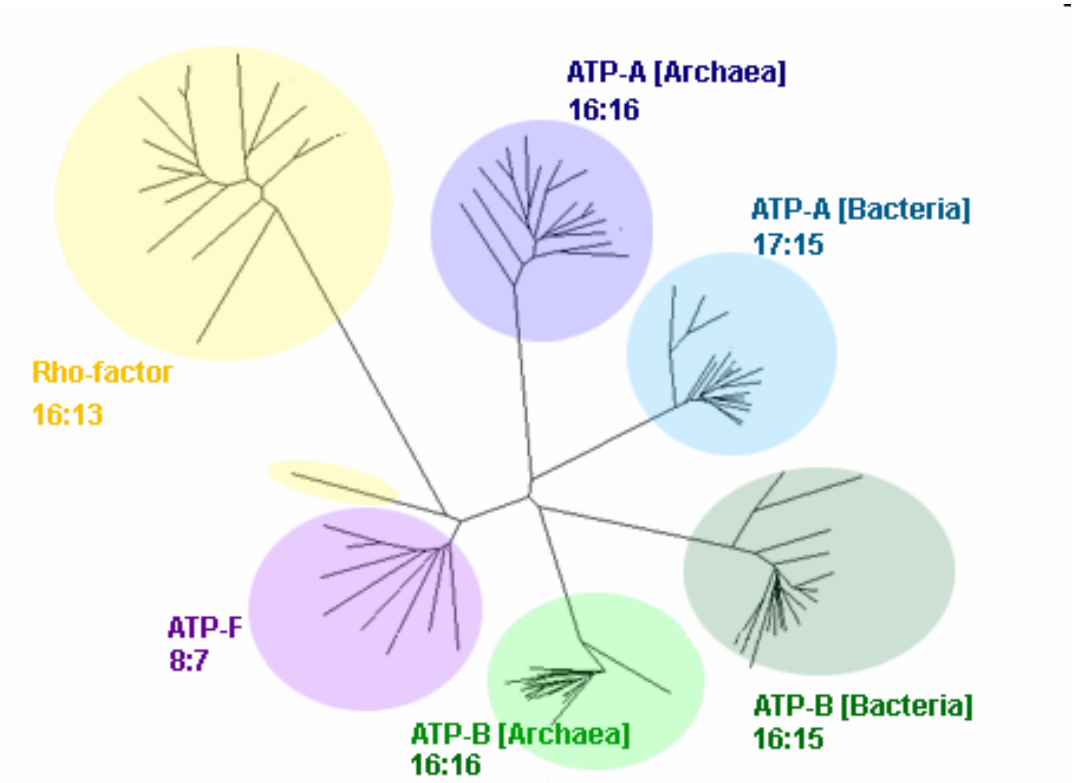
Handout zum Thema

“Rekonstruktion von Genfamilien, BranchClust”



16.07.07

Problemseminar Bioinformatik SS07

Universität Leipzig

Christian Arnold

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):
  1. Vertikale Speziation (speciation) mit Modifikation
  2. Genduplikation, gefolgt von Speziation
  3. Genverlust (gene loss)
  4. Horizontaler Gentransfer (HGT)
  5. 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
- **Ortologie:**
  - **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 (**siehe Bildanhang am Ende, da findet man Beispiele**):**
  - **Co-Orthologs (**zwei oder mehr Gene in einer Abstammung, die aufgrund Duplikation zusammen ortholog zu einem oder mehren Genen einer anderen Linie sind)
  - **Pseudoorthologs** (Gene, die eigentlich paralog sind, aber wegen abstammungs-spezifischen Genverlust ortholog erscheinen)
  - **Outparalogs (**Paraloge Gene, deren Duplikation **vor** einem Speziationsereignis stattgefunden hat)
  - **Inparalogs** (Paraloge Gene, deren Duplikation **nach** einem Speziationsereignis stattgefunden hat)
  - **Pseudoparalogs** (Gene, die paralog erscheinen, in Wirklichkeit aber durch HGT und vertikalen Gentransfer entstanden sind)
  - **Xenologs** (Xenologous gene displacement, Austausch eines Gens in einer bestimtmen Abstammung durch ein Gen aus demselben orthologen Cluster aus einer weiter entfernten Abstammung)

- **Horizontaler Gentransfer (HGT)**
  - auch lateraler Gentransfer (LTG)
  - Übertragung von Genen außerhalb der geschlechtlichen Fortpflanzung und über Artgrenzen hinweg
  - bildet in der Evolutionstheorie eine Möglichkeit zur Erklärung von Sprüngen in der Entwicklung vor allem von Prokaryoten
  - relativ häufig, deshalb kann meist kein exakter Spezies tree angegeben werden, sondern höchstens ein Konsensus tree
  - wird von Programmen bisher ignoriert
  - kann geschehen durch:
    - Tranformation
    - Transfektion
    - Konjugation

- Im wesentlichen 2 Methoden, um Genfamilien zu ermitteln:
  - Tree reconciliation (Baumabgleich)
  - Genome-specific best Hits (SymBets)

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 „incongruences“ nur durch Duplikation und Deletionen erklärt, aber nicht durch HGTs

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

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 best Hits, 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)**

- Strikte Reziproke Natur wird entschärft und durch eine triangulare best Blast Hit Methode ersetzt (Verbesserung des Inparalogs Problems)
- Algorithmus findet Cluster von best Hits, in denen aber nicht zwischen In- und Outparalogen und Orthologen unterschieden wird
- Funktioniert gut bei Abstammungen, die nur wenige Inparaloge aufweisen
- Ist auf eine beschränkte Anzahl von Spezies beschränkt

- HOBAC-GEN
- INPARANOID

3. BranchClust

- Kombiniert Ansätze beider Methoden:
  - Tree reconciliation: Baum wird erstellt, der orthologe Cluster 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 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
- Ziel: Erstellen von orthologen Genfamilien
- **6 Schritte:**
  - Genomdownload und Erstellung signifikanter Hits
  - Erzeugung der taxa identification Tabelle
  - Zusammenstellen der superfamilies
  - Rekonstruktion der superfamily trees
  - Selektion der orthologen Familien
  - 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 Posititon 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

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

3.2 Diskussion

- 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

- Im Vergleich zur reziproken, best Blast Hit Methode wesentlich höhere Anzahl an gefundenen Clustern

3.3 Installation und Bedienung

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

4. Quellen

- BranchClust: a phylogenetic algorithm for selecting gene families, Maria S Poptsova and J Peter Gogarten, BMC Bioinformatics 2007, 8:120 , 10 April 2007 (http://www.biomedcentral.com/1471-2105/8/120)
- Orthologs, Paralogs and Evolutionary Genomics, Eugene V. Koonin, Annual Review of Genetics 39:309-38, August 30, 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)

5. Bildanhang

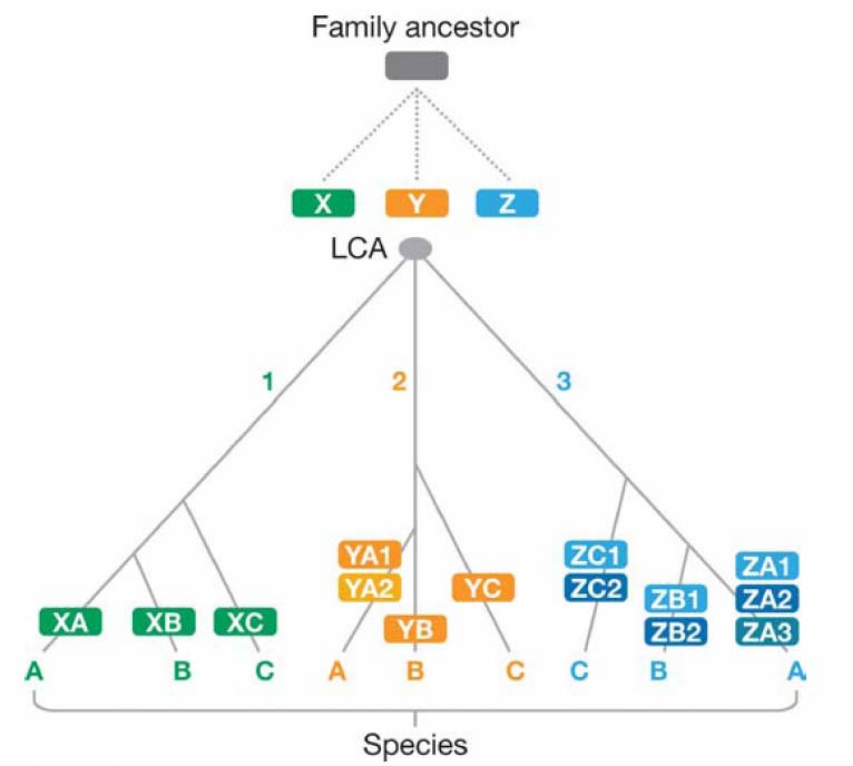


Figure 2

A hypothetical phylogenetic tree illustrating orthologous and paralogous relationships between three ancestral genes and their descendants in three species. LCA, last common ancestor (of the compared species).

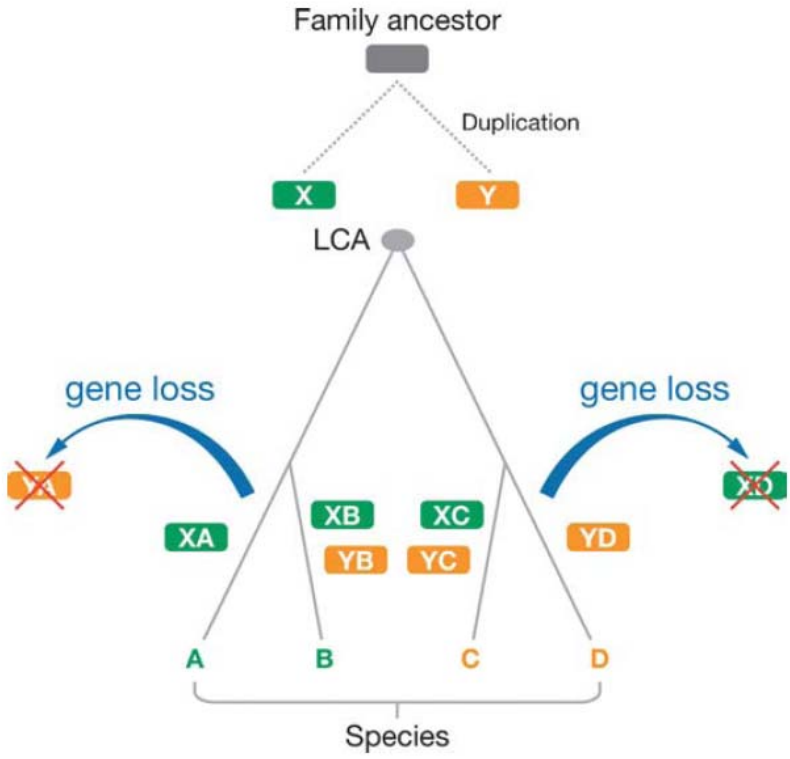


Figure 3

A hypothetical phylogenetic tree illustrating emergence of pseudoorthologs via lineage-specific gene loss.

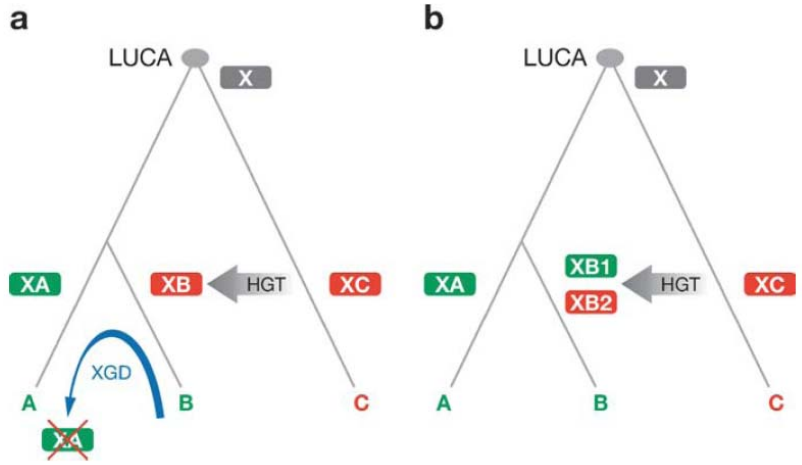


Figure 4
Effect of horizontal gene transfer on orthology and paralogy. (*a*) A hypothetical evolutionary scenario with HGT leading to xenology. (*b*) A hypothetical evolutionary scenario with HGT leading to pseudoparalogy. LUCA, Last Universal Common Ancestor (of all extant life forms).

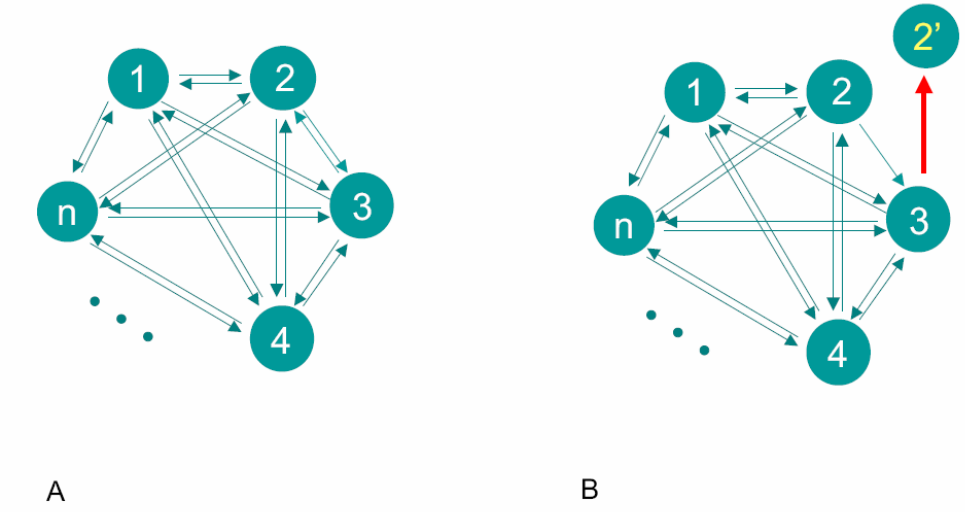


Figure 1
The reciprocal best BLAST hit method. Circles represent genes from n different taxa, arrows signify best BLAST hit relationship; (A) – case of strict reciprocity, (B) – failing of reciprocity in the presence of paralogs.

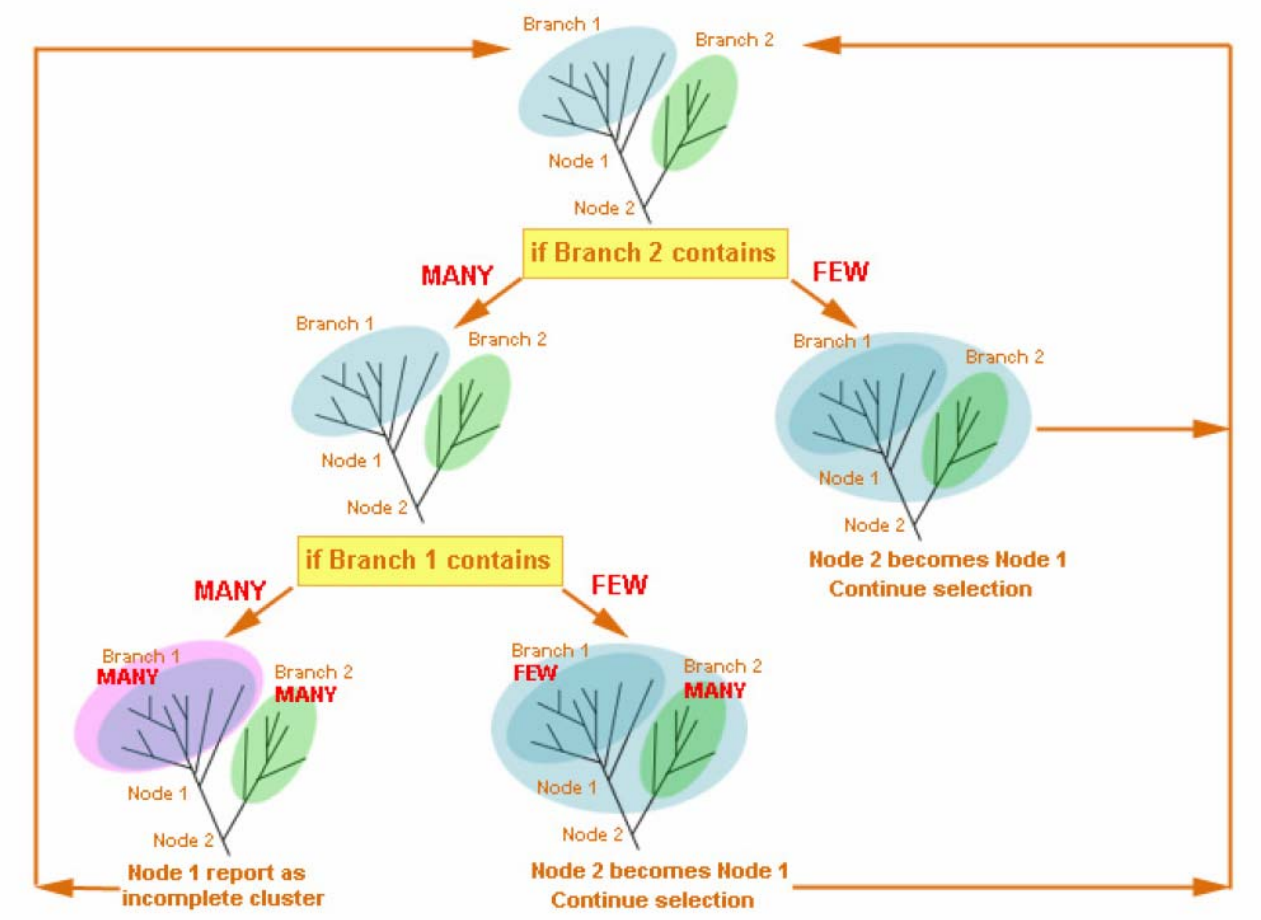


Figure 4
Schematic representation of the BranchClust algorithm.

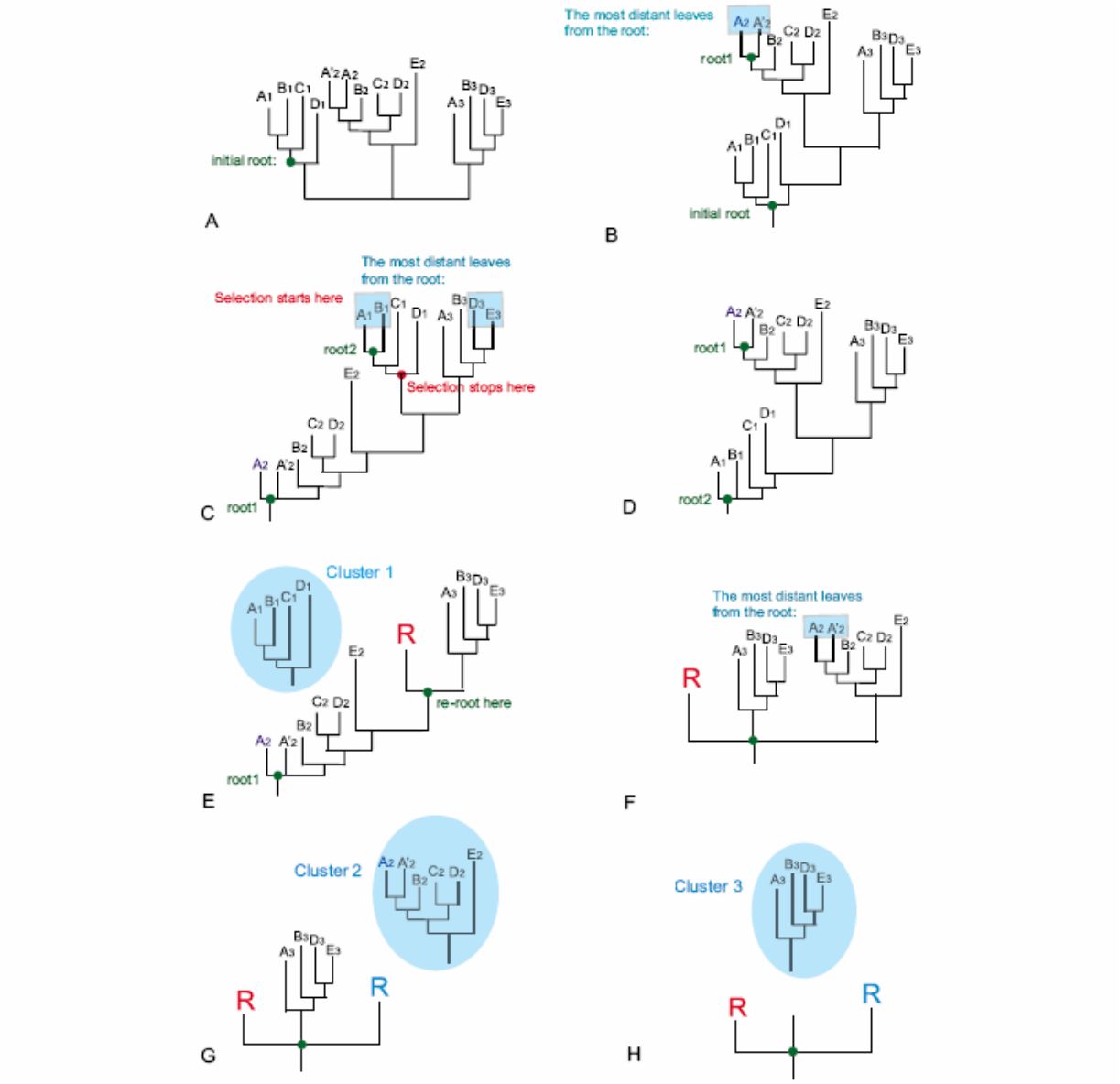


Figure 5

Example of BranchClust selection steps for a superfamily tree for 5 different taxa with 3 clusters. Figure 5A. Unrooted original tree. Figure 5B. Tree is initially rooted inside the cluster. The ancestor of the most distant leaf from the root is set to be root 1. Figure 5C. Tree is re-rooted with the root 1. The ancestor of the most distant leaf from the root 1 is set to be root 2. Selection starts from the most distant from the root 1 leaf and ends when it encounters branch containing MANY (here 4) species. First cluster is selected and removed from the tree. The node is marked with a leaf "R". Figure 5D. Tree is re-rooted with the root 2. Figure 5E. Tree is re-rooted at the ancestor of the selected cluster and the selection continues. Figure 5F. Cluster 2 is selected, removed from the tree, the node of cut is marked with a leaf "R". Figure 5G. Tree is not re-rooted because the ancestor of the removed cluster is already the root. Selection continues. The last cluster is selected. End of selection.