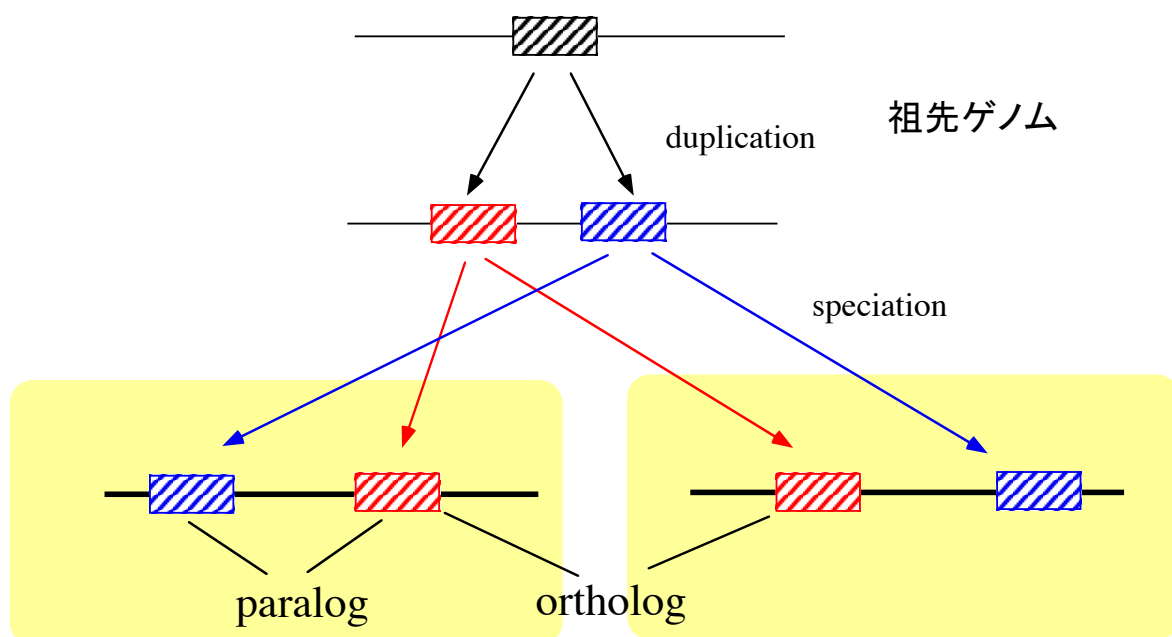


BLASTによるオーソログ解析

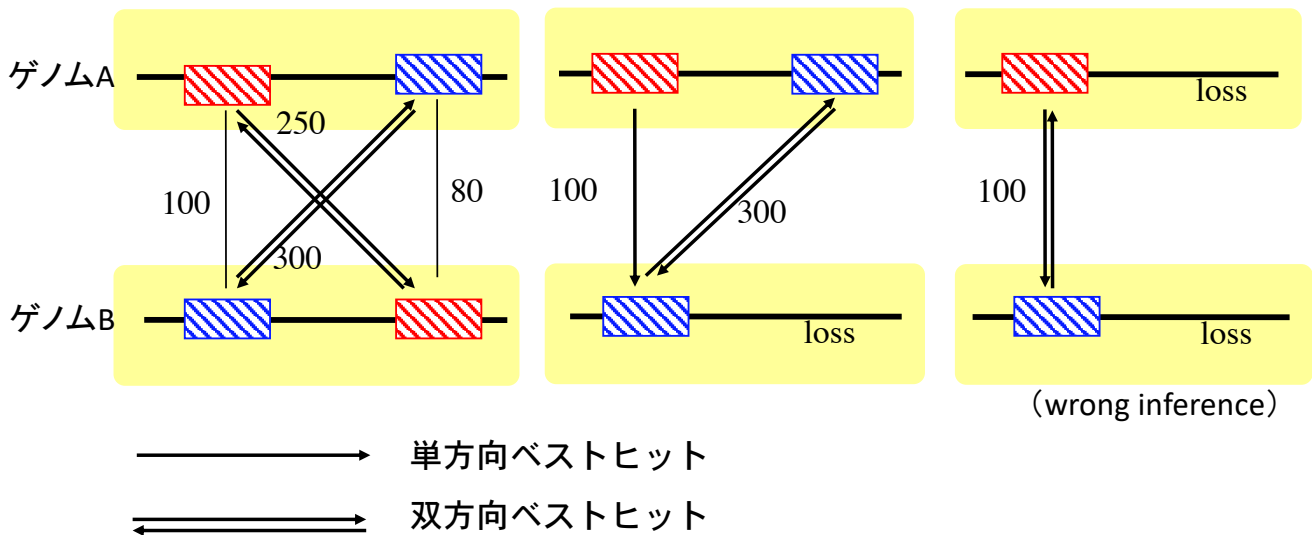
内山郁夫

オーソログとパラログ



オーソログの操作的定義

双方向ベストヒット (bi-directional best hit/reciprocal best hit)



双方向ベストヒットの検出

ゲノム1の遺伝子 類似性スコア
ゲノム2の遺伝子

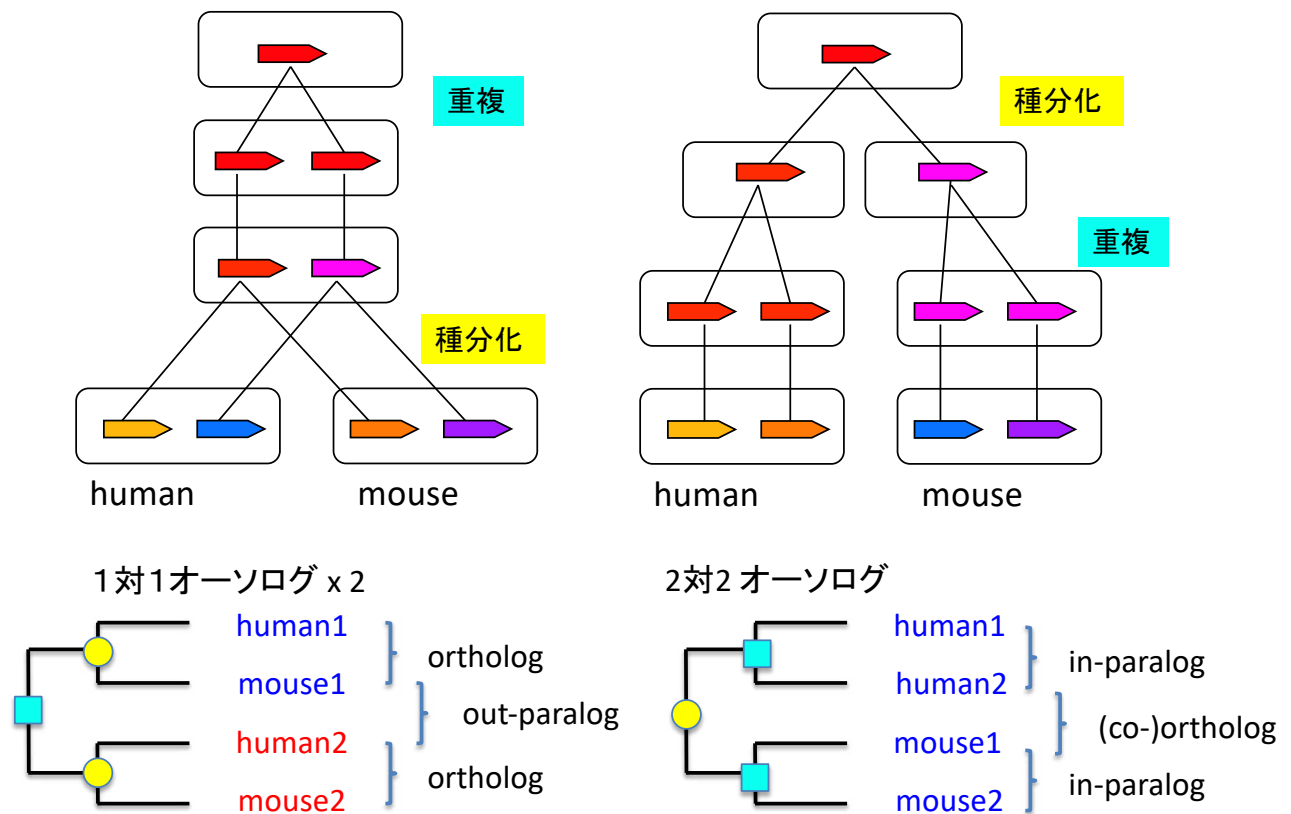
spo:SPAC4F8.12C	sce:YHR165C	3060
spo:SPAC22G7.06C	sce:YJL130C	2939
spo:SPAC56E4.04C	sce:YNR016C	2714
spo:SPAPB1E7.07	sce:YDL171C	2568
spo:SPBC216.07C	sce:YKL203C	2296
spo:SPBC216.07C	sce:YJR066W	2276
spo:SPAC4A8.11C	sce:YPL231W	2247

Lines はスコア順にソートされたファイルの行 のリスト

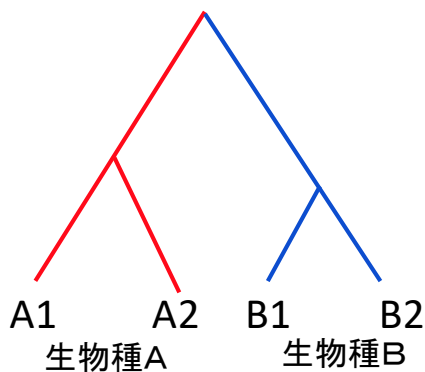
```
for line in Lines do
    (name1, name2, score) = split(line)
    Rank1[name1]++
    Rank2[name2]++
    if (Rank1[name1] == 1
        && Rank2[name2] == 1) then
        print line
    fi
done
```

入力ファイル: ゲノム間の総当りの
類似性スコアのリスト。
スコアの大きい順にソートされているとする。

in-paralog と out-paralog



多対多の関係を考慮した拡張



A1-B1, A1-B2, A2-B1, A2-B2は類似度がほぼ同じ
→いずれもオーソログの関係

RATIO = 0.9; #類似度を同じと見なす許容範囲

Lines はスコア順にソートされたファイルの行のリスト

for line in Lines do

(name1, name2, score) = split(line)

if (Best1[name1]が未定義) then

Best1[name1] = score

fi

if (Best2[name2]が未定義) then

Best2[name2] = score

fi

スコアがベストのRATIO倍以上だとベストヒットと見なす

if (score >= Best1[name1] * RATIO

&& score >= Best2[name2] * RATIO) then

print line

fi

done

実習：出芽酵母と分裂酵母の オーソログ解析

bit-score の順にソートする

```
% sort -k 12,12nr sce-spo.blast > sce-spo.blast.sorted
```

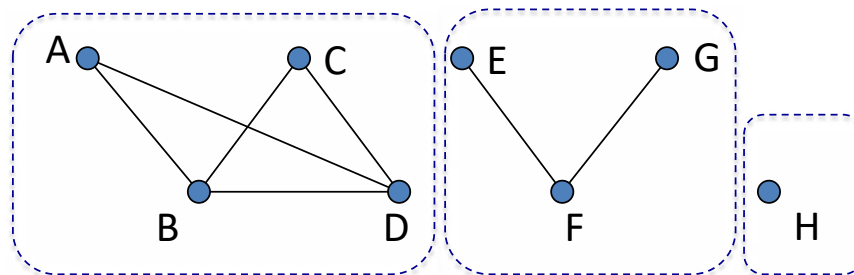
双方向ベストヒットをとる

```
% bbh.pl sce-spo.blast.sorted > sce-spo.bbh
```

双方向ベストヒットをとる(条件を緩めたバージョン)

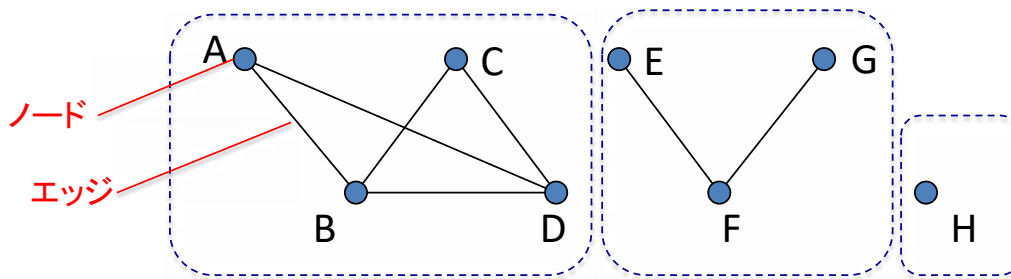
```
% bbh2.pl sce-spo.blast.sorted > sce-spo.bbh2
```

単連結クラスタリング



- 関係で結ばれた遺伝子対をすべてつなぐ
- 一般に、ホモロジー検索結果の整理に有効
 - 相同関係の推移性に基づく
「AとBが相同でBとCも相同なら、AとCも相同である」
 - ホモロジー検索では、類似性が低い相同関係を取りこぼす可能性がある
→ 単連結クラスタリングは検索のとりこぼしを補ってくれる
- アルゴリズムはシンプル(グラフの連結成分connected componentをとる)

2項関係のグラフによる表現



A	B
A	D
B	C
B	D
E	F
F	G

2つのノードがエッジでつながっていることを2次元ハッシュを用いて表す

$$\text{Link}["A"]["B"] = \text{Link}["B"]["A"] = 1$$
$$\text{Link}["A"]["D"] = \text{Link}["D"]["A"] = 1$$

`keys(Link["A"])` (ハッシュ `Link["A"]` におけるキーの集合)

== ノード "A" とつながっているノードの集合

→ "B" と "D"

入力ファイル: 関連を持つ遺伝子対のリスト

単連結クラスタリング

データを読み込んでグラフを構築

for line in Lines do

(node1, node2) = split(line)

node1 と node2 がつながっていることを2次元ハッシュで表す

Link[node1][node2] = Link[node2][node1] = 1;

done

nodeSet = keys(Link)

for node in nodeSet do

if (Mark[node]==0) then

Cluster (配列) を空にする

Traverse(node)

Cluster を出力する

fi

done

サブルーチン Traverse

node1 につながるノードを再帰的に

たどって Cluster に加える

Traverse (node1) {

if (Mark[node1] > 0) then

マークされたノードはスキップ

return

fi

Cluster に node1 を加える

Mark[node1] = 1 # 出力済みマー

ク

nodeSet = keys(Link[node1])

for node2 in nodeSet do

Traverse(node2);

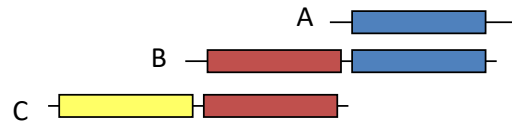
done

```
RPOB SA0500 4628
RPOC SA0501 4507
POLC SA1107 4390
NARG SA2185 3997
GLTA SA0430 3898
PYCA SA0963 3830
PYRAB SA1046 3712
UVRA SA0714 3397
VALS SA1488 3350
```

入力ファイル: 双方向ベストヒットとなる類似遺伝子対のリスト。ソートされている必要なし。

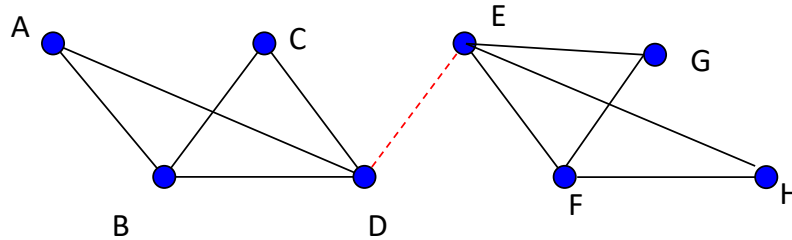
単連結クラスタリングの問題点

- マルチドメイン蛋白質の場合、推移律が満たされないことがある



→アライメントのカバレッジを上げる

- ひとつでも間違った関係があると、分類を大きく間違える可能性がある



→類似性スコアの閾値を上げる

実習: オーソログ結果のクラスタリング

双方向ベストヒット(条件を緩めたバージョン)のクラスタリング

```
% slink.pl sce-spo.bbh2 > sce-spo.oclust
```

タイトルをつける。まずFASTAファイルからタイトル行を抜き出したファイルを

作成して、add_title.plを使ってジョインする。

```
% grep -h '^>' sce.fas spo.fas | sed 's/^>/'  
| sed 's/ /<tab>/' > sce-spo.tit
```

```
% add_title.pl sce-spo.oclust sce-spo.tit  
> sce-spo.oclust_title
```

類似性に関する指標

1. bit score
 2. E-value 統計的評価
 3. percent identity
 4. percent positive score (ppos)
 5. score/length
 6. query coverage ((qend-qstart+1)/qlen)
 7. subject coverage ((send-sstart+1)/slen)
- 長さ当たりの類似性
→進化距離を反映
- 全長が
マッチ
するか?

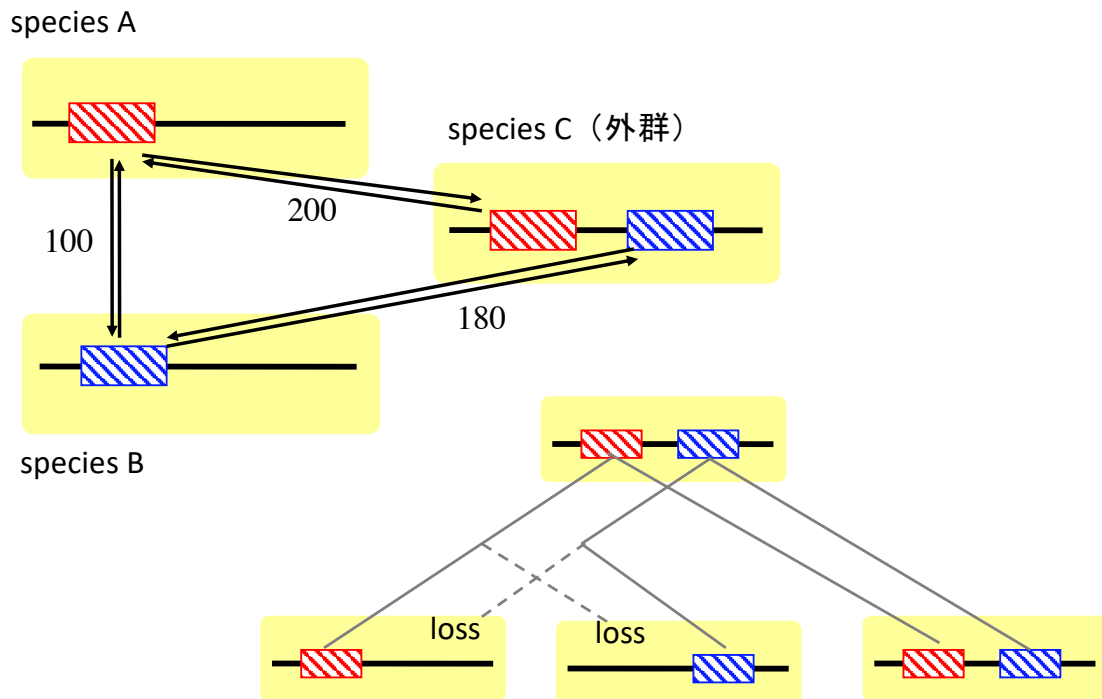
4,6,7は -outfmt 6 で追加のカラム指定が必要
例) -outfmt "6 std qlen slen"

多対多オーソログをより正確にとるには

1. 種間比較だけでなく、種内比較の結果も考慮する
 - specA-specB に加えて、specA-specA、specB-specBの比較も行う(→2つのファイルを連結して自分自身に対して相同性検索を行う)

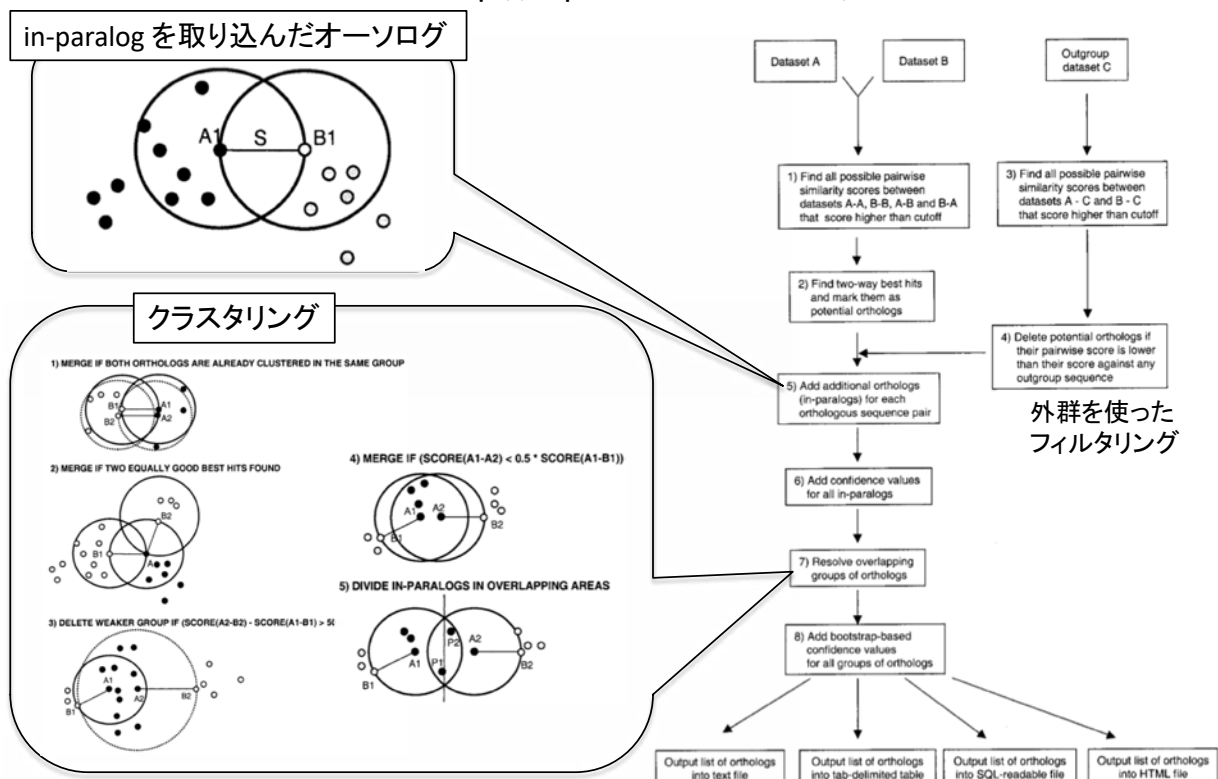
```
% cat sce.fas spo.fas > sce+spo.fas
% makeblastdb -in sce+spo.fas -out sce+spo
% blastp -db sce+spo -query sce+spo.fas
```
2. オーソログの同定基準やクラスタリング手順を工夫する

外群を加えたオーソログ解析

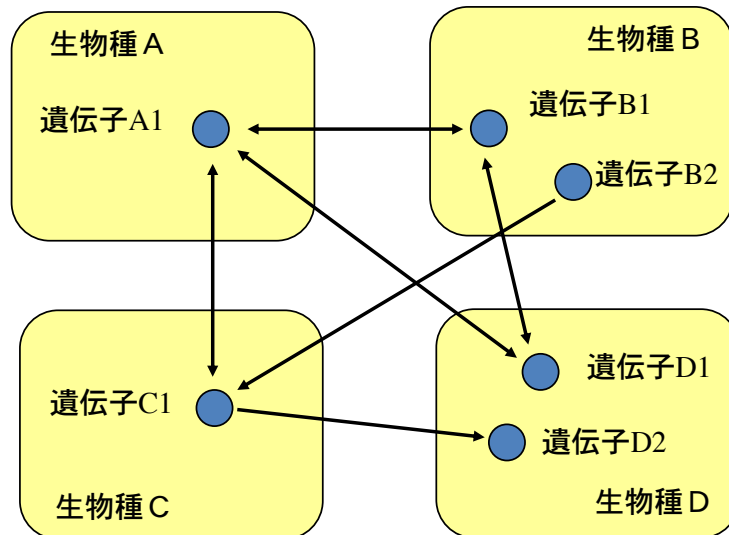


Inparanoid

<http://inparanoid.sbc.su.se/>



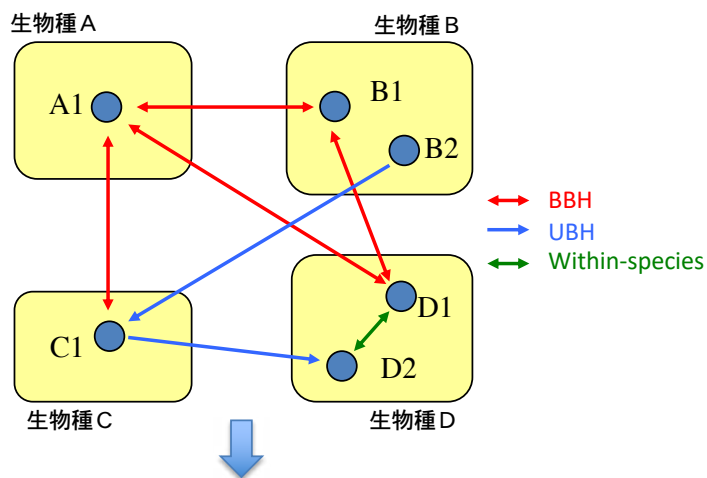
多種間のオーソログ解析



オーソログ推定手法

Graph-based method

ペアワイズ比較→グラフ作成

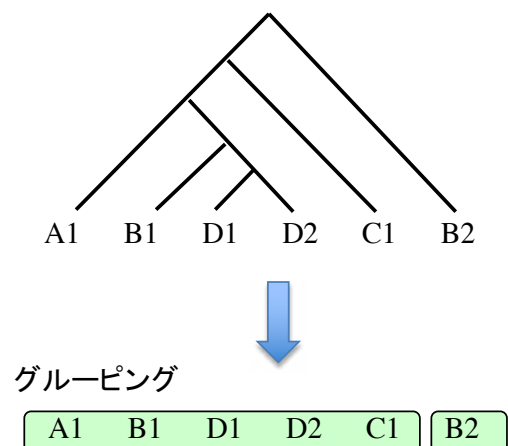


クラスタリング

- Triangle linkage (COG)
- Hierarchical clustering
- Markov clustering

Tree-based method

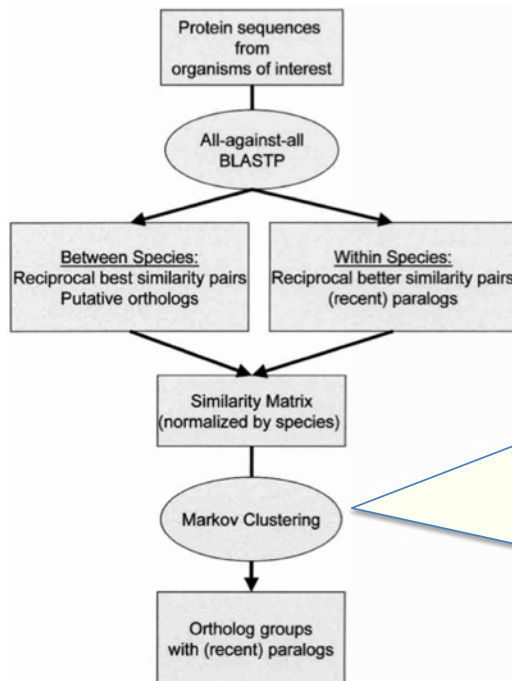
系統樹作成→グルーピング



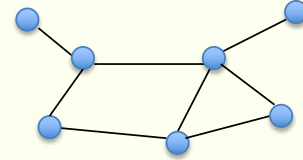
OrthoMCL

<http://orthomcl.org/>

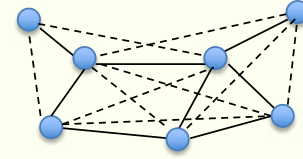
Markov clustering (MCL)



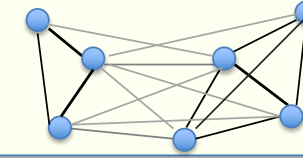
original graph
(確率値による重み付きグラフ)



Markov expansion
(推移確率による拡張)

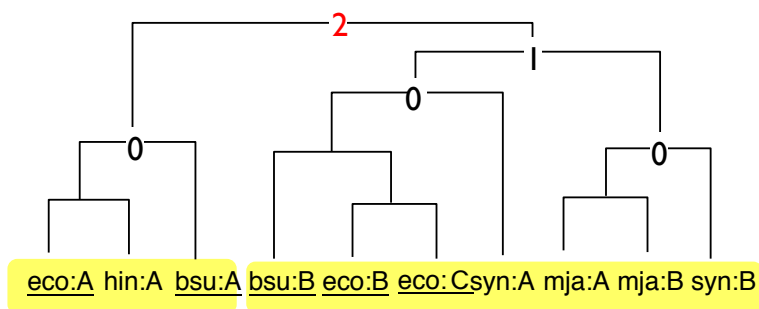
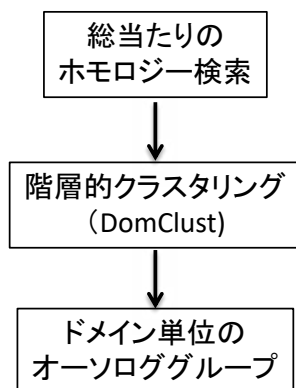


inflation
(重みのコントラストを強調)

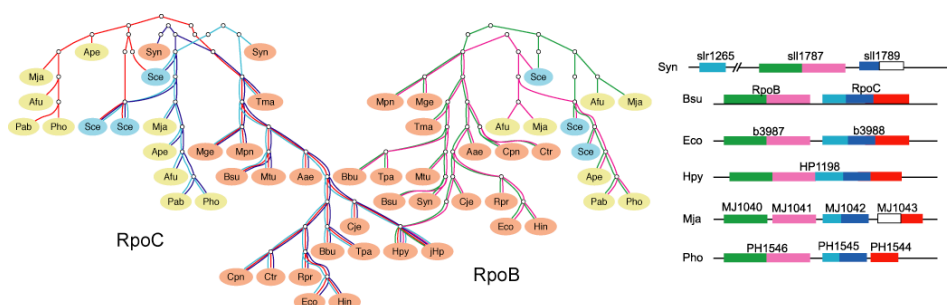


DomClust

<http://mbgd.genome.ad.jp/domclust/>



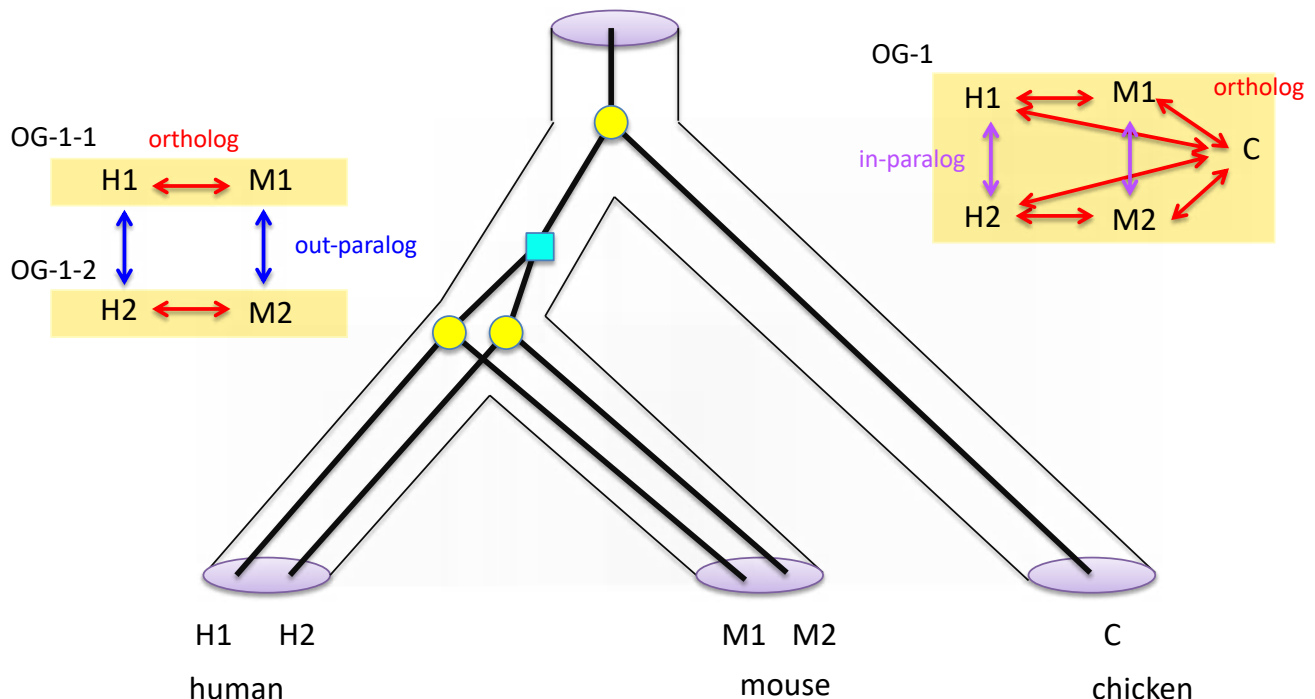
Cut if $\frac{|Spec(A) \cap Spec(B)|}{\min(|Spec(A)|, |Spec(B)|)} > p$
(種の重複度チェック)



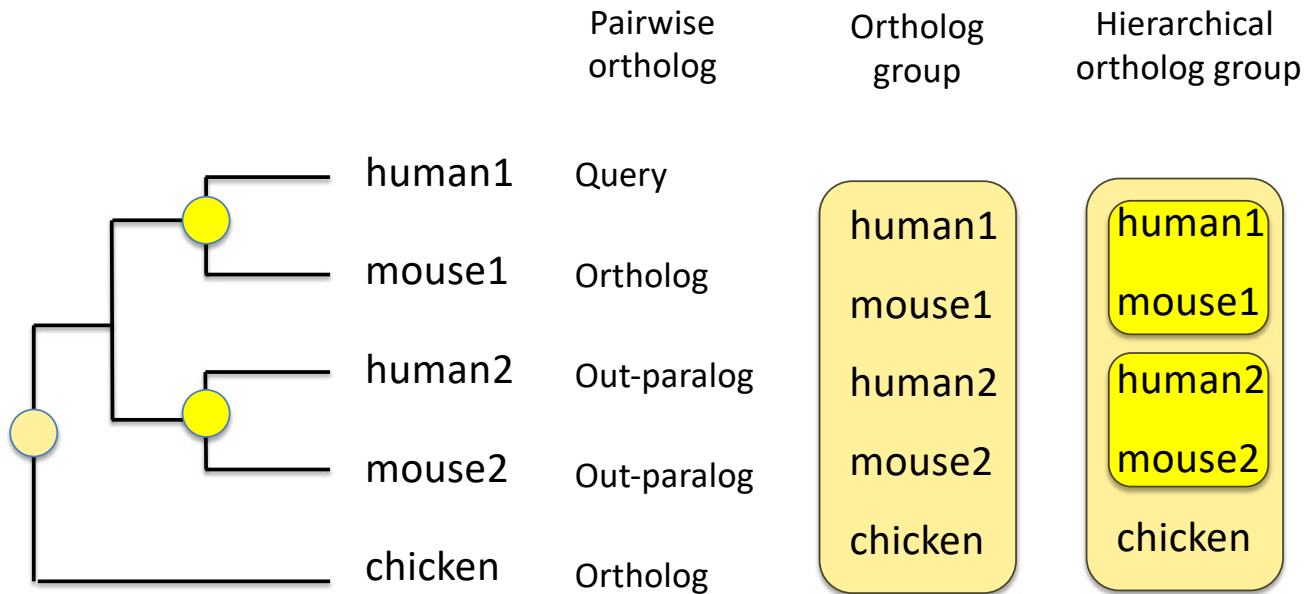
Ortholog databases

- マニキュアルキュレーションによるもの
 - COG/KOG** Manually curated after comprehensive classification with an automated method.
 - Protein Clusters (NCBI)** Probably based on COG, but the groups are more compact than COG.
 - TIGRFAMs** Hidden Markov model for functional prediction
 - KO (KEGG)** Integrated into the pathway database
 - HAMAP** Prokaryotic protein classification based on the UniProt database
 - TreeFam** Orthologs based on phylogenetic trees
- 自動分類によるもの
(グラフベース)
 - InPranoid**
Ortholog definition between two species considering In-paralogs
 - OrthoMCL-DB**
Clustering using TribeMCL
 - MBGD**
Clustering using a hierarchical clustering with domain splitting (DomClust).
 - eggNOG**
Updated version of COG. Constructed using the same method used for COG construction.
 - OMA**
Ortholog database of a broad range of organisms constructed based on a graph based algorithm
 - OrthoDB**
Comprehensive catalog of orthologs
- (ツリーベース)
 - Berkeley PHOG**
Phylogenetic tree based method is applied to PhyloFacts database.
 - PhylomeDB**
Orthology assignment based on complete collection of phylogenies (phylomes)

オーソロジー関係は対象とする系統群の範囲 (taxonomy range) によって変わる




多生物間のオーソログ関係の表現



MBGD: Microbial Genome Database for Comparative Analysis

<http://mbgd.genome.ad.jp/>



Microbial Genome Database for Comparative Analysis

http://mbgd.genome.ad.jp/

Welcome to MBGD

MBGD is a database for comparative analysis of completely sequenced microbial genomes, the number of which is now growing rapidly. The aim of MBGD is to facilitate comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison.

References: *Nucleic Acids Res.* 31:58-62 (2003) / *Nucleic Acids Res.* 35:D343-D346 (2007) / *Nucleic Acids Res.* 38:D361-D365 (2010) / *Nucleic Acids Res.* 41:D631-D635 (2013) / *Nucleic Acids Res.* 43:D270-D276 (2015)

Complete genome sequences

(Total 4742 genomes, Last update 2016/05/19)

Taxonomy Browser

Set Default

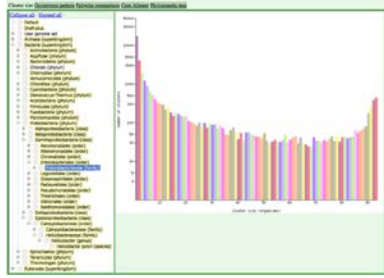
Draft-plus version

Currently selected organisms (868) are highlighted in green.
Please press "Reload" button when you return here by "Back" button.

Bacteria (4150)	Deinococcus (1)	Legionella (2728)	Thermococcus (720)
Deinococcus (1)	Deinococcus-Thermus (22)	Methylobacterium (23)	Methanomonas (373)
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus-Thermus (22)	Thermoplasma (15)
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus-Thermus (22)	Deinococcus (1)	Deinococcus (1)	
Deinococcus			

MBGD: microbial genome database for comparative analysis

Ortholog table selection/overview



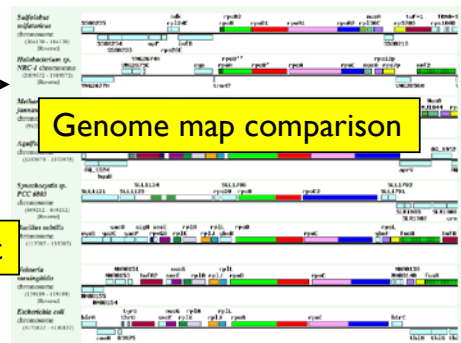
Ortholog table

[illegible]

Ortholog group

[illegible]

Genome map comparison



Sequence alignment

