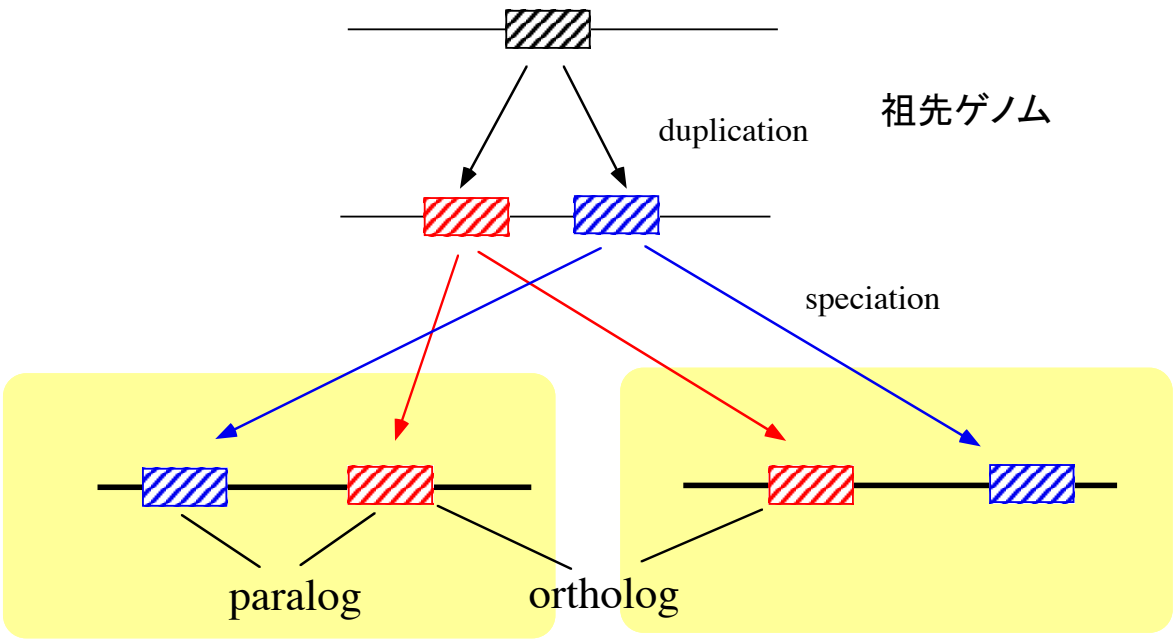
NIBB GITC「BLAST 自由自在」
September 14-15, 2017@Okazaki

BLASTによるオーソログ解析

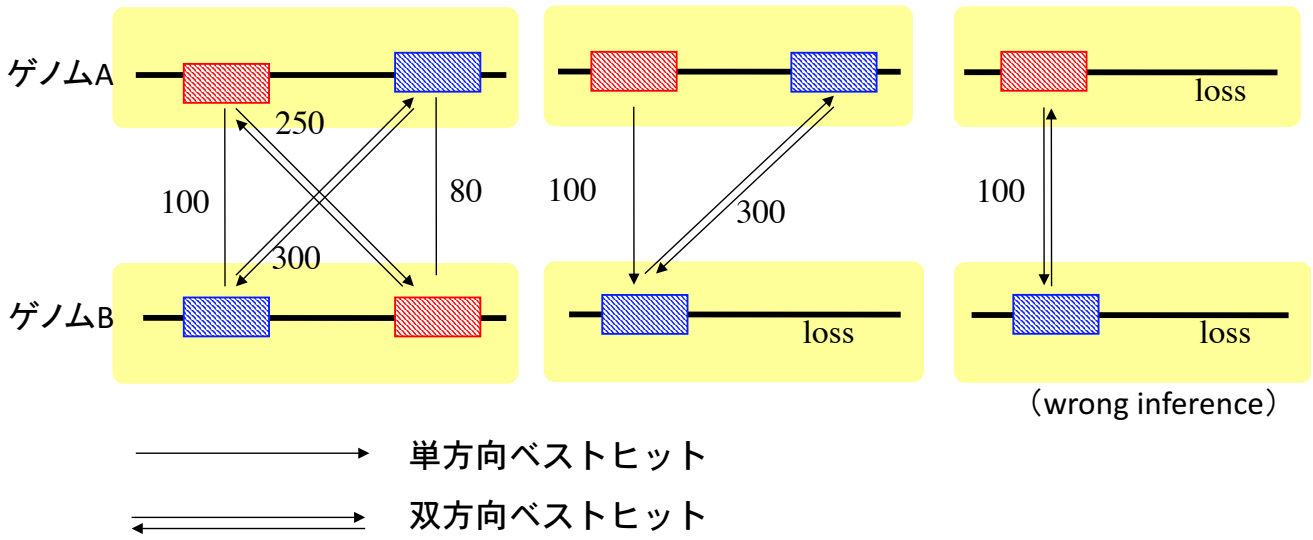
内山郁夫

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

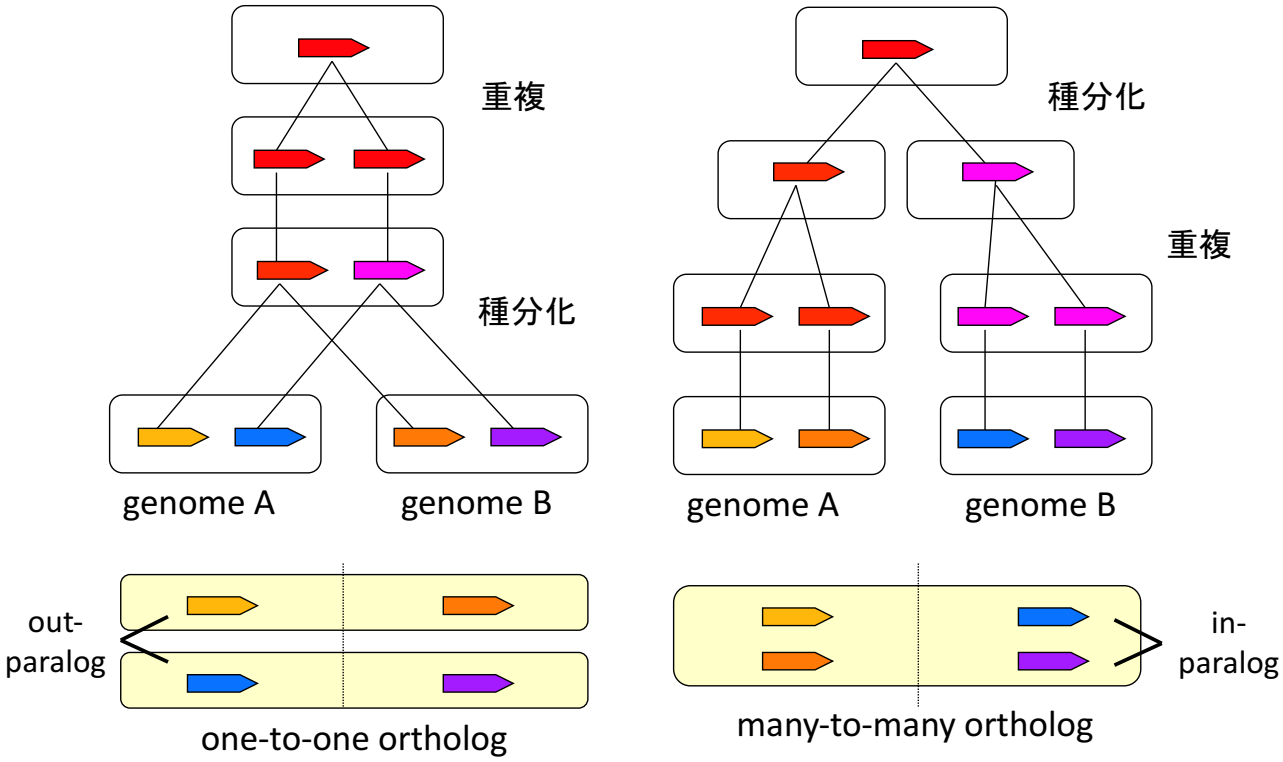
```
spo:SPAC4F8.12C   sce:YHR165C   3060
spo:SPAC22G7.06C  sce:YJL130C   2939
spo:SPAC56E4.04C  sce:YNR016C   2714
spo:SPAPB1E7.07   sce:YDL171C   2568
spo:SPBC216.07C   sce:YKL203C   2296
spo:SPBC216.07C   sce:YJR066W   2276
spo:SPAC4A8.11C   sce:YPL231W   2247
```

Lines はスコア順にソートされたファイルの行 のリスト

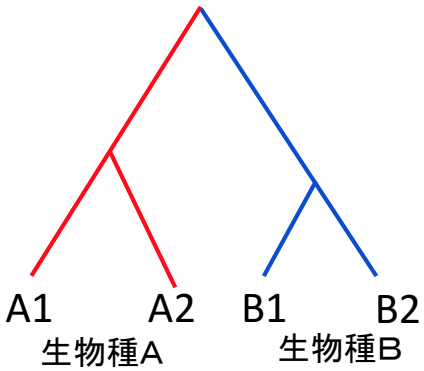
```
for line in Lines do
    (name1, name2, score) = split(line)
    Rank1[name1]++
    Rank2[name2]++
    if (Rank1[name1] == 1
            && Rank2[name2] == 1) then
        print  line
    fi
done
```

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。

種内パラログ in-paralog が存在する場合



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

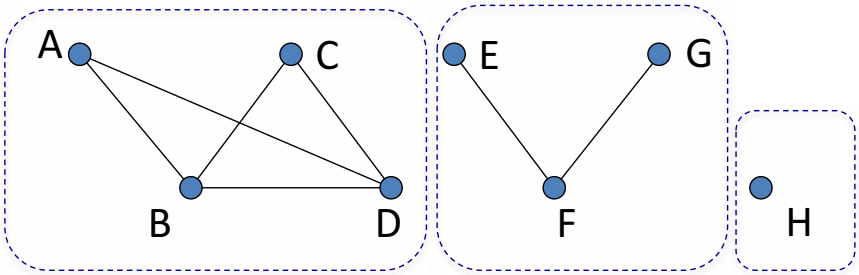
Lines はスコア順にソートされたファイルの行のリスト

```
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

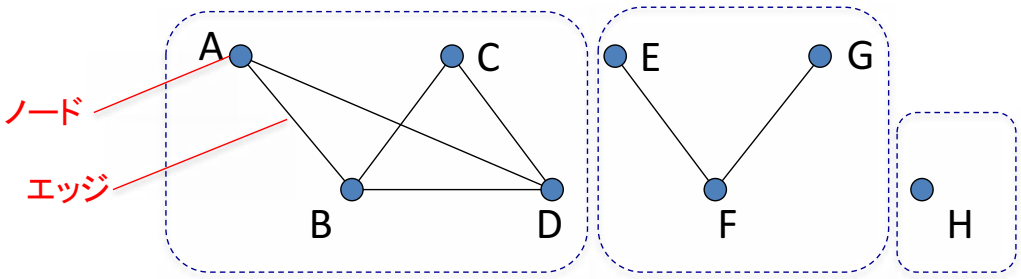
```
# bit-score の順にソートする
% sort -k 12,12nr sce-spo.blast > sce-spo.blast.sorted

# 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

Link["A"]["B"] = *Link*["B"]["A"] = 1
Link["A"]["D"] = *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
```

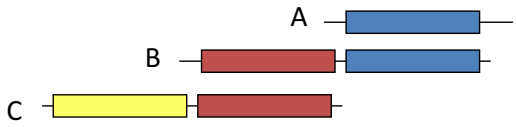
```
# サブルーチン Traverse
# node1につながるノードを再帰的に
# たどって Cluster に加える
Traverse (node1) {
    if (Mark[node1] > 0) then
        # マークされたノードはスキップ
        return
    fi

    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。

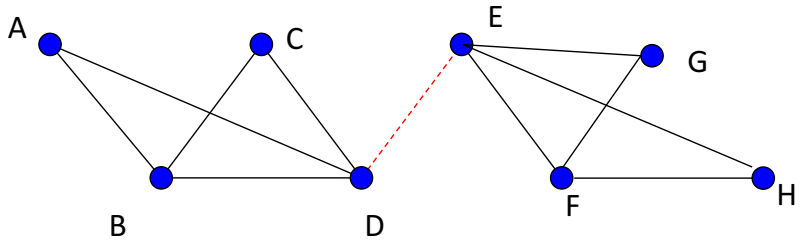
単連結クラスタリングの問題点

- 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 spo | 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

長さ当たりの類似性
→進化距離を反映

6. query coverage ((qend-qstart+1)/qlen)
7. subject coverage ((send-sstart+1)/slen)

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"`

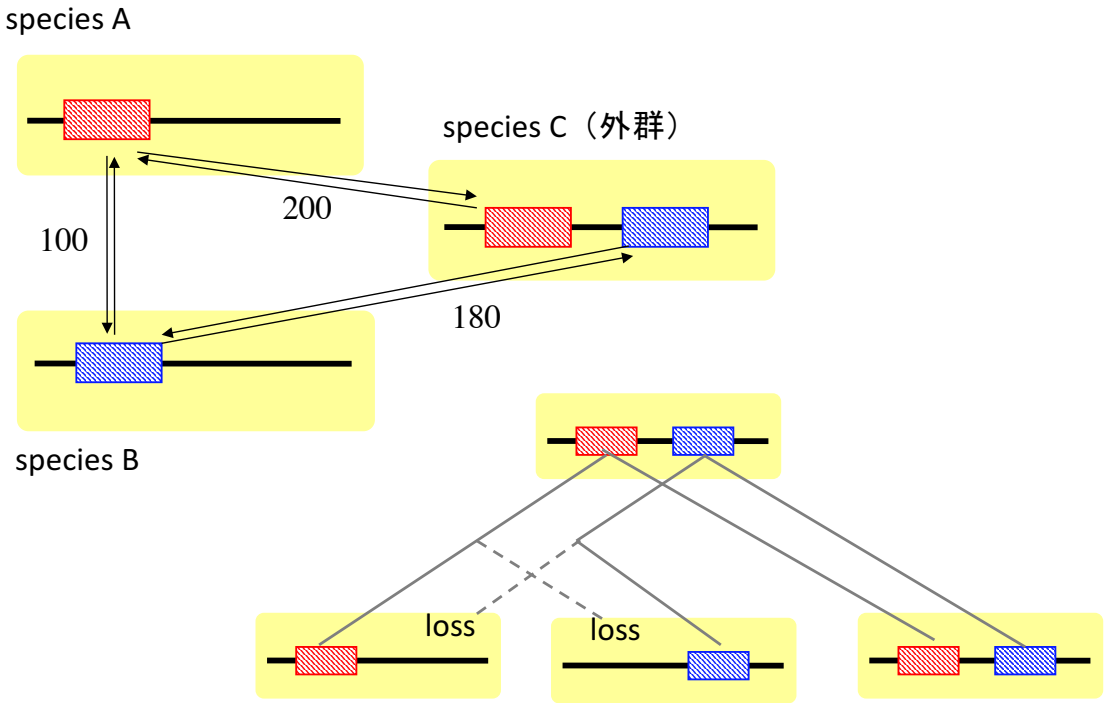
多対多オーソログをより正確にとるには

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 spo > sce+spo
% blastp –db sce+spo –query sce+spo
```

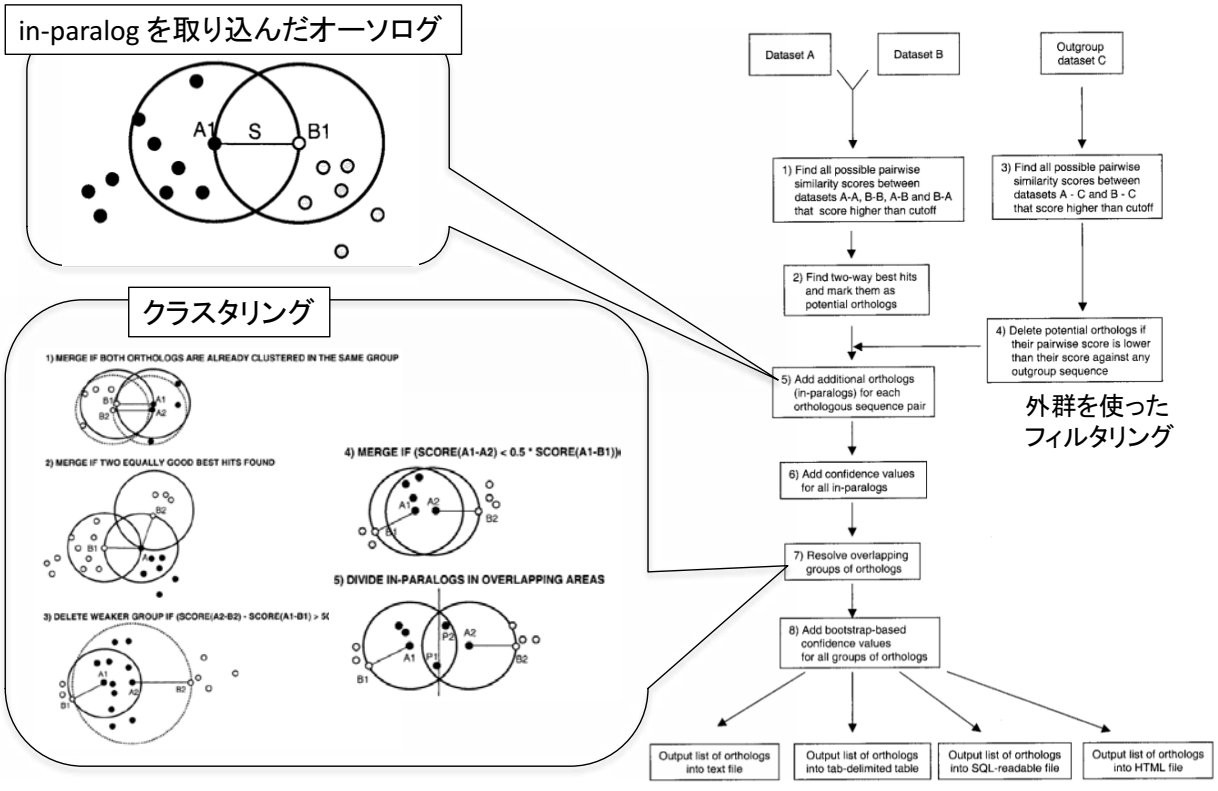
2. オーソログの同定基準やクラスタリング手順を工夫する

外群を加えたオーソログ解析



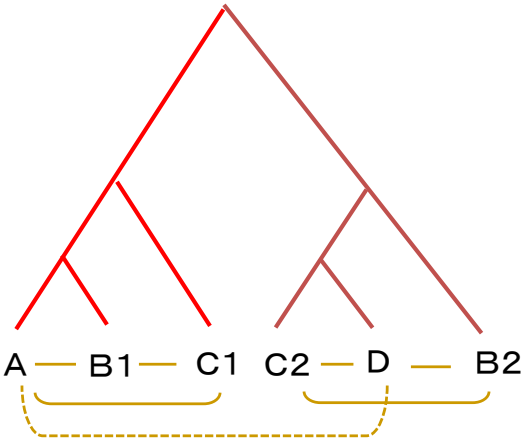
Inparanoid

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

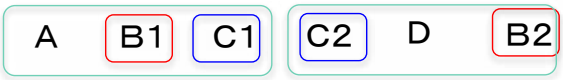


3種以上のオーソログ解析

遺伝子系統樹

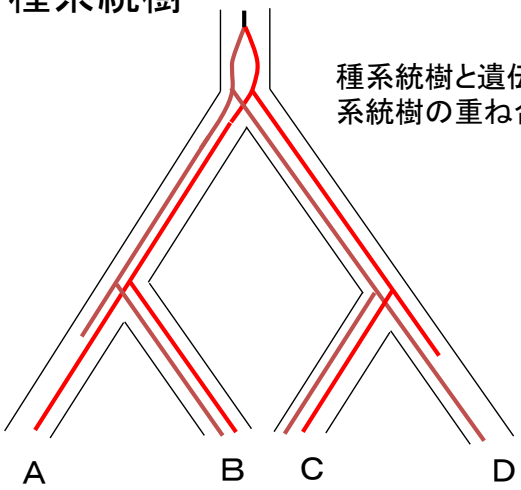


双方向ベストヒットのクラスタリング



種の重複度チェック

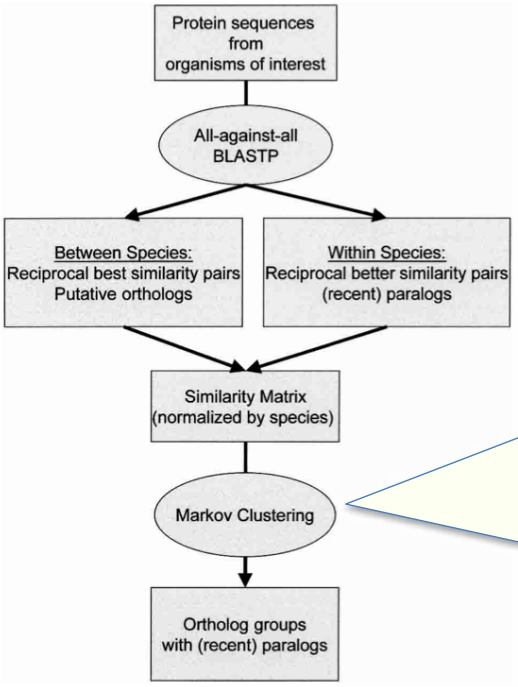
種系統樹

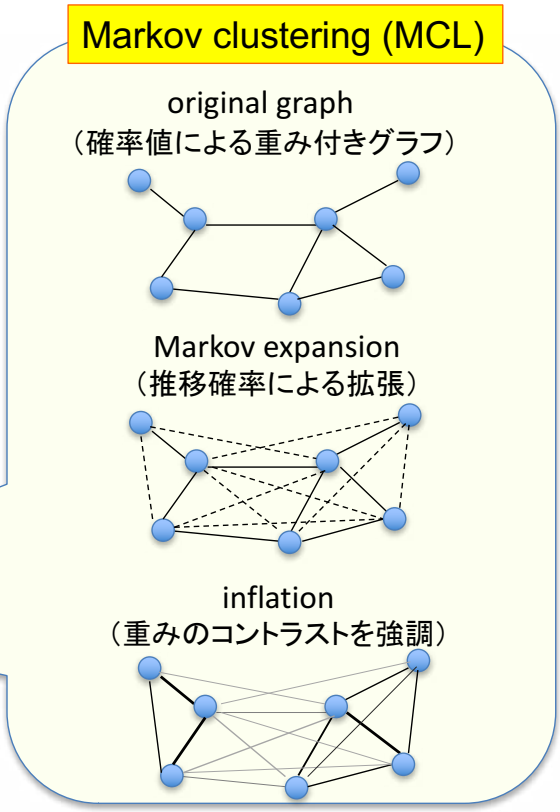


種系統樹と遺伝子系統樹の重ね合わせ

OrthoMCL

http://orthomcl.org/





DomClust

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

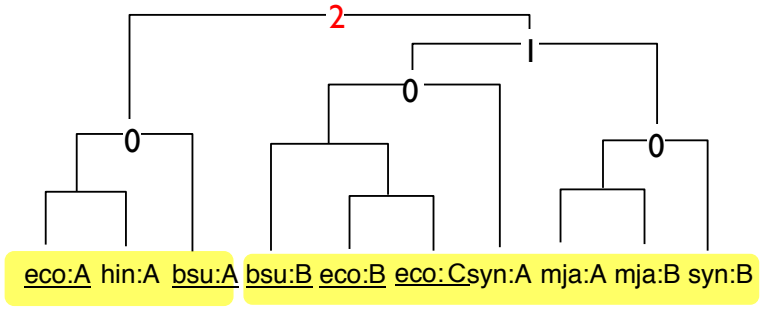
総当たりの
ホモロジー検索

↓

階層的クラスタリング
（DomClust)

↓

ドメイン単位の
オーソロググループ



Cut if $\frac{|Spec(A) \cap Spec(B)|}{\min(|Spec(A)|, |Spec(B)|)} > p$

（種の重複度チェック）