

ゲノム情報利用ワークショップ2005 10/11-12
実戦系バイオインフォマティクス@かずさ

Noncoding配列の探索と 比較ゲノム

片山俊明 <k@bioruby.org>

東大医科研ヒトゲノム解析センター
ゲノムデータベース分野



ヒトゲノムの9割はゴミ



というのは

C+O+A+8

ガゼ

「ヒトゲノムの9割はゴミ」
というのは ガセ

うそつき

大人の事情により
内容を一部割愛

FANTOM3

- The Transcriptional Landscape of the Mammalian Genome
- Science 309:1559-1563 (2 Sep 2005)

プレスリリースより:

哺乳動物のトランスクリプトームの総合的解析による「RNA新大陸」の発見

- タンパクを作り出さないRNA(非タンパクコードRNA)について、予想をはるかに超える23,000個以上を発見した。
 - 二重鎖DNAの両方の鎖が転写されたRNA(センス／アンチセンスのRNAペア)について考えられてきた数をはるかにしのぐ31,422個を発見。
- 「ヒトゲノムの7割くらいは転写されているっぽい」

The background of the slide is an abstract, artistic composition. It features a series of concentric, wavy lines in shades of red and pink, radiating from a central point. These lines are interspersed with small, bright pink dots, creating a sense of depth and movement. The overall effect is reminiscent of a microscopic view of a biological structure or a stylized representation of a flower or star.

Noncoding重要



話は変わりますが
リピートも重要

2次構造つながり

- Noncoding RNA とかは2次構造をとって機能するらしい
- リpeatにも色々あって中には2次構造を取りそうなものも

The background of the slide is an abstract composition of vibrant red and pink wavy lines that create a sense of depth and movement. These lines are interspersed with numerous small, bright pink dots, giving the overall effect a textured, almost digital or particle-like appearance. The colors are rich and saturated, with some areas appearing darker due to the way the lines and dots overlap.

探索方法は流用できそう

リピートの分類

- Tandemly repeated DNA - 反復型リピート
 - Simple sequence repetitions - マイクロサテライト, ミニサテライトなど
 - Blocks of tandemly repeated segments - セントロメアなど長めのサテライト
- Dispersed repetitive DNA - 分散型リピート
 - Segmental duplications - ゲノム内の配列重複など
 - Transposable elements
 - RNA intermediate (Class I) - コピー&ペースト
 - Retrotransposon - LTRを持つ、+env でレトロウィルスの仲間など
 - Retroposon - LTRを持たない、LINE, SINE など
 - DNA intermediate (Class II) - カット&ペースト
 - IS (insertion sequence) - transposase (integrase) を持つ
 - Transposon - IS + さらにいろいろ持つ
 - 不明 → 萌え要素
 - MITE (miniature inverted repeat trasposable element)

分類不明 == 萌え(たぶん)

- 動物？植物？
 - ミドリムシ
- 菌類？動物？
 - 粘菌 (南方熊楠 - ChemRuby)
- 節足動物？環形動物？
 - 緩歩動物門 (クマムシ - BioRuby)

大人の事情により内容を一部割愛

クマムシとは

- 苔などに棲んでいて、ヨチヨチ歩きます
- 東大の駐車場から南極までどこにでもいます
- 乾くと tun 状態に変形します→水に戻すと蘇生

変形すると

- 耐乾燥: 何年間も乾眠します
- 耐温度: マイナス272℃～プラス151℃の環境に耐えます
- 耐X線: 57万レントゲンでもOK (ヒトの致死線量は500)
- 耐圧力: 微生物も死に絶える3000気圧の倍6000気圧OK
- 耐真空: 真空中に晒してもOK

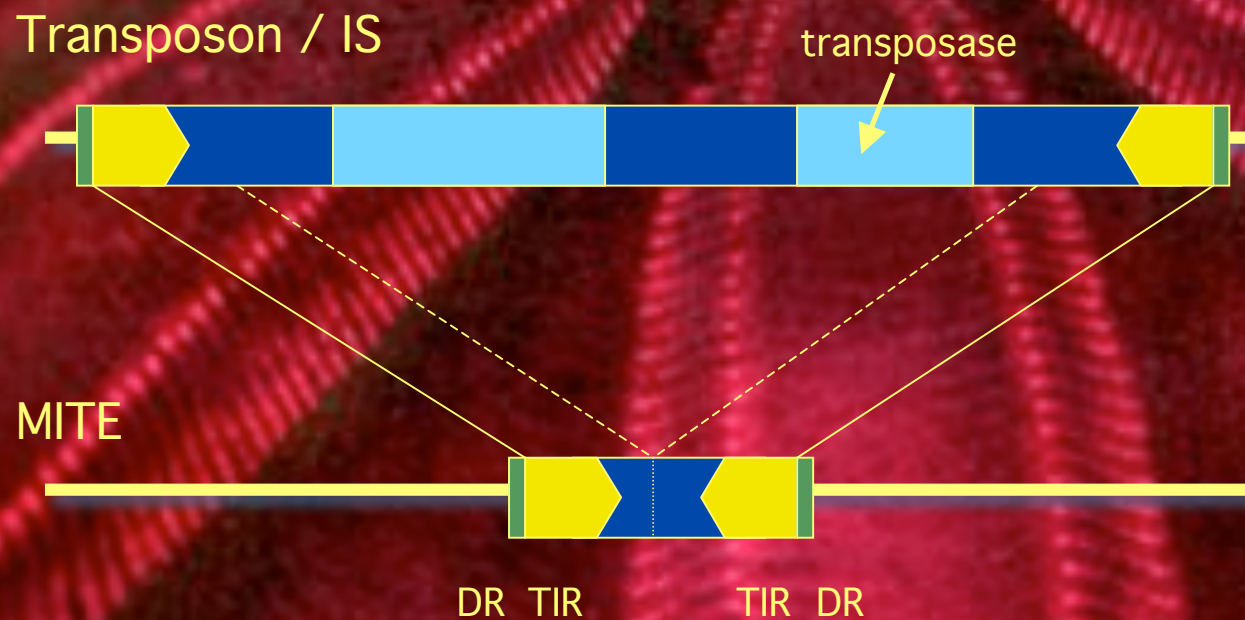
詳しくはコチラ → <http://kumamushi.net/>



それはさておき

MITEとは

- Miniature inverted transposable element





こんなものが
ゲノムの
あちこちにある



みつけた

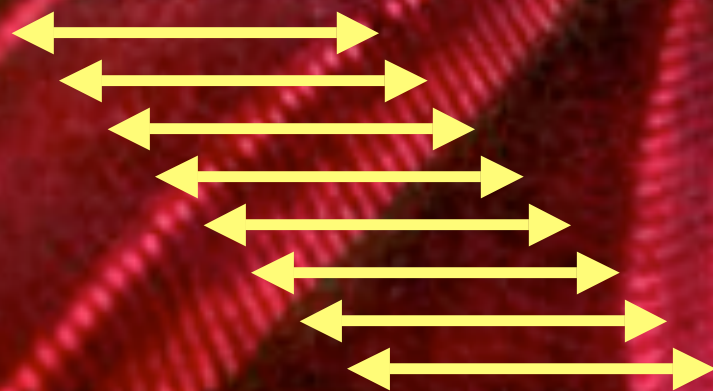
そのまえに

- ゲノム配列はどれくらいランダムか
 - サルはシェークスピアを書けるか
 - サルがランダムに ATGC をずらずら並べてもゲノムっぽくなるのか

ウィンドウサーチ

ウィンドウサイズ (例: 8bp) ごとにスライドしながらナニカをするコト

ゲノム CCTAGGCGAACCTTTAGCAGTAGCGACAAAAGCTA



CCTAGGCG 2回

CTAGGCGA 5回

TAGGCGAA 4回

... ランダムならゲノム中の出現頻度は
ほぼ一定のはず、、だが？

BioRubyでは

```
#!/usr/bin/env ruby
```

```
require 'bio'
```

```
count = Hash.new(0)
```

```
genome = Bio::Sequence::NA.new("
  CCTAGGCGAACCTTTAGCAGTAGCGACAAAAGCTA
")
```

```
genome.window_search(8) do |subseq|
  count[subseq] += 1
end
```

```
sorted = count.sort_by { |subseq, num| num }
```

```
sorted.each do |subseq, num|
  puts "#{subseq}¥t#{num}"
end
```



caagcccc	1
ttccaaac	1
:	
agcgatcg	19
cgatcgct	20
ggcgatcg	21
cgatcgcc	27
gcgatcgc	51

実際は偏りがある

ていうかBioRubyとは

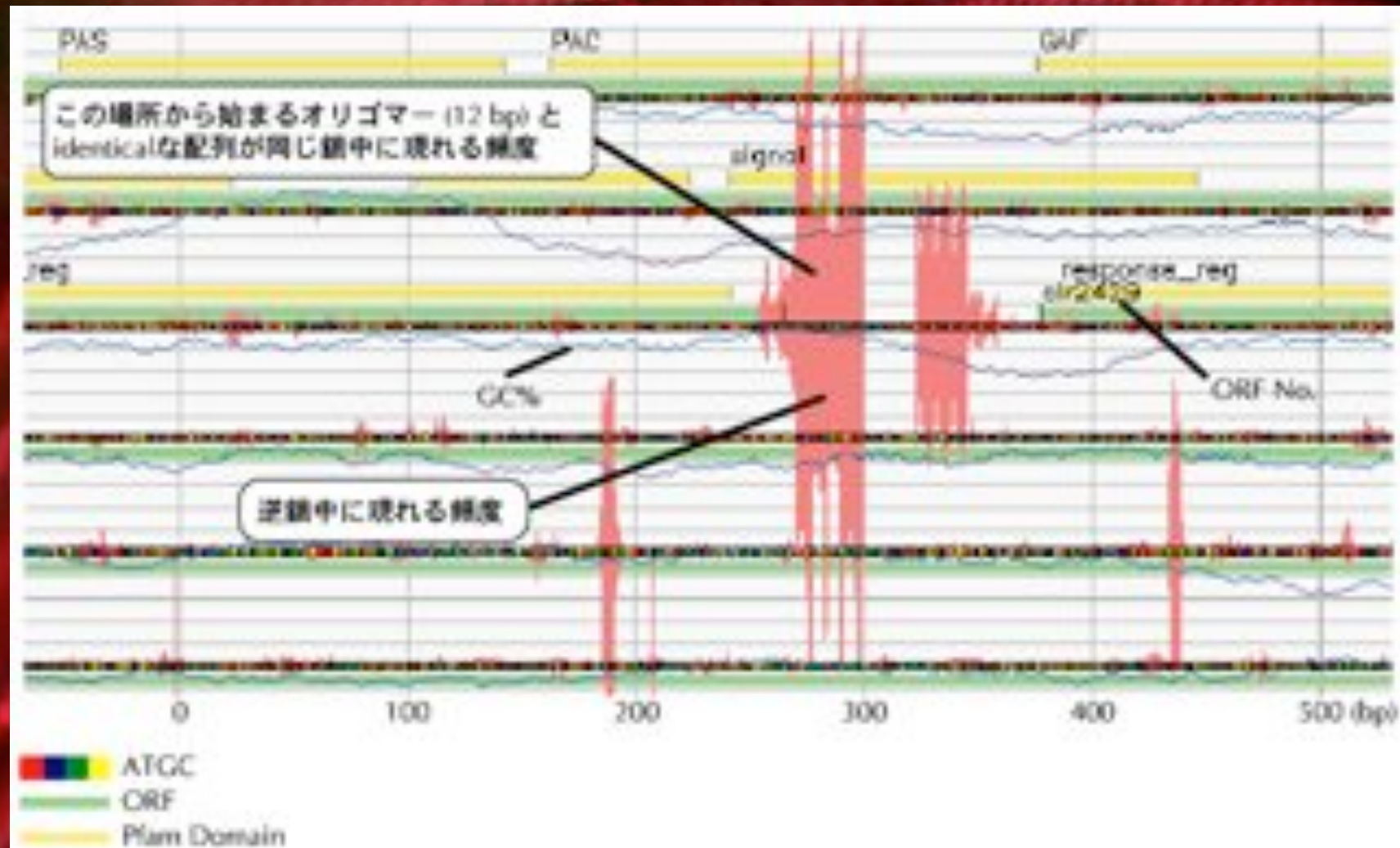
- Ruby言語
 - オブジェクト指向
 - スクリプト言語
 - 日本発(まつもとゆきひろさん作)の言語
 - Ruby on Rails のおかげで？世界中でブレイク中
 - Mac OS X にも最初から入ってます
- BioRubyは、それ用のバイオインフォライブラリ
 - 配列解析とか KEGG いじりとか
 - 今年は ChemRuby と共にIPA未踏ソフトに採択
 - そして、今日はコア開発者3人がココに来てます。。



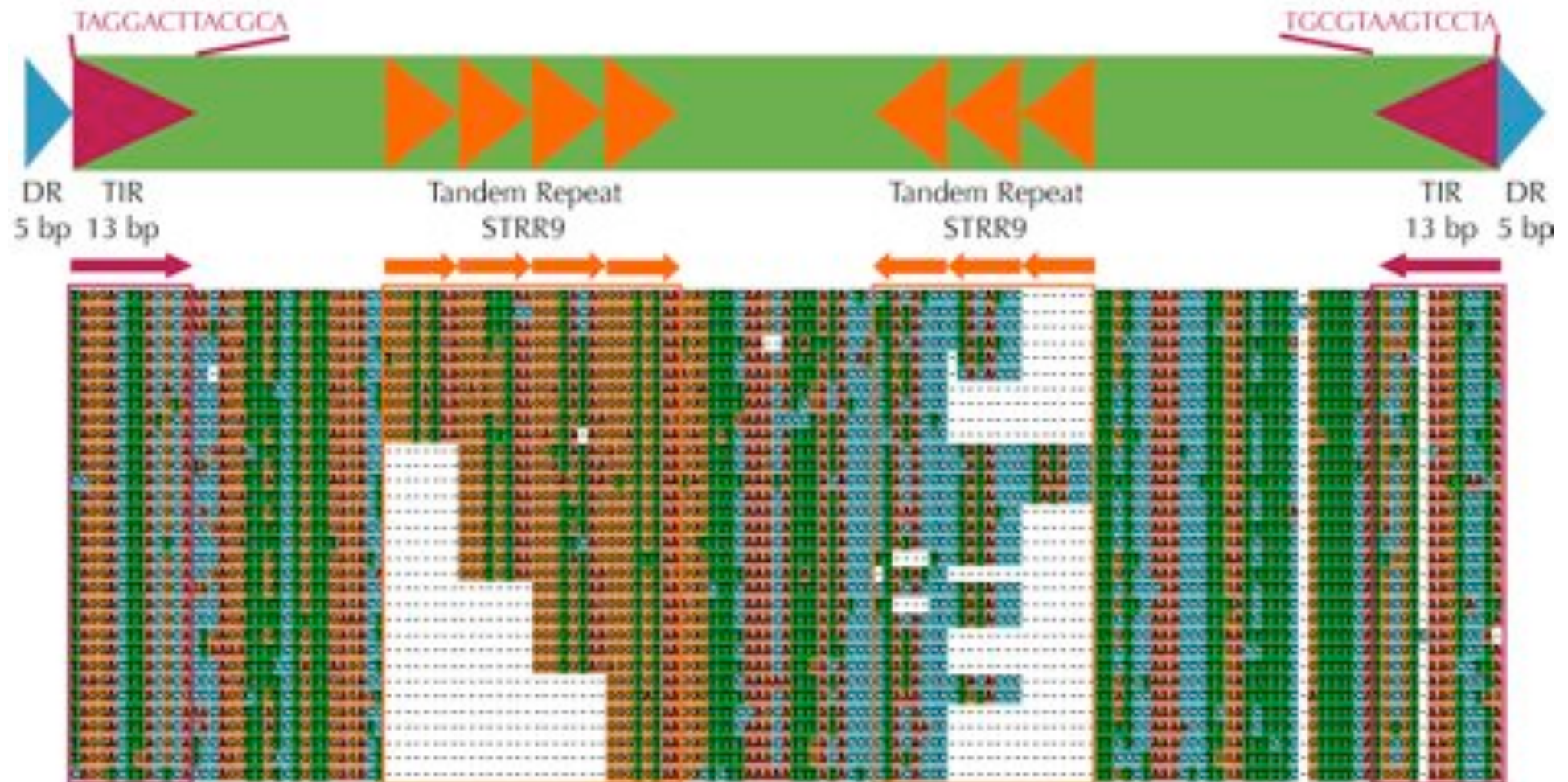


それはさておき

藍藻 *Anabaena* sp. PCC7120 の インタージェニックに高頻度な配列？

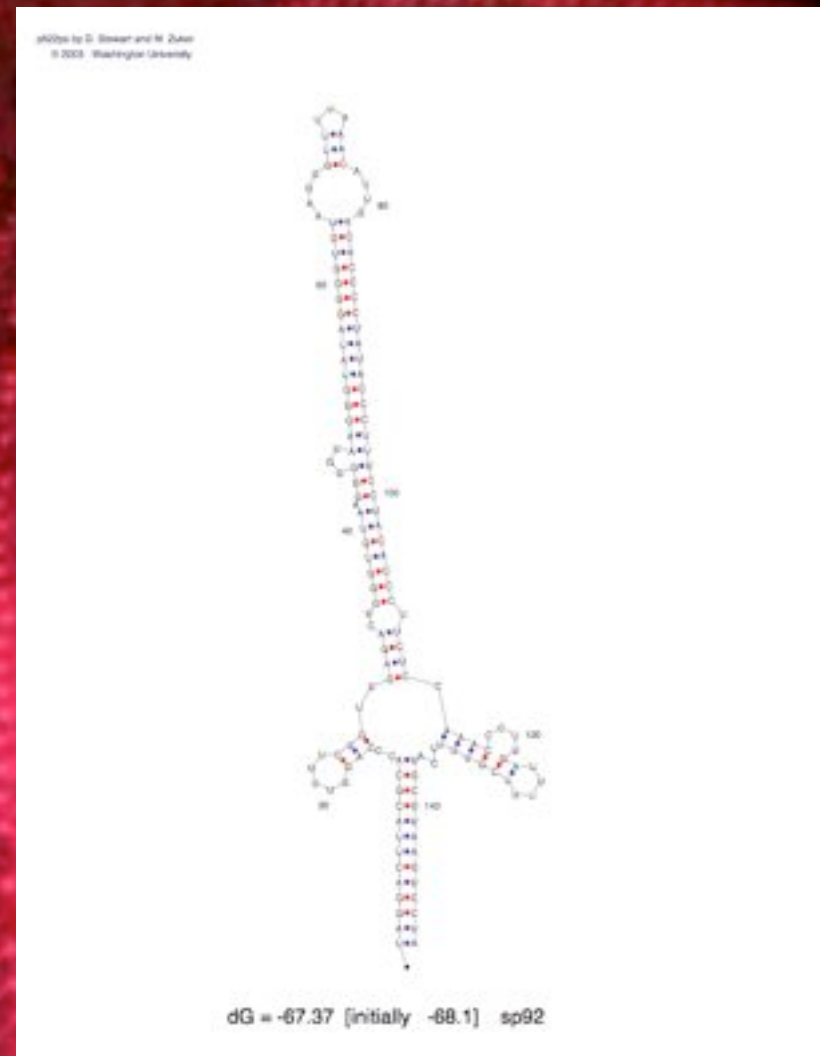
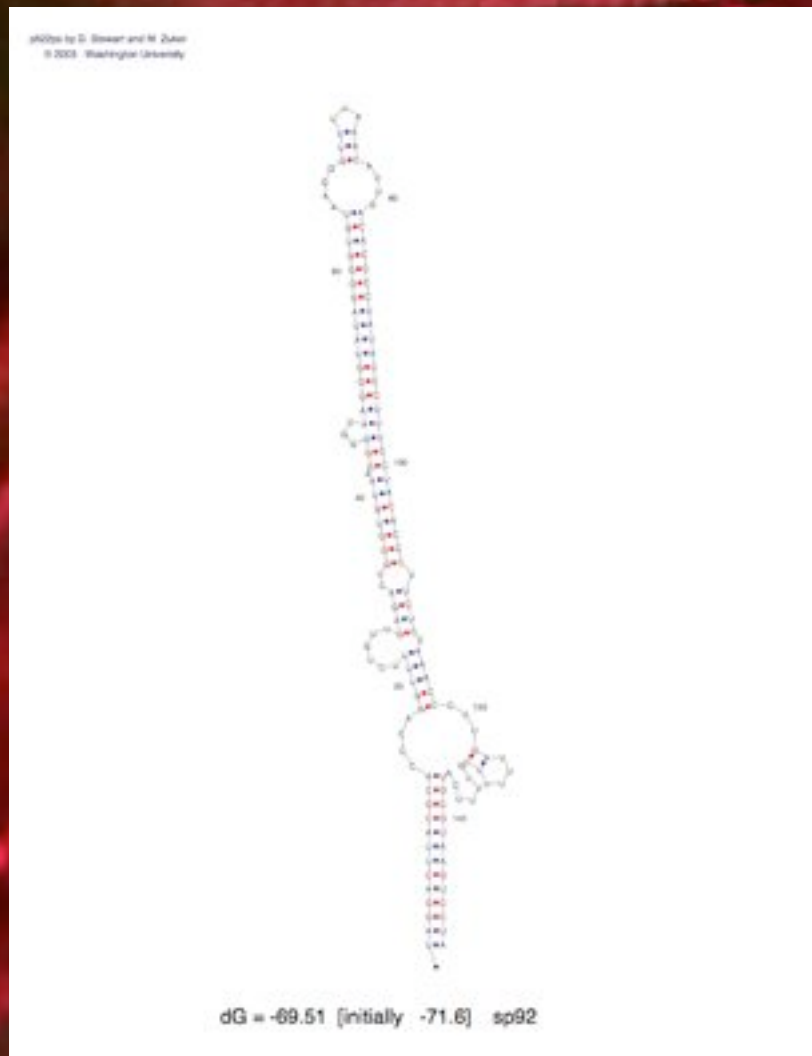


BLASTすると数十コピーの保存配列



DR: Direct repeat, TIR: Terminal inverted repeat

2次構造もありそうだ



手アラインメント

- Stockholm フォーマットで2次構造を表現
- Emacs (ralee-mode.el) を利用
 - <http://www.sanger.ac.uk/Users/sgj/ralee/>

```
;;; ~/.emacs の設定例
```

```
;; Emacs lisp は ~/lib/lisp/ 以下にディレクトリを掘って置くと整理できる
```

```
(let ((default-directory "~/lib/lisp"))  
  (normal-top-level-add-subdirs-to-load-path))
```

```
;; ~/lib/lisp/ralee/ 以下に ralee-*.el をインストール  
;; .stk ファイルを開いたら ralee-mode が起動するように設定
```

```
(autoload 'ralee-mode "ralee-mode" "Yay! RNA things" t)  
(setq auto-mode-alist (cons '("\\.stk$" . ralee-mode) auto-mode-alist))
```

File Edit Colour Structure Buffers Help

Emacs (ralee-mode.el) 作者のページよりコピペ)

Emacs (rablee-mode.el)

(作者のページよりコピペ)

[illegible]

Stockholm フォーマット

```
# STOCKHOLM 1.0
```

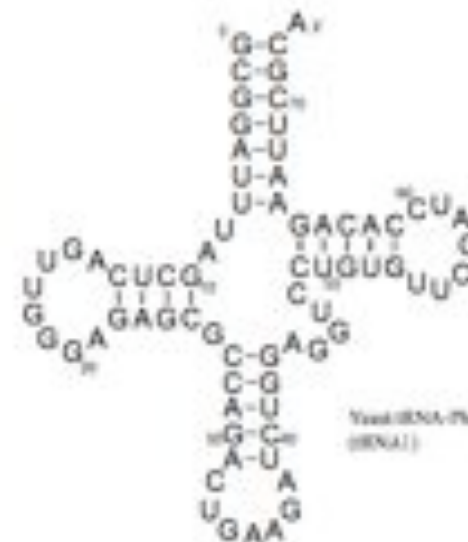
```
seq1      catgcgaaacgtcaagctgggcatc
seq2      cattcgaaagggtcatggtgcgcaat
seq3      catgggaaaccacacagtggccatt
#=GC SS_cons  .<<<<...>>.<<...>>..>>>.
//
```

もう少しちゃんとした例？

tRNA1 GCGGAUUUAGCUCAGUUGGG . AGAGCGCCAGACUGAAGAUCUGGAGGUCC
tRNA2 UCCGAUAUAGUGUAA C . GGCUAUCACAUCACGCUUUCACCGUGGAGA . CC
tRNA3 UCCGUGAUAGUUFUAAU . GGUCAGAAUGGGCGCUCUGUCGCGUGCCAGA . UC
tRNA4 GCUCGUAUGGCCGCAGU . GGU . AGCGCAGCAGAUUGCAAUUCUGUUGGUCC
tRNA5 GGGCACAUUGGCCGCAGUUGGU . AGCGCGCUUCCCUGCAAGGAAGAGGUCA
#=GC SS_cons <<<<<< , <<<< , , , , , , , , , >>>> , <<<<< , , , , , , , , , >>>> , , , , , <

tRNA1 UGUGUUCGAUCCACAGAAUUCGCA
tRNA2 GGGGUUCGACUCCCCGUAUCCGAG
tRNA3 GGGGUUCAAUUCCCCGUCGCGGAG
tRNA4 UUAGUUCGAUCCUGAGUGCGAGCU
tRNA5 UCGGUUCGAUUCGUGUUGCGUCCA
#=GC SS_cons <<<< , , , , , , , , , >>>>>>>>>>>> ,
//

Yeast tRNA-Phe
(tRNA1)



(INFERNAL の Userguide.pdf より)

欠点? : pseudoknot は扱えないらしい

ralee mode のキーバインド

C-f, C-b, C-n, C-p

いつもより速い移動(20文字,20行単位)

1文字ずつの移動は矢印キーで

C-c C-p, C-c C-o

対応するペアにジャンプ(C-c C-oの方は画面分割も)

. (ピリオド), C-d

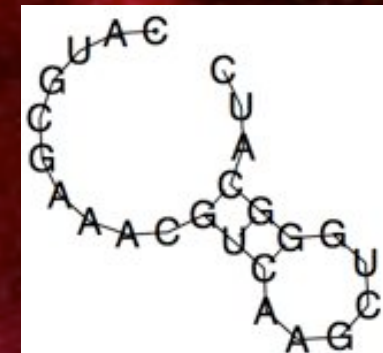
配列ごとにギャップの挿入、削除

C-c C-i, C-c C-d

全配列に対してギャップの挿入、削除

C-c C-f

RNAfold で画像ファイル rna.ps を出力



INFERNAL

- できあがったアラインメントから CM ファイル作成
- ゲノム配列に対してサーチ
- Rfam の作成や検索にも使われている
 - <http://www.genetics.wustl.edu/eddy/infernal/>

INFERNAL cmbuild

% cmbuild mite.cm mite.stk

- Stockholm フォーマットの mite.stk から
CM ファイル mite.cm が生成される

INFERNAL cmbuild

% cmsearch mite.cm genome.fa

- 配列に対して、2次構造を考慮した検索

INFERNALの良い点

簡単

INFERNALの悪い点

遅い、

make による大規模解析 (バカパラ)

- 遅いのでスパコンで並列計算させる
 - http://www.sanger.ac.uk/Software/Rfam/help/scripts/search/rfam_scan.pl
 - rfam_scan は配列に対し Rfam を効率的に検索するための Perl スクリプト
- ヒットしそうな Rfam をあらかじめ BLAST で絞り込む

手順1: shotgun.rb

- KEGG の全生物種のゲノムを 1000bp オーバーラップしつつ 11000bp 毎に切る

```
#!/usr/bin/env ruby
require 'bio'
Bio::FlatFile.auto(ARGV) do |ff|
  ff.each do |entry|
    name, seq = entry.entry_id, entry.naseq
    begin
      i = 0
      seq.window_search(11000, 10000) do |subseq|
        puts subseq.to_fasta("#{name}:segment_#{i*10000+1} (11000bp)", 60)
        i += 1
      end
      puts seq.subseq(i*10000+1).to_fasta("#{name}:segment_#{i*10000+1}", 60)
    rescue
      puts seq.to_fasta("#{name}:segment_all", 60)
    end
  end
end
```

手順2: Makefile

- 適当なエントリ数毎にファイルに分割して
% `make -j 200`

```
# 検索対象の配列ファイルをリストアップ
QUERIES := $(wildcard query/*/*)

# 配列ファイルから結果ファイル生成するルール
result/%: query/%.fa
    bin/rfam_scan.pl -d ./rfam -f tab $< > $@
    echo $@ >> finished.txt

# 作りたい結果ファイルのファイル名一覧を生成
infernal: ${QUERIES:query/%.fa=result/%}
```

結果

7月末から始めて 266 生物種中

現在まだ 137 生物種め
orz...

もし終わったら KEGG DAS に載せます

<http://das.hgc.jp/>

しかも

まさに明日、スパコンが止まります。。。
(計算中なのにー)

でも make だから大丈夫。スパコン再起動後に

`% make -j 200`

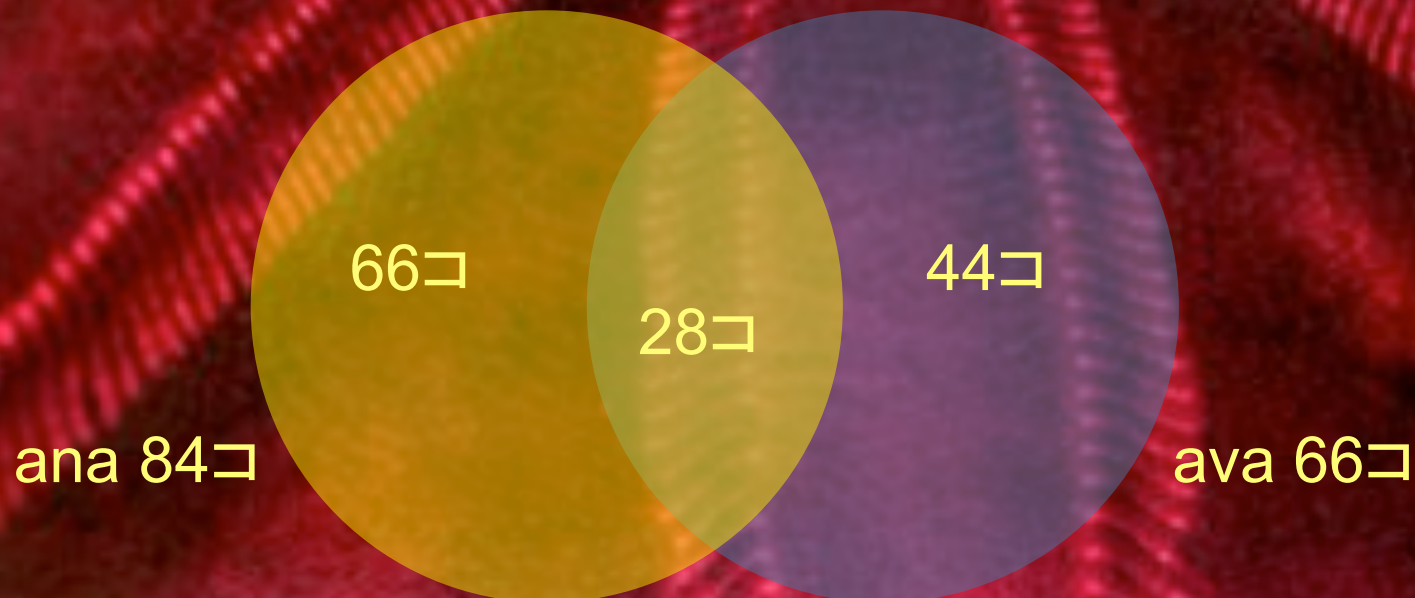
とするだけで続きをやってくれる、、はず



さて話は MITE に戻って

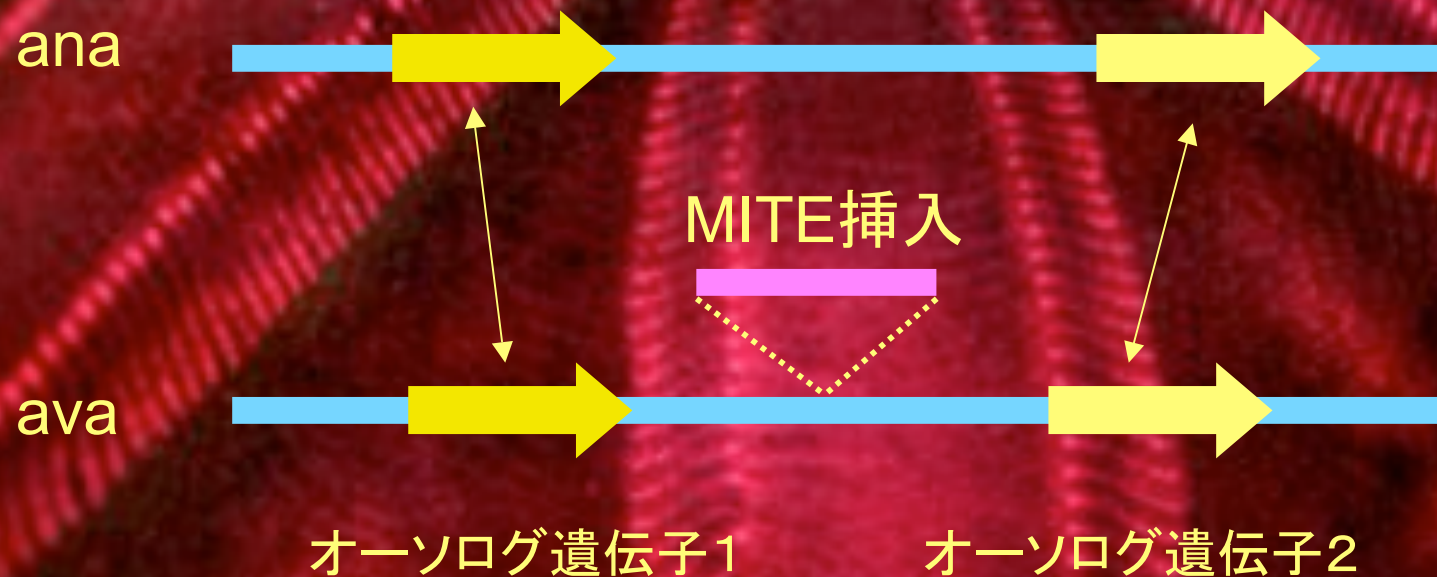
INFERNAL で釣れた MITE

- BLAST, HMMER よりも高感度、高精度
 - 取りこぼしがなく、崩れかけた端も拾えている
 - 近縁の藍藻2種で検索してみた → 計150コ



シンテニー領域の比較ゲ

- MITE 挿入部位の前後 500bp を種間比較



EMBOSS でアライメント

- なぜに EMBOSS で？
- グローバルアライメントがしたい
 - needle
 - stretcher (needle でメモリが足りない場合？)
- ローカルアライメント
 - water (blast, ssearch でいいとおもいます、、)

needle は大文字小文字を保存

あらかじめ興味のある領域を大文字にしておくで見やすい

グローバルアラインメントなので両端を合わせようとしてくれる

197313..19841	451	gtcaagttggtgagtcagcaatctgatcaatttcaccataatgccaatat	500
		. . .	
1250971..1251	451	gtcaagatggtgagtcagcaatctgatcaatttcaccacaatgtcaat--	498
197313..19841	501	TAGGACTTACGCATCCAGGTTGTCTGTTAAGACTGGGTGTAAGGGGGTAA	550
1250971..1251	499	-----	498
197313..19841	551	GGGTATAGCCCTGTACCCCTACACCCTTCTCCAAACCCTTGATTTTTCGT	600
1250971..1251	499	-----	498
197313..19841	601	TTTCATGCGTAAGTCCTAaatataagtaaggatgtcaacaagatgcgag	650
			
1250971..1251	499	-----gaaggatgtcaacaagatggatg	522
197313..19841	651	gttcttgggaaggcatgattcaatcttgctcaatctt-----	687
		. .	
1250971..1251	523	gttcttgggaaggatgattcaatggtgctcaatcttgggtgtctttgct	572

結果 ana-specific

ana 固有な AnMITE1: 挿入配列の前後に 5 bp のDRが生成

ana_sp8	TATCGTGTATTTTCCCAAGTGAAAGCAGACGCAAATATACCTGATTAAAT	DR
ava	TATCGTGTATTTTCCCAAGTGAAAGCGGATGCAAATATACCTGAT----	
ana_sp8	IR	TRU TRU
	TAGGACTTACCTACTCGATTGTTTGTTCGCGCTCGATTATAGGGGTGTA	
ava	-----	
ana_sp8	TRU TRU	
	GGGTTTTCARCATTTATACCCCGACACCTTCTCCAAACCCCTTGATTTTTC	
ava	-----	
ana_sp8	IR DR	
	GTTTTCATGGGTATGTCCTTAAATTATATTTGCTAATTTCTATTGGTAA	
ava	-----TAAATGATACTTGCCAATTTCTGTTGGTAA	

結果 ana-ava-common

ana, ava 共通な AnMITE1: タンデムリピートのユニット数が異なる

ana_sp32	GTAGTTTTOGATCTAAAACTGAGTACTAAGTGAAGAGTAAGAAGAAATTAG	DR
ava_sp76	GTAGTTTTCGGTCTAGAAATGAGTACTGAGTGAGGAGTAAGAAGAAATTAA	
ana_sp32	TAGGACTCACGTACCCAGGTTGTCCTTTGAGAGCTGGGTGTAA	TRU TRU
ava_sp76	TAGGACTTACACGCCAGGTTGTCCTTTGAGAGCTGGGTGTAA	
ana_sp32	GGGTGTAAAGAAATTTTCAAAGTTGTCCAAACCCCTTGATTTT	TRU TRU
ava_sp76	GGGTATAGGGGTGTAAAGAAATTTCAAACCCCTTGATTTT	
ana_sp32	CTTTTCATCGGTAAAGTCATTATTAAATCATTCTTCTTATCCCCACTTTTAC	TRU DR
ava_sp76	CTTTTCATCGGTAAAGTCATTATTAAATCATTCTTCTTATCCCCACTTTTAC	

こいつらの伝播時期は？

- MITE のマルチプルアラインメントを作成
 - 中が崩れている MITE 群のアラインメントは ClustalW ではつらい
 - さらに、内部のリピート数の変化など進化速度が違いそうな部分は削りたい
- 手でやるしか？

XCED

- 宮田研(隈、加藤 et al.)の統合環境
 - <http://www.biophys.kyoto-u.ac.jp/~katoch/programs/align/xced/>
- XCED 内蔵の高精度マルチプルアライメント(MAFFTも)をベースに手でモディファイ

xced の図
(xced のページよりコピペ)



ClustalW

- XCED のアラインメントを FASTA フォーマットで吐き出す
- ClustalW で読み込んでツリーだけ作成

% clustalw ごによごによ

- mite.dnd ファイルができた

R - ape のインストール

<http://www.r-project.org/>

統計パッケージだが ape ライブラリで系統樹も描ける

<http://cran.r-project.org/src/contrib/Descriptions/ape.html>

```
% R  
> install.packages("ape")  
> example(plot.phylo)  
> q()
```

勝手にネットからダウンロード、コンパイル、インストールまでしてくれる。すげー。

パーミッションの問題が出る場合は root で R を起動したり `sudo R` としたりする。

ape で系統樹を描く

```
> tree = read.tree("tree.dnd")

# 無根系統樹を描く
> postscript("tree_unrooted.ps", horizontal=F, pointsize=5)
> plot(tree, "u", lab4ut="axial", font = 4, cex = 0.5)
> dev.off()

# 有根系統樹を描く
> postscript("tree_rooted.ps", horizontal=F, pointsize=5)
> plot(tree, font = 4, cex = 0.5)
> dev.off()

# PDF に変換
% ps2pdf tree_unrooted.ps
```

- lab4ut が unrooted tree 用のラベルの向き axial か horizontal を指定
- font 0〜4で bold とか italic とか(なんでやねん)
- cex と font の大きさを試行錯誤で設定

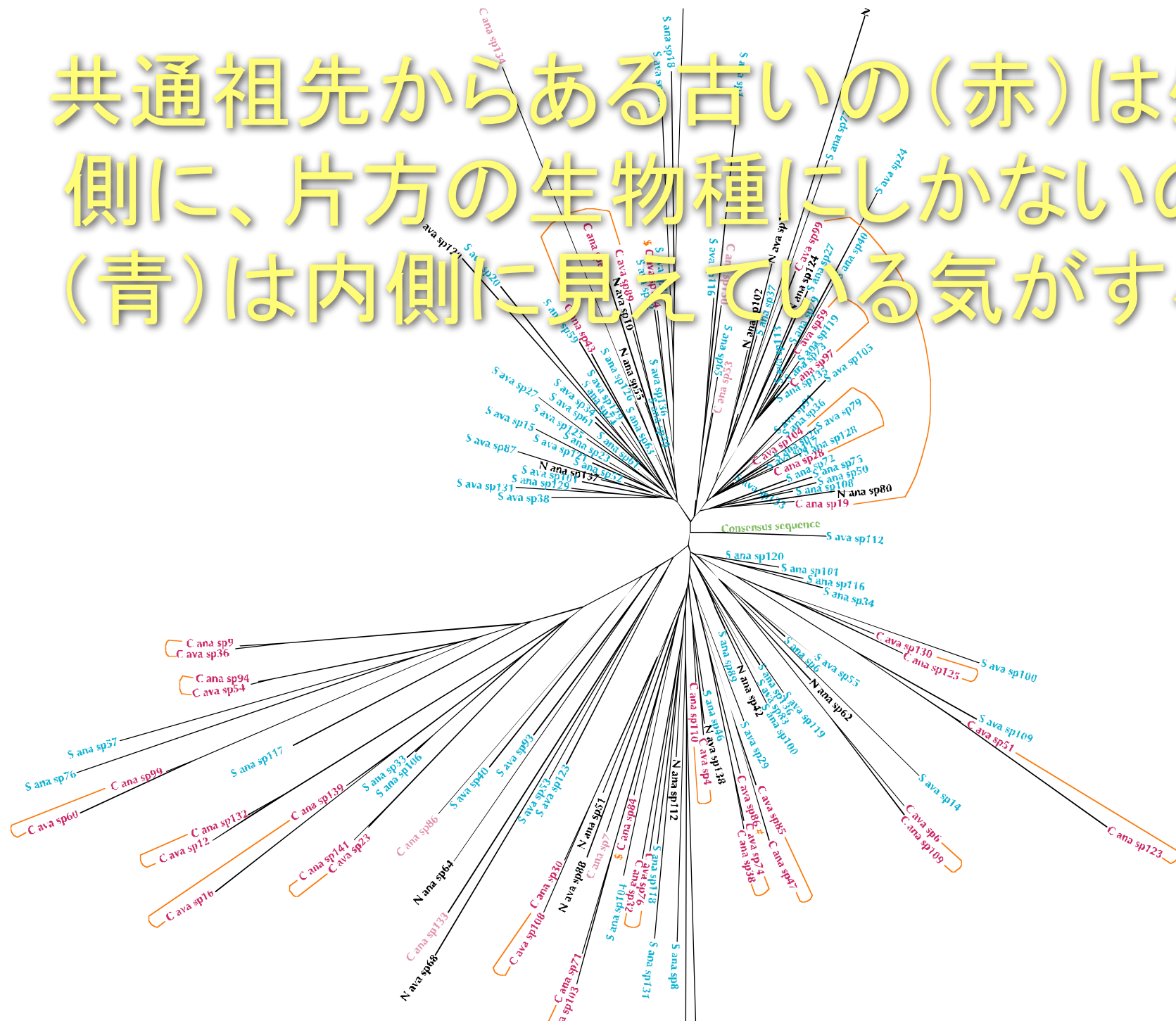


こんな感じ:あとはイラレなどで加工

そういえばMITEの伝播時期は...?



共通祖先からある古いの(赤)は外側に、片方の生物種にしかないの(青)は内側に見える気がする



あとは**Ruby** の新隠し機能

あー、もう時間がないっ

続きは GIW 2005 で！(ソフトデモ)

2005/12/19-21 GIW2005@パシフィコ横浜
今年もたぶん「オープンバイオ研究会BoF」
やります。

オープンバイオ研究会 [http://open-bio.jp/
jointo:open-bio-info@lists.sourceforge.jp](http://open-bio.jp/jointo:open-bio-info@lists.sourceforge.jp)