://www.ncbi.nlm.nih.gov/Genomes/
  - Entrez Genome
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome
  - Entrez Genome Project
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
    - いずれも全部網羅しているわけではない？
    - 逆に、一部の染色体のみ決定された生物も掲載
    - 配列データをダウンロードするまでに何段階かリンクをたどる必要があるかもしれない

真核生物ゲノム — 酵母・真菌

- NCBI
  - http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi?p3=11:Fungi&taxgroup=11:Fungi|12:
    - この表の status が complete のものについては、データのダウンロードが可能
    - ftp://ftp.ncbi.nih.gov/genomes/Fungi/
      - RefSeq（NCBIの手が入ったデータベース）
    - ftp://ftp.ncbi.nih.gov/GenBank/genomes/Fungi/
      - GenBank（登録者のデータをそのまま掲載）

真核生物ゲノム — 動物

- Ensembl
  - http://www.ensembl.org/
    - 全データのダウンロードが可能
  - ftp://ftp.ensembl.org/
    - ftp://ftp.ensembl.org/pub/data/生物名-リリース番号/
    - 最新版のショートカット: pub/data/current_生物名/
      - FASTA形式 data/fasta/
        - data/fasta/dna ゲノム配列
        - data/fasta/pep タンパク質（アミノ酸配列）
      - GenBank形式 data/flatfiles/genbank/
      - EMBL形式 data/flatfiles/embl/

真核生物ゲノム — 動物

- UCSC Genome Browser
  - http://genome.ucsc.edu/
  - 全データのダウンロードが可能
  - ftp://hgdownload.cse.ucsc.edu/goldenPath/
    - ダウンロードに関するFAQ
    - http://genome.ucsc.edu/FAQ/FAQdownloads

真核生物ゲノム — Arabidopsis

- NCBI
  - ftp://ftp.ncbi.nih.gov:/genomes/Arabidopsis_thaliana/
- TAIR (The Arabidopsis Information Resource)
  - NCBIよりデータが新しい？
  - http://www.arabidopsis.org/
  - ダウンロード ftp://ftp.arabidopsis.org/home/tair/Sequences/

真核生物ゲノム — その他

- Genomes at the EBI
  - 真核生物ゲノムの一覧表
    http://www.ebi.ac.uk/genomes/eukaryota.html
  - Whole Genome Shotgun entries
    http://www.ebi.ac.uk/genomes/wgs.html
- NCBI Genomic Biology
  - http://www.ncbi.nlm.nih.gov/Genomes/
  - Entrez Genome
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome
  - Entrez Genome Project
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj

ソフトウェアの開発

- 多数のゲノム配列から、保存されている配列を検索するソフトウェア
  - BLASTや Clustal W では困難
  - 私の知る限り適したソフトは存在しなかった

新ソフトウェアを開発

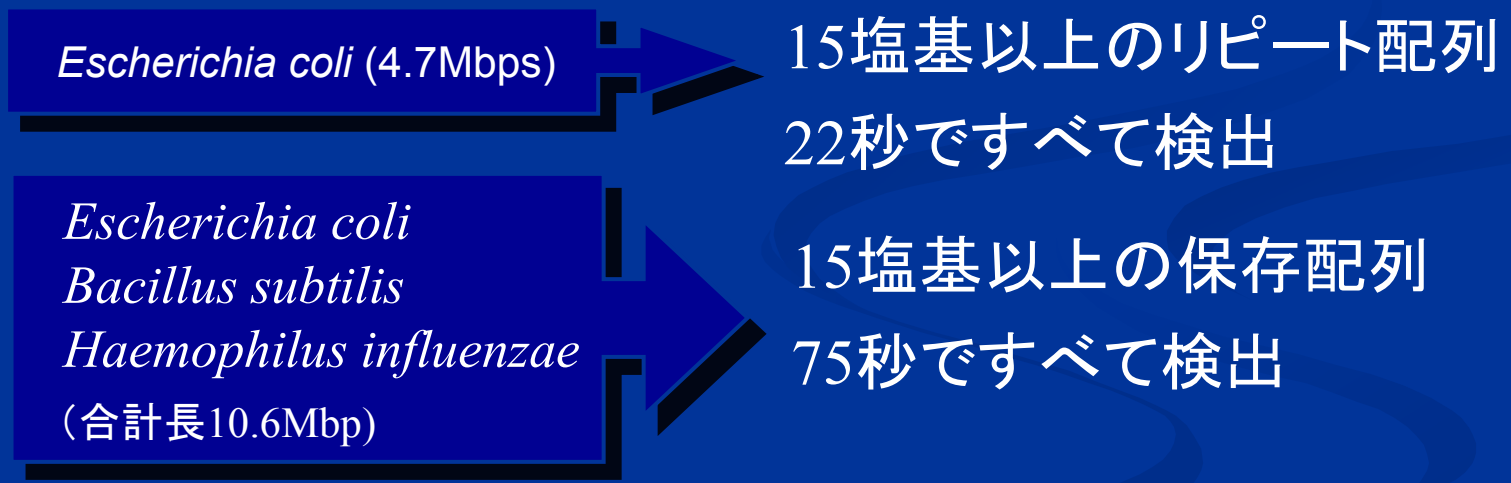
CONSERV

複数のゲノム全配列に保存されている
指定した長さ以上のすべての配列を検出

```
>genome01
…GGCAGGGGCAGGTGGCCACCGAAGTCGTAACAAGGTATCCTCTCTGCCCCCGCCAAAATGATGACCTTG
CTAAAGTTCTTCACCCCCGCACCATTATGTTGGGTTAAGTCCCGCCCCCATCGCCCAGTCCGAAAAATAC
CATCGTATCTAAATGCTAGCTTTCGTCACATTATTTTAATAATCCAACTAGTTGCATCATACAACTACG…
>genome02
…CGCAGTAACAAGCCTTCGCTGTTGGGTTAAGTCCCGTCCGCCCGCCTGACAGATCGCTGCGACCTTGGA
GCGCTCTACCGCTGAGCTACGGCGGCCCTCATCCTTGGGTTTACACTTATTCATCCGAGGGTTTAAGGGT
CCGGCCAGCCTCGCCATAGTCTATATACTAAGTCGTAACAAGGTACGGCCGTTCCCACTCGACACTTCT…
>genome03
…CCAATGATAGCTTTAAGTCGTAACAAGGTACTAATGGGACACTTAAGGCGTACTGTGAAGAATAATCTG
CTTATCTCGGGCTTTGAGAGCAAACCCTCAACAAGACTGGCGGCAACCTCATTTCTGAGAGTGGAGAAGA
TTGCTGTTCAGGATATTTTGTTGGGTTAAACTTTTGTGAATGTTGGGTTAAGTCCCGGTGTCGCGGAAT…
>genome04
…ATAGCAACTTCCAAGTCGTAACAAGGTATCTTGCCGCGTCAGCTGTTGGGTTAAGTCCCGCGATGACTC
CTTCCGCAAGTGATCCACCAGTCGAGTTGATGACCCGGTCATAGGTCTCGACATCATCCCCCCAATCAAC
CAGCTCAAGCGCGGCGTCACCGACGATCATCGGAAGTCGTAACAAGGTACGAGCCGGTGAAAGCCGACG…
```

CONSERVの特長

- 複数のゲノム配列に適用した場合は保存配列を、単一のゲノム配列に適用するとリピート配列を検出
- 完全一致配列のみ検出可能
- 高速な処理



- 複数の染色体を持つ真核生物にも対応
- 塩基配列だけでなくアミノ酸配列にも適用可能

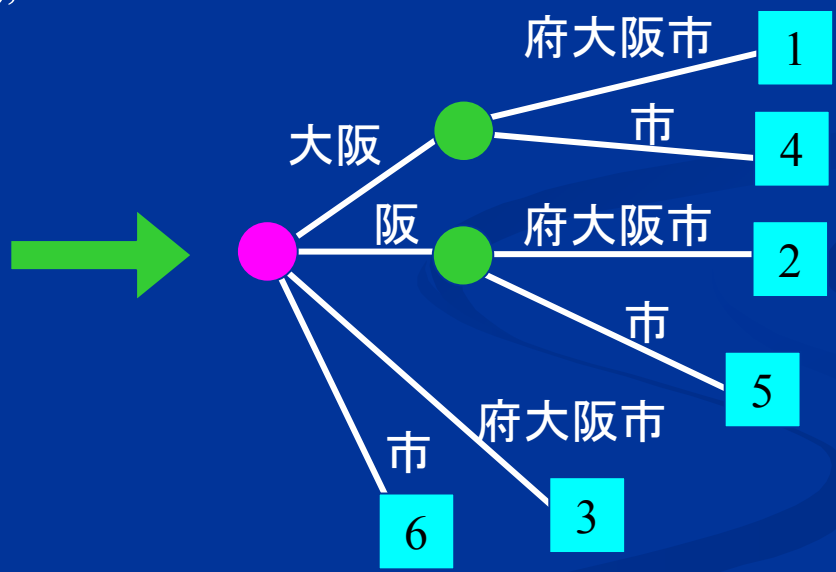
CONSERVの動作原理

Suffix Tree

文字列のすべての Suffix (n文字目から終端までの部分文字列)を全部まとめてツリー状にしたデータ構造

例: “大阪府大阪市”

大阪府大阪市
阪府大阪市
府大阪市
大阪市
阪市
市



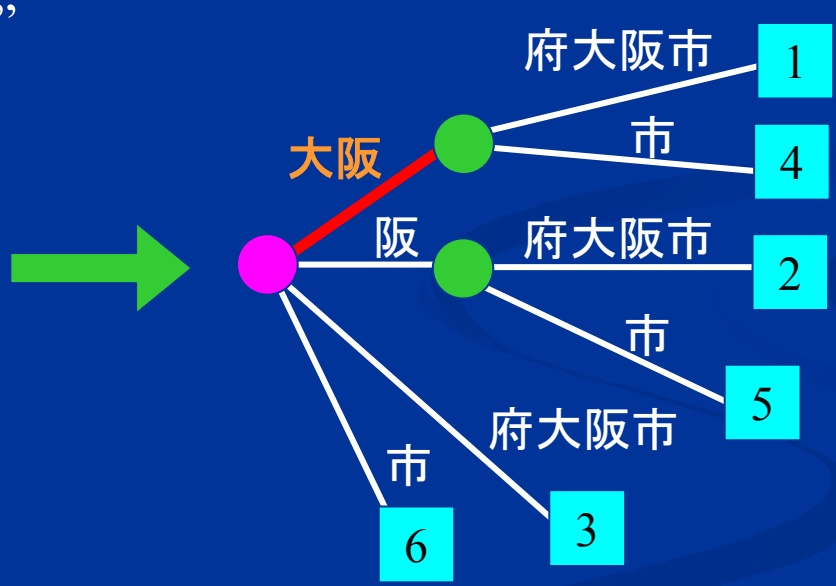
CONSERVの動作原理

Suffix Tree

文字列のすべての Suffix (n文字目から終端までの部分文字列)
を全部まとめてツリー状にしたデータ構造

例: “大阪府大阪市”

大阪府大阪市
阪府大阪市
府大阪市
大阪市
阪市
市

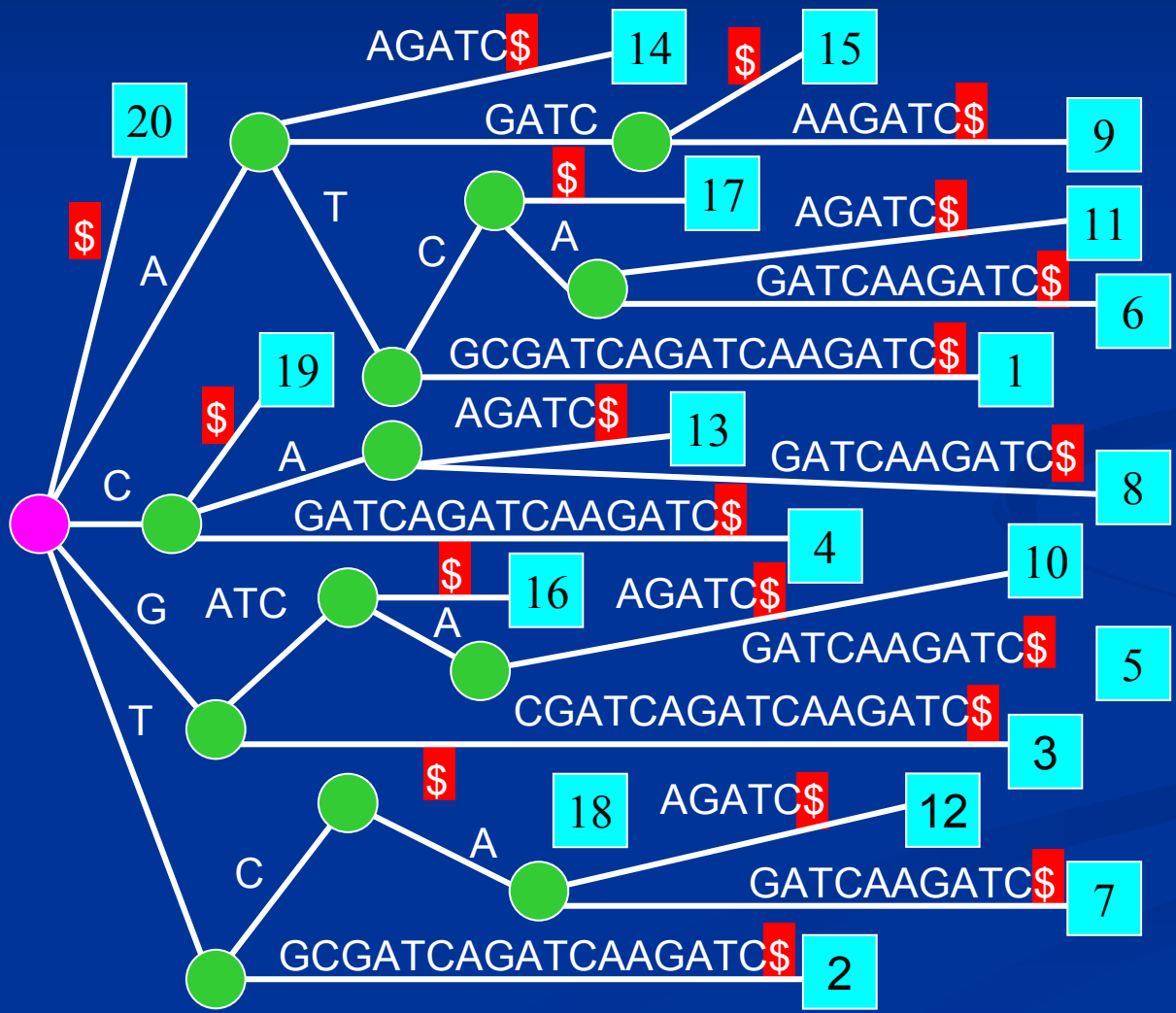


Suffix Tree により効率的なリピート検出が可能

Ukkonen(1995)のアルゴリズム

文字列の長さに比例した計算時間とメモリ使用量でSuffix Treeを構築

例: ATGCGATCAGATCAAGATC$



$ を追加

完成！

CONSERVの欠点

- 完全一致しか検出できない
  - 曖昧さを許すように現在研究中
- メモリを大量に消費する
  - ゲノムサイズの約20～40倍
  - 現在改良中（約10～20倍）
- まだ公開していない
  - 近日公開予定

Complete Genomes used in this Analysis

Bacteria(70)

Corynebacterium glutamicum ATCC 13032
Mycobacterium tuberculosis H37Rv (lab strain)
Mycobacterium tuberculosis CDC1551
Mycobacterium leprae TN
Streptomyces coelicolor A3(2)
Chlamydia trachomatis serovar D
Chlamydia muridarum strain Nigg
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae AR39
Chlamydia pneumoniae J138
Chlorobium tepidum TLS
Synechocystis sp. PCC6803
Nostoc sp. PCC 7120
Deinococcus radiodurans R1
Bacillus subtilis 168
Bacillus halodurans C-125
Listeria innocua CLIP 11262
Listeria monocytogenes EGD-e
Staphylococcus aureus COL
Staphylococcus aureus N315
Staphylococcus aureus Mu50
Staphylococcus aureus MW2
Clostridium perfringens 13
Thermoanaerobacter tengcongensis MB4(T)
Enterococcus faecalis V583
Lactococcus lactis subsp. lactis IL1403
Streptococcus pneumoniae TIGR4
Streptococcus pneumoniae R6
Streptococcus pyogenes MGAS8232
Streptococcus agalactiae 2603V/R
Streptococcus pyogenes SF370 serotype M1
Mycoplasma genitalium G-37
Mycoplasma pneumoniae M129
Ureaplasma urealyticum parvum biovar serovar 3
Mycoplasma pulmonis UAB CTIP
Fusobacterium nucleatum ATCC 25586
Caulobacter crescentus CB15
Brucella suis 1330
Brucella melitensis 16M
Sinorhizobium meliloti 1021
Agrobacterium tumefaciens C58 Cereon
Agrobacterium tumefaciens C58 UWash
Rickettsia prowazekii Madrid E
Rickettsia conorii Malish 7
Neisseria meningitidis MC58
Neisseria meningitidis serogroup A Z2491
Ralstonia solanacearum GMI1000
Campylobacter jejuni NCTC 11168
Helicobacter pylori 26695
Helicobacter pylori J99
Shewanella oneidensis MR-1
Escherichia coli K12-MG1655
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7 VT2-Sakai
Salmonella typhimurium LT2 SGSC1412
Salmonella enterica serovar Typhi CT18
Yersinia pestis CO92
Buchnera sp. APS
Haemophilus influenzae KW20
Pasteurella multocida PM70
Pseudomonas aeruginosa PAO1
Vibrio cholerae El Tor N16961
Xylella fastidiosa 9a5c
Xanthomonas campestris pv. campestris ATCC33913
Xanthomonas axonopodis pv. citri 306
Magnetococcus sp. MC-1
Borrelia burgdorferi B31
Treponema pallidum Nichols
Thermotoga maritima MSB8

Archaea(16)

Aeropyrum pernix K1
Sulfolobus solfataricus P2
Sulfolobus tokodaii strain 7
Pyrobaculum aerophilum IM2
Archaeoglobus fulgidus DSM4304
Halobacterium sp. NRC-1
Methanobacterium thermoautotrophicum delta H
Methanococcus jannaschii DSM2661
Methanosarcina mazei Goe1
Methanosarcina acetivorans C2A
Methanopyrus kandleri AV19
Pyrococcus horikoshii shinkaj OT3
Pyrococcus abyssi GE5
Pyrococcus furiosus DSM 3638
Thermoplasma acidophilum DSM 1728
Thermoplasma volcanium GSS1

Eukarya(2)

Saccharomyces cerevisiae
Schizosaccharomyces pombe

真正細菌70種,古細菌16種, 酵母2種の計88種すべてに存在する配列

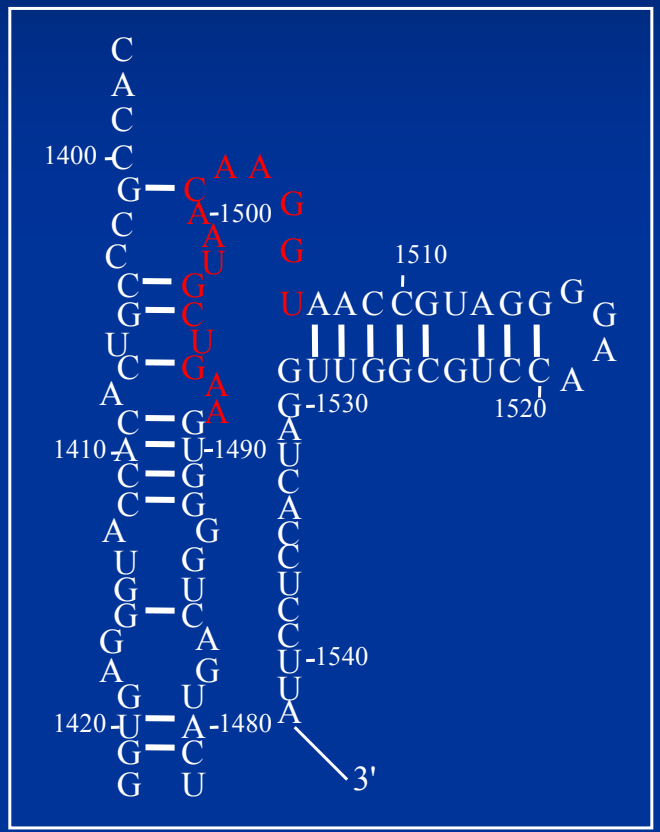
（長さ13塩基以上）*

長さ (bases)	配列	遺伝子	遺伝子内部の位置**
15	**AAGTCGTAACAAGGT**	16S/18S rRNA	1492

* より長い保存配列の一部分となっている配列は記載していない.

** *Escherichia coli*の遺伝子における値. 複数の遺伝子に存在する場合は代表的なものを示した.

AAGTCGTAACAAGGT



- 16S/18S リボソームRNA上に存在
- 1,492塩基め(*Escherichia coli*の値)に存在
- この領域はmRNAのコドンをtRNAのアンチコドンが認識するデコーディング機能に関与
- 16S rRNAのよく保存されている領域のひとつであることは従来知られていた

今回の解析はゲノム全配列が対象

88種のゲノムにおける最長の保存配列

88種のゲノム全配列に共通して存在する最長の配列

AAGTCGTAACAAGGT

この配列が88種以外のゲノムに存在するかを調べた

現在までにゲノム全配列が決定された生物のほぼ全て

真正細菌 217種
古細菌　　20種
真核生物　24種

} のゲノム配列への存在を確認

Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio,
Drosophila melanogaster, Anopheles gambiae,
Caenorhabditis elegans, Plasmodium falciparum, Arabidopsis thaliana,
Saccharomyces cerevisiae, Schizosaccharomyces pombe, ...

BioRubyによる配列の簡易な検索

塩基配列（複数可）に指定した配列が存在するかどうか調べるBioRubyスクリプト

```
#!/usr/bin/env ruby
require 'bio'
pat = Regexp.new(ARGV.shift, true, "n")
Bio::FlatFile.auto($<) do |f|
    f.each do |e|
        e.naseq.scan(pat) do |x|
            pos = $~.offset(0)[0] + 1
            print "#{e.entry_id}¥t#{pos}¥t#{$&}¥n"
        end
    end
end
```

使い方

```
% ruby search02.rb AAGTCGTAACAAGGT file01.fst file02.fst
```

BioRuby

バイオインフォマティクスにおいて

頻繁に使用する機能・あったら便利な機能

- 塩基・アミノ酸配列の処理・解析
- データベースのデータ処理
- 解析ソフトウェアの結果処理
- ファイル入出力・ネットワークとの通信
- …

統一されたインターフェース・使用法

個別に深く理解する必要なく使える

Ruby言語で実装したライブラリ
（ソフトウェア部品集）

http://bioruby.org/

ゲノム間の保存配列の解析(2)
転写開始点付近の保存配列の解析

- 目的:発現制御に関与する配列の候補探索
  - 同一発現パターンを示す遺伝子の転写開始点付近(主に上流数百～数千bp)に保存されている配列の探索
- 方法
  - (0) mRNA, cDNA, EST などを収集
  - (1) ゲノムに貼り付ける
  - (2) ゲノムから上流配列を切り出す
  - (3) 保存配列を見つける

mRNAのゲノムマッピング

- BLAST (http://www.ncbi.nlm.nih.gov/blast/)
  - GT-AG を考慮しないので下記のソフトを使うほうがよい
- sim4 (http://globin.cse.psu.edu/html/docs/sim4.html)
  - 定番
  - 後継の SIBsim4 (http://sibsim4.sourceforge.net/) 開発中
- BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat)
  - 速い
  - ソースのダウンロード http://www.soe.ucsc.edu/~kent/src/
- Spidey (http://www.ncbi.nlm.nih.gov/spidey/)
  - NCBI謹製
- exonerate (http://www.ebi.ac.uk/~guy/exonerate/)
  - Ensemblで採用

BLAST結果処理の実行速度比較

	所要時間(s)	S.D.	速度(MB/s)	速度比
Bio**Ruby** (Ruby1.8.0)	35.325	0.032	2.83	21.3
BioPerl (Perl5.6.1)	751.067	2.915	0.133	1

BioRubyはBioPerlの**20倍**速い！

sim4, BLAT, Spidey の使い方

- sim4

  % sim4 クエリー配列 ゲノム配列 > 出力ファイル

- BLAT

  % blat ゲノム配列 クエリー配列 出力ファイル

- Spidey

  % spidey -i ゲノム配列 -m クエリー配列 -o 出力ファイル

※ゲノム配列、クエリー配列はそれぞれ単一のFASTA形式の配列を格納したファイル
（マルチFASTA形式への対応状況はソフトによって異なる）

BioRubyで出力ファイルを処理する例

```
#!/usr/bin/env ruby
require 'bio'

ARGV.each do |fn|
  Bio::FlatFile.auto(fn) do |ff|
    ff.each do |entry|
      prog = entry.class.to_s.sub(/¥ABio¥:¥:/,
'').sub(/(¥:¥:Default)?¥:¥:Report.*/, '')
      entry.each do |hit|
        hit.each do |hsp|
          print [ prog, entry.query_def.split[0],
            hit.target_def.split[0],
            hsp.query_from, hsp.query_to,
            hsp.hit_from, hsp.hit_to ].join("¥t"), "¥n"
        end
        break
      end
    end
  end
end
```

BioRubyで出力ファイルを処理する例

- 使用方法
  - % ruby sample_mapping.rb file…
  - エクソン毎に以下の情報をタブ区切りで出力
    - 使用したソフトの名称
    - クエリー配列（cDNAなどの配列）の説明
    - ゲノム配列の説明
    - クエリー配列上のアライメント開始位置
    - クエリー配列上のアライメント終了位置
    - ゲノム配列上のアライメント開始位置
    - ゲノム配列上のアライメント終了位置
  - 使用上の注意点
    - Spideyの結果処理にはCVS先端が必要（バグがあった）
    - ゲノム - cDNAが逆方向の鎖の場合は、ソフトによって数字の扱いが異なるため要注意

BioRubyのいいところ

- 入力ファイル形式は自動判別可能
  - Bio::FlatFileクラスの機能
  - いちいち指定しなくていいので楽
  - ファイル形式を覚えなくても大丈夫
- 複数ファイル形式に対応するスクリプトをわりと簡単に書ける
  - オブジェクト指向のおかげ
  - 先ほどのサンプルの場合は4つのソフトの出力に対応

    sim4, BLAT, Spidey, BLAST

ゲノム間の保存配列の解析（2）
転写開始点付近の保存配列の解析

- 目的：発現制御に関与する配列の候補探索
  - 同一発現パターンを示す遺伝子の転写開始点付近（主に上流数百～数千bp）に保存されている配列の探索
- 方法
  - （0）mRNA, cDNA, EST などを収集
  - （1）ゲノムに貼り付ける
  - （2）ゲノムから上流配列を切り出す
  - （3）保存配列を見つける

転写開始点上流の配列の切り出し

例:「ファイル名　転写産物名　鎖の方向(+/-)　開始点の座標」というタブ区切りのファイルを元に、上流 XXX bp の配列を切り出すスクリプト

```
#!/usr/bin/env ruby
require 'bio'
len = ARGV.shift.to_i
prev_fn = nil; prev_seq = nil
$<.each do |x|
    fn, name, strand, pos = x.split(/¥t/)
    pos = pos.to_i
    next unless fn
    if prev_fn == fn then
        seq = prev_seq
    else
        seq = Bio::FlatFile.auto(fn) { |ff| ff.next_entry.naseq }
    end
    if strand == '-' then
        s = seq.splicing("complement(#{pos+1}..#{pos+len})")
    else
        s = seq.splicing("#{pos-len}..#{pos-1}")
    end
    puts s.to_fasta("upstream_#{name}", 70)
    prev_fn = fn; prev_seq = seq
end
```

転写開始点上流の配列の切り出し

- サンプルの使用方法

  たとえば上流1500塩基を切り出す場合

  - % ruby sample_splicing.rb 1500 test.tsv
  - サンプルの工夫している点
    - 毎回ファイルをオープン→配列を読み出し、を繰り返すと非常に遅いので、直前の行と同じファイル名だったら、配列を使いまわすようにした。
    - それでも、何千配列も連続して出力させるとメモリ不足でエラーになることがある。本格的にやるなら、さらに工夫をするか、DASサーバを立てたほうがよい。

ゲノム間の保存配列の解析(2)
転写開始点付近の保存配列の解析

- 目的:発現制御に関与する配列の候補探索
  - 同一発現パターンを示す遺伝子の転写開始点付近(主に上流数百～数千bp)に保存されている配列の探索
- 方法
  - (0) mRNA, cDNA, EST などを収集
  - (1) ゲノムに貼り付ける
  - (2) ゲノムから上流配列を切り出す
  - (3) 保存配列を見つける

パターン・モチーフ抽出ソフト

- 多数のソフトウェアが存在
  - CONSENSUS
    (ftp://ftp.genetics.wustl.edu/pub/stormo/Consensus)
  - MEME (http://meme.sdsc.edu/)
- 複数のソフトに同時にデータを投げるツール
  - MELINA (http://melina.hgc.jp/)
    - ウェブ上のサービス
  - BEST (http://webster.cs.uga.edu/~che/BEST/)
    - Linux/UNIX用ソフトウェア

最後に主張したいこと

BioRubyはとても便利。
どんどん使おう！