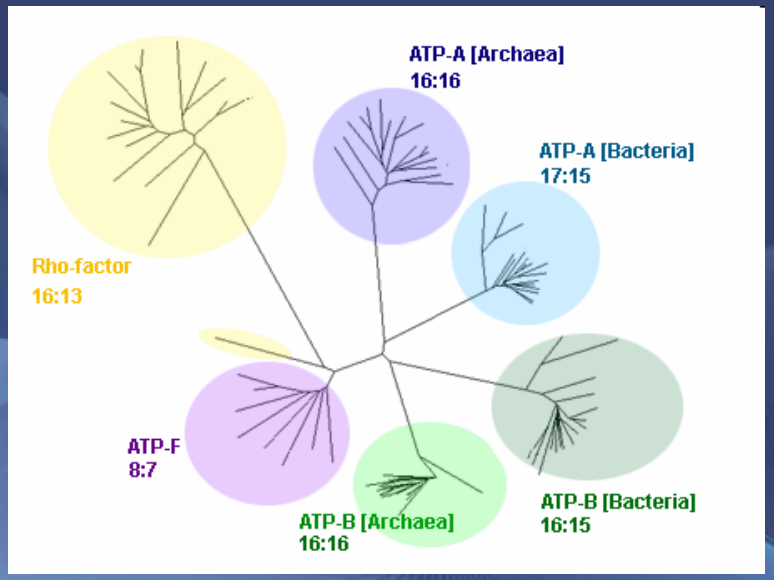
Bioinformatik Problemseminar SS07

Thema: „Rekonstruktion von Genfamilien, BranchClust“



16.07.07

Universität Leipzig

Christian Arnold

Gliederung

1. Motivation und Einführung
2. Bisherige Ansätze
    - 2.1. Tree reconcilitation
    - 2.2. Genome-specific best Hits (SymBets)
3. BranchClust
    - 3.1. Algorithmus
    - 3.2. Bewertung
    - 3.3. Installation und Benutzung
4. Quellen

1. Motivation und Einführung

- die Selektion von Genfamilien ist eines der fundamentalen Probleme in der evolutionären Genomik
- korrekte Identifikation von Orthologen und Paralogen ist äußerst entscheidend für die Rekonstruktion und Interpretation von phylogenetischen Bäumen und den damit verbundenen Fragen:
  - Wie viele gemeinsame Gene besitzen verschiedene Spezies?
  - Welche Gene haben eine gemeinsame Geschichte, welche sind erst neu entstanden?
  - Welche Gene wurden evolutionär modifiziert durch Duplikation, Deletion, oder
  - horizontalem Gentransfer

- Gegeben: Set von Genomen / Genen
- Ziel: automatisierte Methode, um Genfamilien von orthologen (und paralogen) Genen herauszufinden

2. Bisherige Ansätze

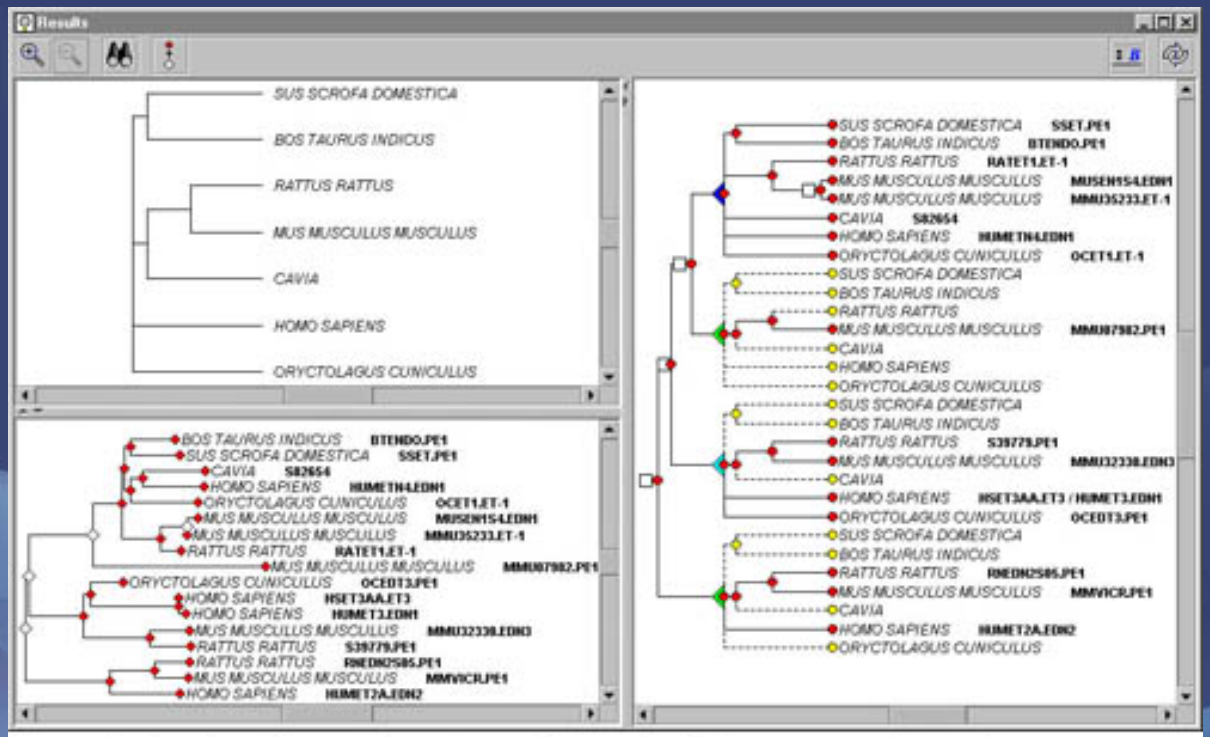
- 5 Arten von Genevolution (grob geordnet nach Wichtigkeit und Vorkommen):
  - Vertikale Speziation (Speciation) mit Modifikation
  - Genduplikation, gefolgt von Speziation
  - Genverlust
  - Horizontaler Gentransfer (HGT)
  - Verschmelzen, Spalten und andere Rearrangement Ereignisse
- Zusammen ergeben diese Veränderungen ein komplexes Netzwerk, die evolutionär über einen langen Zeitraum wirken und gewirkt haben
- Homologie: sehr allgemeiner Begriff, bezeichnet eine Beziehung aufgrund gemeinsamen evolutionärerem Ursprung von zwei oder mehreren Entitäten, ohne genaue Beschreibung, welcher Prozess der Genevolution stattgefunden hat

- Subkategorien davon:
  - Orthologie
  - Paralogie
- Orthologie:
  - Gene, die durch Speziation aus einem einzigen Gen vom gemeinsamen Vorfahren (last common ancestor) entstanden sind
- Paralogie:
  - Gene, die durch Duplikation entstanden sind
- Erweiterte Konzepte (detaillierter im Handout):
  - Co-Orthologs
  - Pseudoorthologs
  - Outparalogs (Paraloge Gene, deren Duplikation vor einem Speziationsereignis stattgefunden hat)
  - Inparalogs (Paraloge Gene, deren Duplikation nach einem Speziationsereignis stattgefunden hat)
  - Pseudoparalogs
  - Xenologs (Xenologous gene displacement)

2.1 Tree reconciliation (Baumabgleich)

- Idee:
  - baumbasierte Methode
  - Topologie eines Genbaumes wird mit einem ausgewählten Spezies tree verglichen
  - Auf Basis von Parsimony Prinzipien werden diese beiden abgeglichen
  - Dabei werden so wenig wie möglich Duplikationen und Genverluste in der Evolution der entsprechenden Gene zugrunde gelegt
  - Der abgeglichene Baum reflektiert dann orthologe Beziehungen
- Praxis:
  - Zuverlässiger als die Genom-specific best Hit Methode
  - Schwieriger zu automatisieren
  - Sehr hohe Komplexität und Artefaktbildung bei Artefakten bei der Baumrekonstruktion
  - benötigt oft bekannten Spezies tree, der oft nur sehr hypothetisch ist (wegen bevorzugtem häufigem Auftreten von HGT in Prokaryoten)
  - Deswegen kann meist höchstens ein Konsensusbaum verwendet werden
  - Zudem werden Inkongruenzen nur durch Duplikation und Deletionen erklärt, aber nicht durch HGTs

- Beispiele und Verbesserungen:
  - HOVERGEN and HOBACGEN
  - RIO
  - Orthostrapper



Reconciliation of a phylogenetic tree (lower-left corner) with the corresponding species tree (upper-left corner). Reconciliation produces a reconciled tree (right) which describes both genes and species history. It allows to deduce the location of gene duplications (white squares) which are the only information needed to distinguish orthologous from paralogous genes.

2.2 Genome-specific best Hits (SymBets)

- Idee:
  - Die Probleme bei der 1. Methode haben zu Vereinfachungen und Annahmen geführt, welche wie folgt sind:
    - 1. Sequenzen von orthologen Genen sind ähnlicher zueinander als zu anderen Genen des Genoms, sie produzieren also symmetrische best Hits (SymBets)
    - 2. Zudem wird angenommen, dass SymBets hauptsächlich von orthologen Genen gebildet werden
  - Man bestimmt also den best Hit, im Symmetriefall sind die Gene ortholog
  - Für n Spezies wird eine orthologe Gengruppe nur dann als ortholog klassifiziert, wenn alle paarweisen Verbindungen symmetrisch sind
- Praxis:
  - Die Hypothese dürfte statistisch gesehen richtig sein
  - Verletzungen der Annahmen kann recht häufig vorliegen:
    - 1. wird bei Inparalogs und schnell eolvierenden Paralogen verletzt
    - 2. wird im Falle von Xenologen und Pseudoorthologen verletzt
  - sehr strikt, geringe false positive Rate
  - False negative Rate kann recht hoch sein im Falle von Paralogen

- Beispiele und Verbesserungen:
  - Clusters of Orthologous Groups (COG): siehe Handout
  - HOBAC-GEN

Table 2 Symmetrical best hits between selected prokaryotic genomes[a]

	Ec	Yp	Hp	Bs	Mg	Aa	Tm	Mj	Ma	Ta
Ec-4289		0.584	0.456	0.305	0.527	0.519	0.428	0.24	0.151	0.308
Yp-4083	2385		0.432	0.28	0.525	0.496	0.423	0.218	0.144	0.275
Hp-1566	714	677		0.403	0.442	0.396	0.321	0.181	0.211	0.176
Bs-4100	1251	1144	631		0.648	0.495	0.465	0.239	0.153	0.306
Mg-480	253	252	212	311		0.469	0.515	0.235	0.281	0.238
Aa-1553	806	771	615	768	225		0.449	0.279	0.33	0.256
Tm–1846	808	780	503	858	247	697		0.245	0.294	0.265
Mj-1770	425	385	284	423	113	434	433		0.489	0.362
Ma-4540	649	589	330	627	135	513	543	866		0.415
Ta-1478	455	406	260	453	114	378	392	535	614	

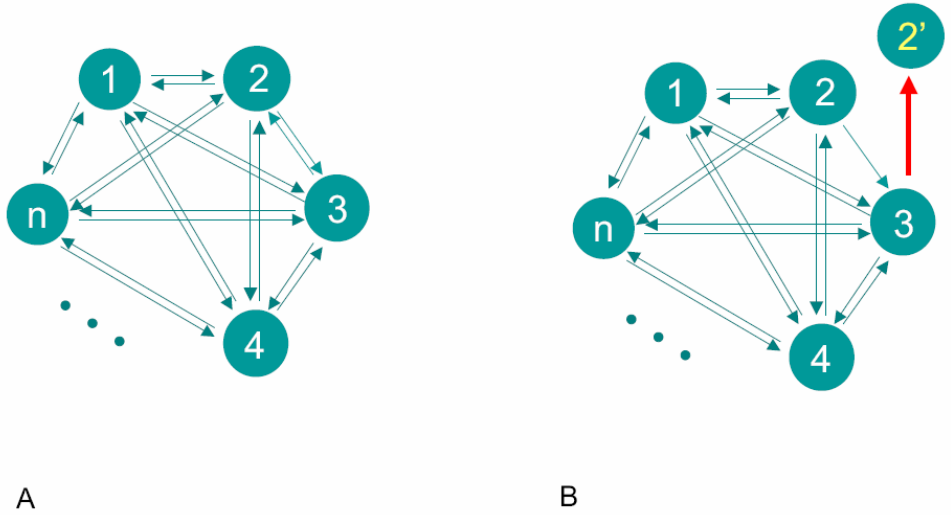


Figure 1
The reciprocal best BLAST hit method. Circles represent genes from n different taxa, arrows signify best BLAST hit relationship; (A) – case of strict reciprocity, (B) – failing of reciprocity in the presence of paralogs.

3. BranchClust

- Kombiniert Ansätze beider Methoden:
  - Tree reconciliation: Baum wird erstellt, der potentielle Genfamilien enthält
  - Best blast hit: aufwändiges Blasten aller Gene gegeneinander und Filtern aller signifikanten (nicht nur der besten) Hits
- Keine Beschränkung bei der Anzahl der Spezies
- Benötigt keinen bekannten Spezies tree
- Kann zwischen Paralogen (In- und Outparaloge) und Orthologen unterscheiden
- Grundsätzliche Idee:
  - nahe verwandte Gene befinden sich in einem Zweig
  - Erkennung von Familien läuft auf Erkennung von Zweigen hinaus, die Gene von allen, oder fast allen, Spezies haben

3.1 Algorithmus

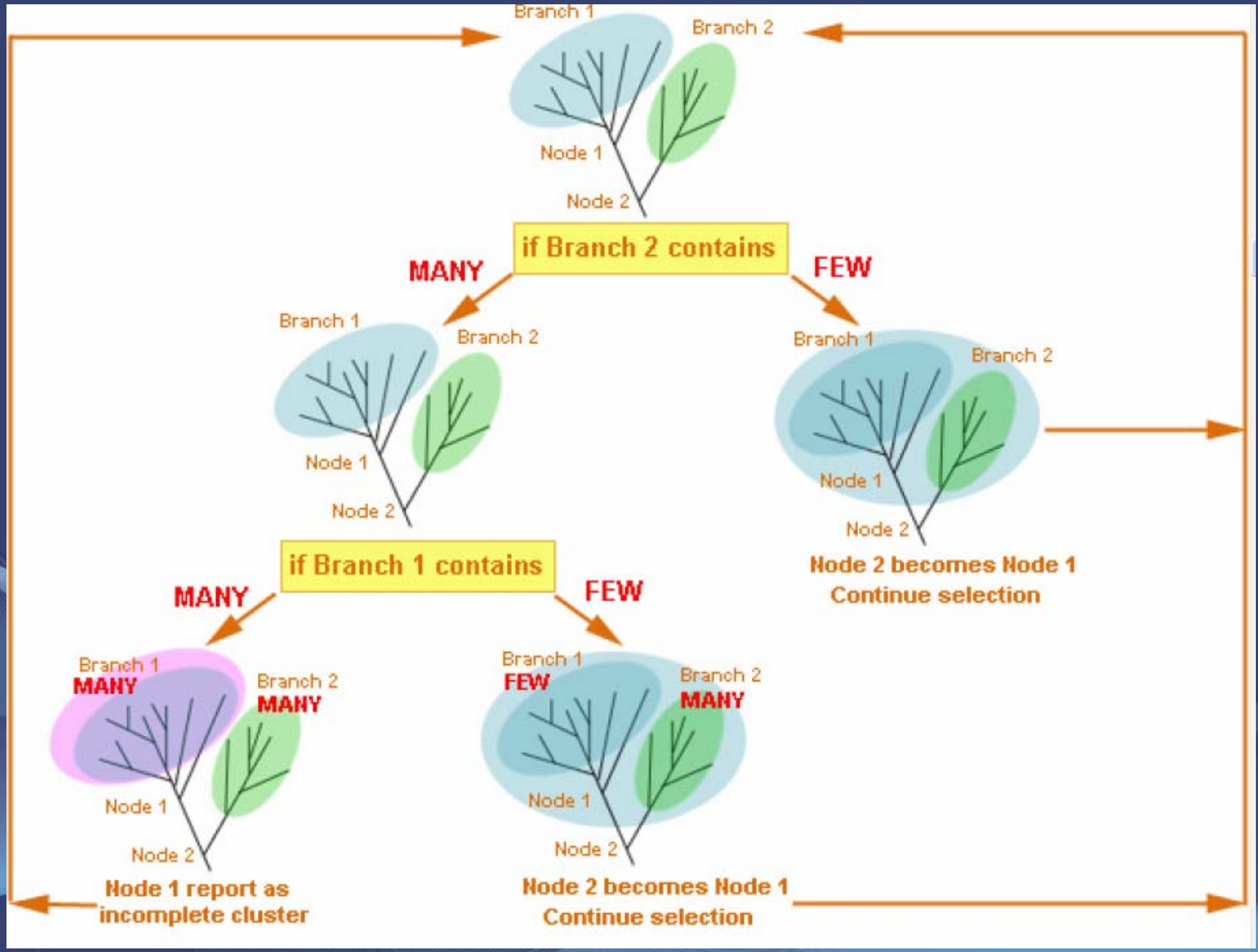
- Ziel: Erstellen von orthologen Genfamilien
- 6 Schritte:
  - 1. Genomdownload und Erstellung signifikanter Hits
  - 2. Erzeugung der taxa identification Tabelle
  - 3. Zusammenstellen der superfamilies
  - 4. Rekonstruktion der superfamily trees
  - 5. Selektion der orthologen Familien
  - 6. Visualisierung mit TreeDyn
- Schritt 1 und 2:
  - Vorbereitende Schritte für Erstellung der superfamilies, detaillierte Erklärung siehe Tutorial

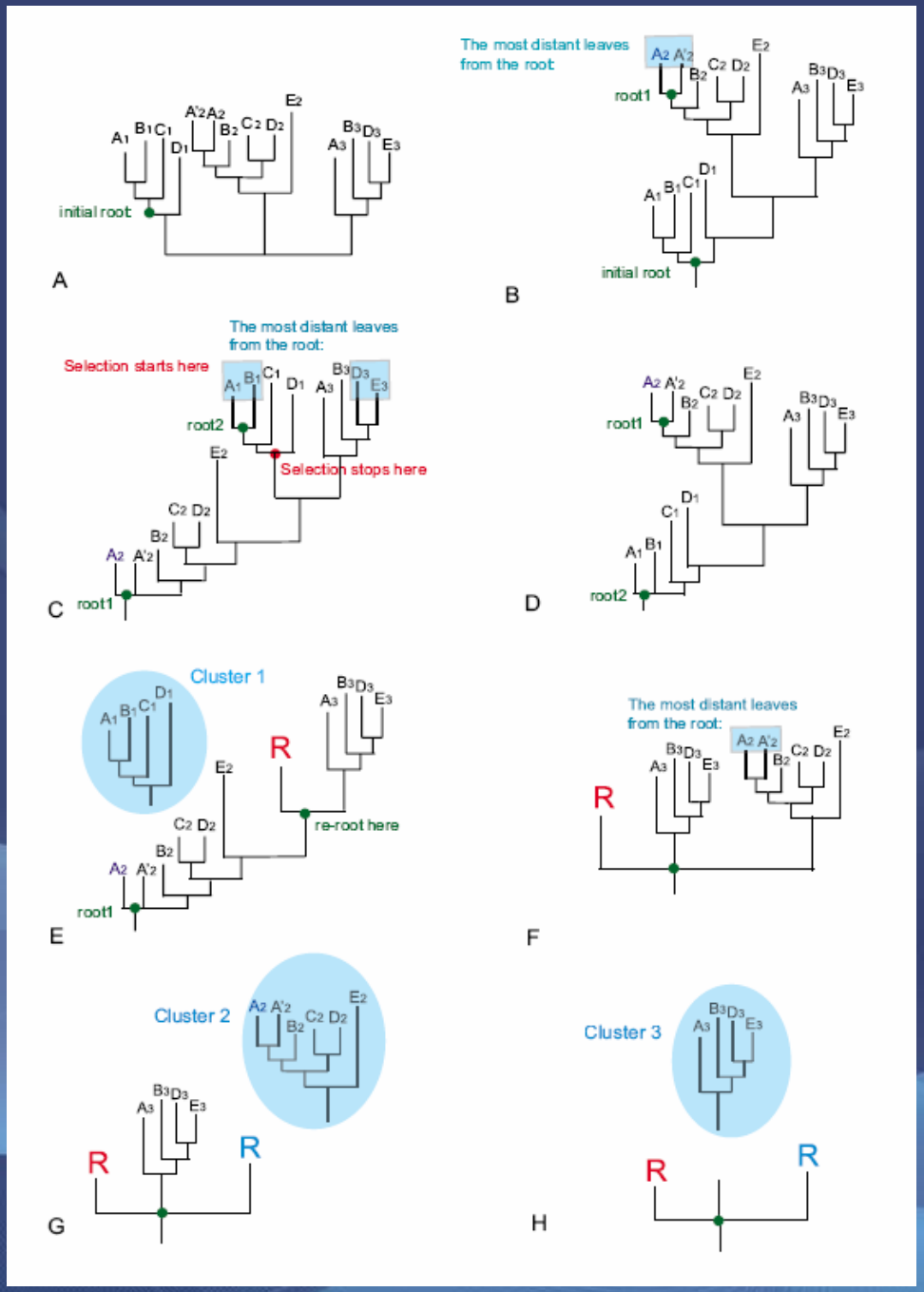
- Schritt 3:
  - n verschiedene Genome von n verschieden Spezies
  - Alle Genome in einem Datenset kombinieren
  - Jedes Gen jeder Spezies wird jetzt gegen das komplette Genom aller Spezies geblasted
  - Dabei werden nicht nur die best hits gespeichert, sondern alle signifikanten Hits -> jede Spezies kann mehrere Arten von Homologen beitragen
  - Signifikante Hits für jedes Gen von verschiedenen Spezies werden in einer Superfamilie zusammengefasst
  - Sehr zeitintensiv

- Schritt 4:
  - Die Sequenzen der superfamilies werden aligned und ein phylogenetischer Stammbaum wird erstellt
  - Dabei kann eine beliebige Methode angewandt werden (default: clustalw 1.83)

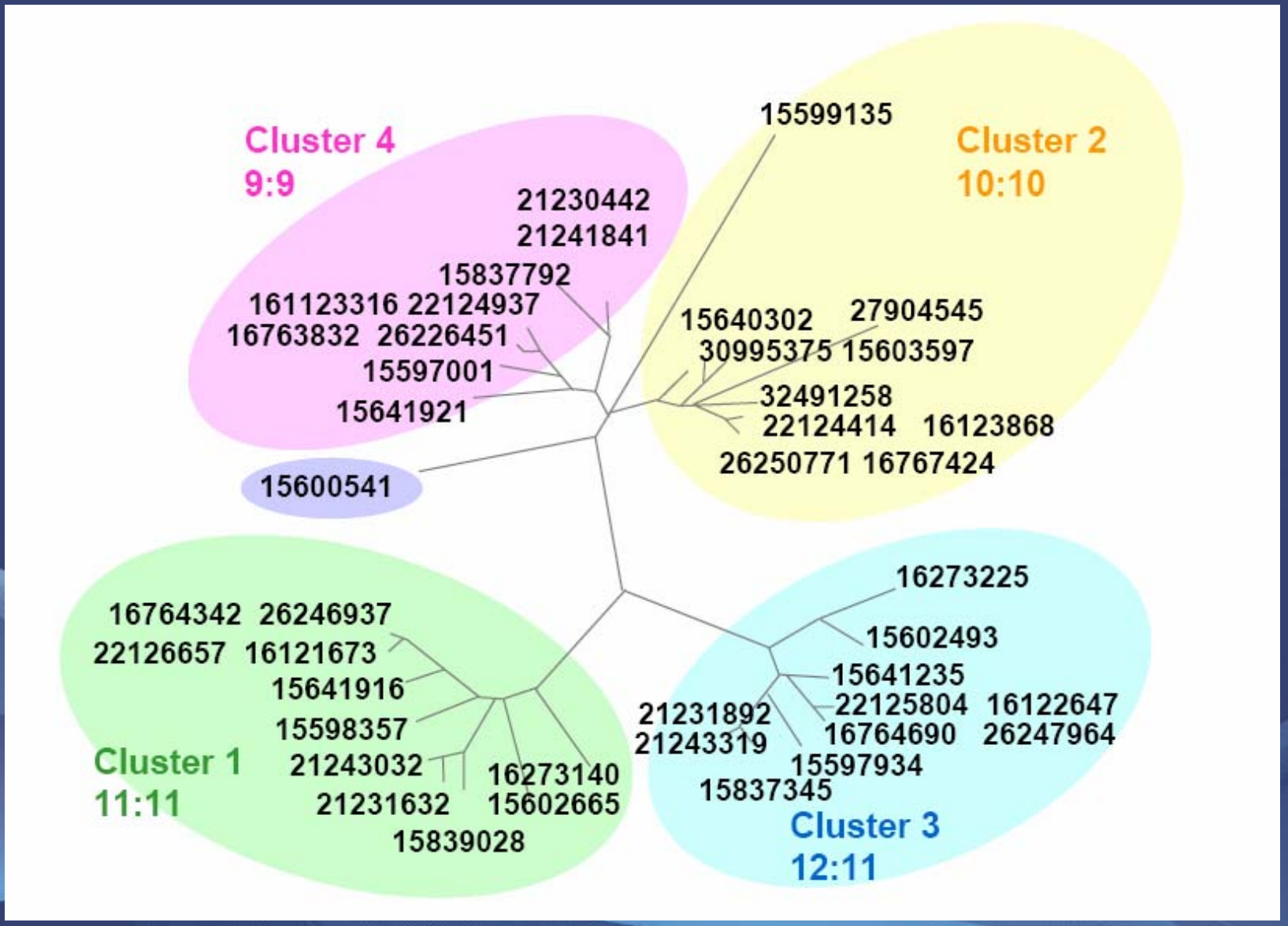
- Schritt 5:
  - Nun wird die eigentliche Grundidee des Algorithmus umgesetzt
  - Schrittweise werden Zweige zu einer bestehenden Familie hinzugefügt, solange, bis dieser Zweig Gene von allen oder fast allen beteiligten Organismen enthält
  - Dabei spielt der einzige Parameter des Programms eine große Rolle:
    - MANY Parameter entscheidet, wann ein Zweig als Stopper fungiert und genügend Spezies enthält, um als eigenes Cluster klassifiziert zu werden
  - Dabei wird am Anfang eine Wurzel zufällig gesetzt
  - Zur Artefaktvermeidung wegen der beliebig gesetzten Wurzel wird der Knoten, der am weitesten von der Wurzel entfernt ist, in einem 2. Durchlauf als Wurzel gesetzt
  - Beide Durchläufe werden dann verglichen und derjenige mit der minimalen Anzahl an Paralogen wird ausgewählt
  - Nachdem ein Cluster isoliert wurde, wird ein repräsentatives Gen ausgewählt von jeder Spezies ausgewählt und in einem output File gespeichert
  - Zudem können Gene als Inparalogs oder Outparalogs klassifiziert werden, je nach Positition zu dem Zweig, der das Cluster enthält
    - Outparalog: Gen gehört zur superfamilie, ist aber nicht in dem Cluster enthalten
    - Inparalog: Gen gehört zur superfamilie, und ist auch in dem Cluster enthalten
  - Der eigentliche Algorithmus ist eine Rekursion

- NEW SELECTION:
- while (Tree T has leaves):
  - INITIATION:Find the leaf most distant from the current, arbitrarily selected root and set Node 1 as the ancestor of the most distant leaf.
  - RECURSION:
    - check if the ancestor of the Node1 has a "stopper", leaf "R", that signifies previously removed cluster.
    - if (ancestor of the Node 1 has leaf "R") ->
      - Branch 1 contains a cluster, report an incomplete cluster, remove it, mark the Node 1 with a leaf "R", re-root the tree with cluster's ancestor and go to NEW SELECTION:
    - calculate number of different taxa on the Branch 1: n1
    - calculate number of different taxa on the Branch 2: n2
    - calculate total number of different taxa on the Node 2:n3
    - if (n1 > = n) ->
      - Node 1 is complete, report a complete cluster, remove it from the tree, mark the Node 1 with a leaf "R", re-root the tree with cluster's ancestor and go to NEW SELECTION:
    - else – Node 1 is incomplete, check the state of the Node 2
      - case (Branch 1, Branch 2) contains combinations:
        - » ((FEW, FEW), (FEW, MANY), (MANY, FEW)) and (n3 < n) -> Node 2 becomes Node 1, go to RECURSION.
        - » ((FEW, FEW), (FEW, MANY), (MANY, FEW)) and (n3 > = n) -> Node 2 is complete, report a complete cluster, remove it, mark the Node 2 with a leaf "R", re-root the tree with cluster's ancestor and go to NEW SELECTION:
        - » (MANY, MANY) -> Branch 1 contains a cluster, report an incomplete cluster, remove it, mark the Node 1 with a leaf "R", re-root the tree with cluster's ancestor and go to NEW SELECTION:
  - END OF RECURSION

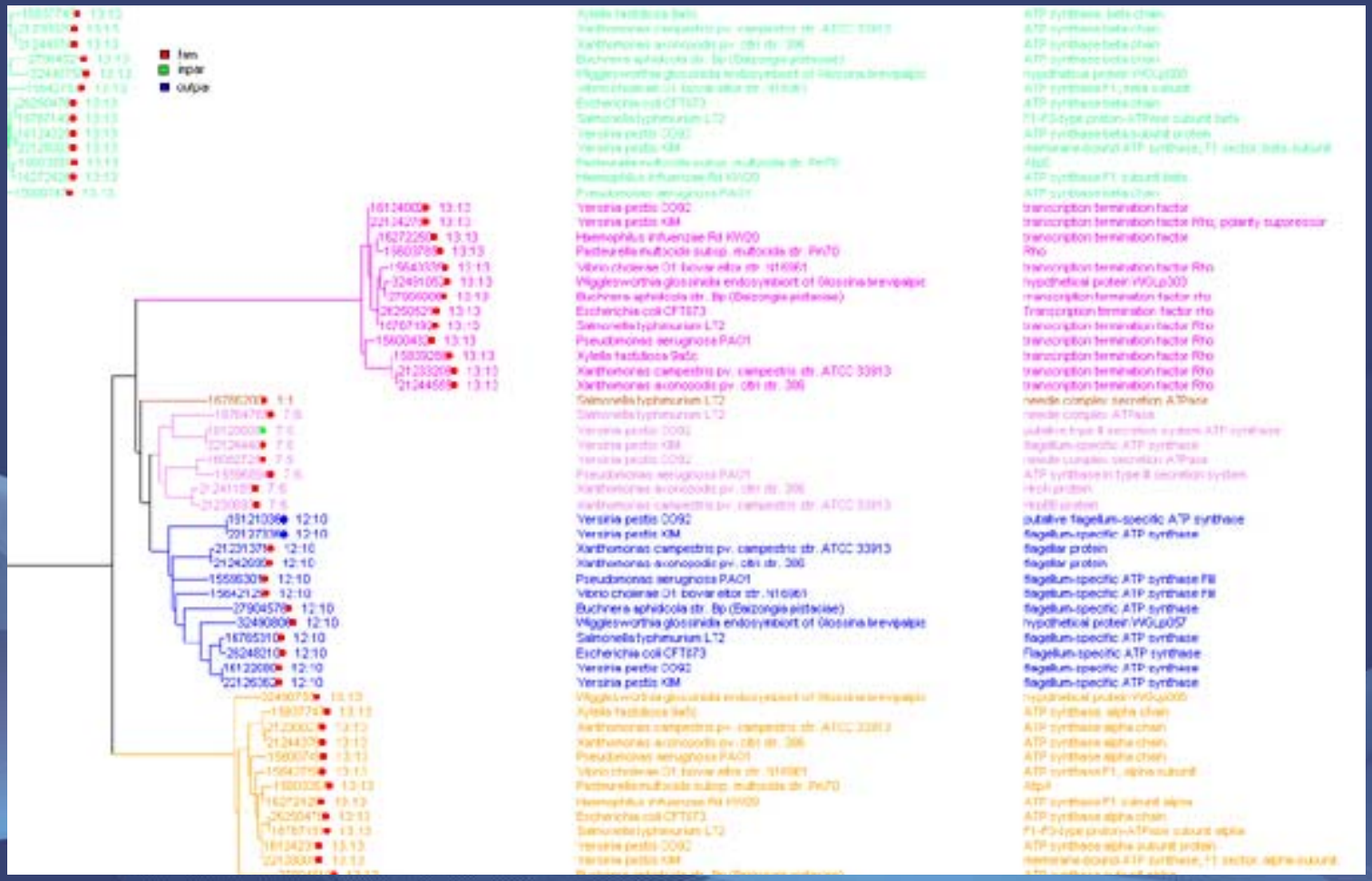

Branch 1
Branch 2
Node 1
Node 2
if Branch 2 contains
MANY
FEW
Branch 1
Branch 2
Node 1
Node 2
Branch 1
Branch 2
Node 1
Node 2
Node 2 becomes Node 1
Continue selection
if Branch 1 contains
MANY
FEW
Branch 1
MANY
Branch 2
MANY
Node 1
Node 2
Node 1 report as
incomplete cluster
Branch 1
FEW
Branch 2
MANY
Node 1
Node 2
Node 2 becomes Node 1
Continue selection


initial root
A
The most distant leaves from the root
root1
initial root
B
The most distant leaves from the root:
Selection starts here
root2
Selection stops here
root1
C
root1
root2
D
Cluster 1
R
re-root here
root1
E
The most distant leaves from the root:
R
F
Cluster 2
R
R
G
Cluster 3
R
R
H

```
---------- CLUSTER 1 ----------
---------- FAMILY ------------
>gi|27904705| peptidoglycan synthetase FtsI [Buchnera aphidicola str. Bp (Baizongia
              pistaciae)]
>gi|26246017| Peptidoglycan synthetase ftsI precursor [Escherichia coli CFT073]
>gi|16273058| penicillin-binding protein 3 [Haemophilus influenzae Rd KW20]
>gi|15602001| FtsI [Pasteurella multocida subsp. multocida str. Pm70]
>gi|15599614| penicillin-binding protein 3 [Pseudomonas aeruginosa PAO1]
>gi|16763512| division specific transpeptidase [Salmonella typhimurium LT2]
>gi|15642404| penicillin-binding protein 3 [Vibrio cholerae O1 biovar eltor str. N16961]
>gi|32490961| hypothetical protein WGLp212 [Wigglesworthia glossinidia endosymbiont of
              Glossina brevipalpis]
>gi|21230194| penicillin-binding protein 3 [Xanthomonas campestris pv. campestris str.
              ATCC 33913]
>gi|21241544| penicillin-binding protein 3 [Xanthomonas axonopodis pv. citri str. 306]
>gi|15837394| penicillin binding protein 3 [Xylella fastidiosa 9a5c]
>gi|16120877| penicillin-binding protein 3 [Yersinia pestis CO92]
>gi|22127506| peptidoglycan synthetase [Yersinia pestis KIM]
COMPLETE: 13

>>>>> IN-PARALOGS -----------
>gi|16765177| putative penicillin-binding protein 3 [Salmonella typhimurium LT2]
>gi|15597468| penicillin-binding protein 3A [Pseudomonas aeruginosa PAO1]


---------- CLUSTER 2 ----------
---------- FAMILY ------------
>gi|26246616| Penicillin-binding protein 2 [Escherichia coli CFT073]
>gi|16272007| penicillin-binding protein 2 [Haemophilus influenzae Rd KW20]
>gi|15603789| Pbp2 [Pasteurella multocida subsp. multocida str. Pm70]
>gi|15599198| penicillin-binding protein 2 [Pseudomonas aeruginosa PAO1]
>gi|16764017| cell elongation-specific transpeptidase [Salmonella typhimurium LT2]
>gi|15640966| penicillin-binding protein 2 [Vibrio cholerae O1 biovar eltor str. N16961]
>gi|32490921| hypothetical protein WGLp172 [Wigglesworthia glossinidia endosymbiont of
              Glossina brevipalpis]
>gi|21232896| penicillin-binding protein 2 [Xanthomonas campestris pv. campestris str.
              ATCC 33913]
>gi|21241430| penicillin-binding protein 2 [Xanthomonas axonopodis pv. citri str. 306]
>gi|15837913| penicillin binding protein 2 [Xylella fastidiosa 9a5c]
>gi|16122817| penicillin-binding protein 2 [Yersinia pestis CO92]
>gi|22125081| peptidoglycan synthetase, penicillin-binding protein 2 [Yersinia pestis KIM]
INCOMPLETE: 12

>>>>> IN-PARALOGS -----------
>gi|16765252| putative penicillin-binding protein [Salmonella typhimurium LT2]
```


Cluster 4
9:9
21230442
21241841
15837792
161123316 22124937
16763832 26226451
15597001
15641921
15599135
Cluster 2
10:10
15640302 27904545
30995375 15603597
32491258
22124414 16123868
26250771 16767424
15600541
16764342 26246937
22126657 16121673
15641916
15598357
Cluster 1
11:11
21243032
21231632
16273140
15602665
15839028
16273225
15602493
15641235
21231892 22125804 16122647
21243319 16764690 26247964
15597934
15837345
Cluster 3
12:11

- Schritt 6
  - Die Ergebnisse können mit dem Programm TreeDyn visualisiert werden
  - Die dafür benötigten Skripte und Dateien werden mitgeliefert



3.2 Bewertung

- Interessanter Ansatz, der Vorteile beider existierender Methoden vereint
- Die Zuverlässigkeit der Ergebnisse lässt sich nur subjektiv bestimmen
- Es werden keine speziellen Genevolutionen betrachtet, nur Paralogie (In- und Out) und Orthologie
- Die Phylogenien in jedem Zweig werden lediglich dazu benutzt, um zwischen Paralogen und Clustern von Orthologen zu unterscheiden
- Sehr zeitintensiv (Rechendauer kann eine Woche und länger dauern)
- Wahl des few/many Parameters beeinflusst Ergebnis wesentlich, hier kann eine Verbesserung erfolgen, indem die Parameter weniger strikt (kontextabhängig) wirken
- Probleme, wenn diverse Gruppen analysiert werden:
  - Orthologe, die nur in einer Subgruppe enthalten sind, als Outparaloge klassifiziert
  - Unterscheidung zwischen Orthologen und Paralogen ist schwierig für Gene, die einer hohen Substitutionsrate unterliegen

Number of taxa – A: Archaea B: Bacteria	Number of selected families:	
	Reciprocal best BLAST hit	BranchClust
2A2B	80	414 (all complete)
13B	236	2066 (369 complete, 1690 with n ≥ 8)
14A	125	1431 (300 complete, 1131 with n ≥ 8)
14B 16A	12	195 (80 complete, 195 with n ≥ 24)

Vergleich mit reziproker, best Blast Hit Methode

3.3 Installation und Benutzung

- In Perl geschrieben
- Konsolenbasiert, recht mühsame Installation (nicht Biologen-kompatibel), aber:
- Sehr verständliches Tutorial auf der Website erhältlich
- Benötigt:
  - BioPerl
  - Blastall
  - Clustalw
  - Optional: TreeDyn

6. Quellen

- BranchClust: a phylogenetic algorithm for selecting gene families, Maria S Poptsova and J Peter Gogarten, 2006 (*http://www.biomedcentral.com/1471-2105/8/120*)
- Orthologs, Paralogs and Evolutionary Genomics, Eugene V. Koonin, 2005 (*http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.genet.39.073003.114725*)
- Programm und Tutorial (*http://bioinformatics.org/branchclust/)*
- TreeDyn (*http://www.treedyn.org/)*
- BioPerl (*http://www.bioperl.org/wiki/Main_Page*)