Gene Duplication, Loss and Transfer

Paweł Górecki

MIM UW

March, 2014

Species trees

Definition

A *species tree* is a rooted binary tree. Leaves of the species tree are called *species*.

Species trees

Definition

A *species tree* is a rooted binary tree. Leaves of the species tree are called *species*.

Definition

A *gene tree* over a species tree S is a rooted binary tree whose leaves are labeled by *species* present in S (i.e. by the leaves of S).

Species trees

Definition

A *species tree* is a rooted binary tree. Leaves of the species tree are called *species*.

Definition

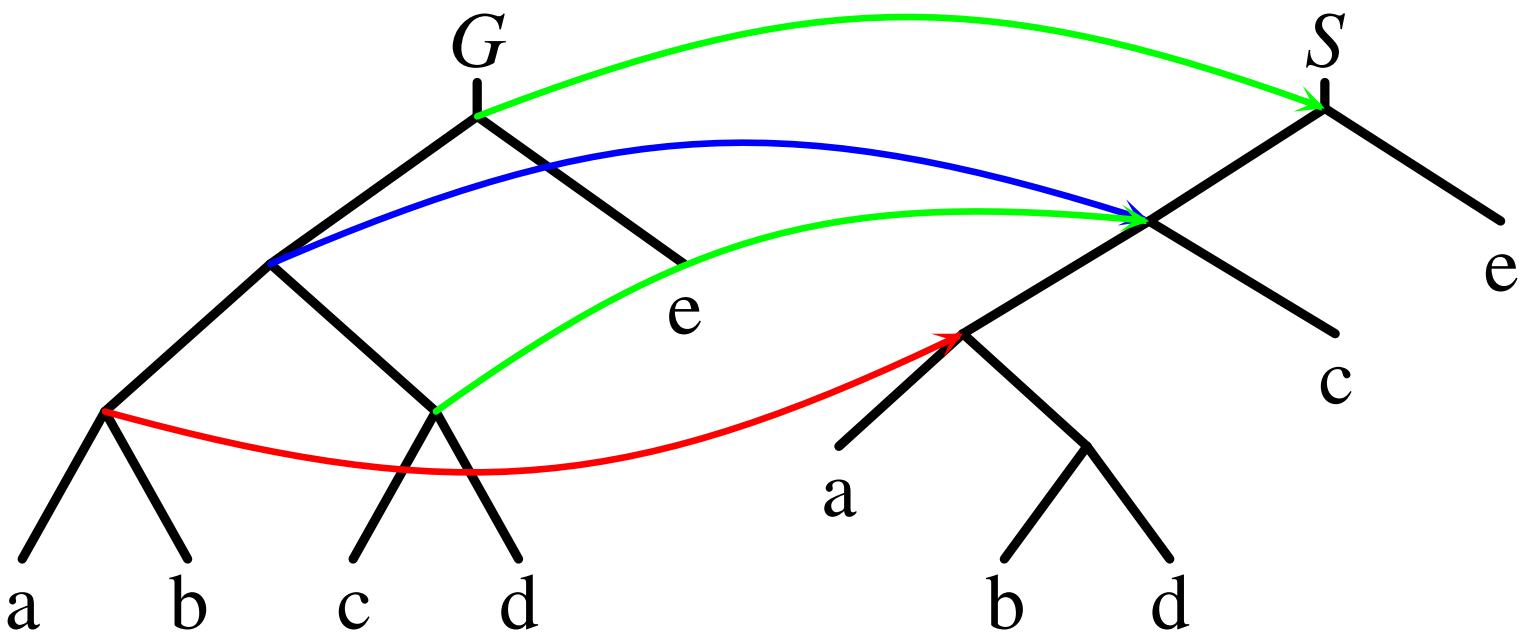
A *gene tree* over a species tree S is a rooted binary tree whose leaves are labeled by *species* present in S (i.e. by the leaves of S).

Formally, if G is a gene tree, then $G = \langle V, E, \Lambda \rangle$, where Λ is the labeling function (from the leaves of G into the leaves of S). $\langle V, E \rangle$ is called *the shape* of G.

LCA mapping

Definition

Given a species tree G and a gene tree over S, let M be the least common ancestor mapping from the nodes of G to the nodes of S that preserves the labels of leaves.



Deep coalescence

Definition

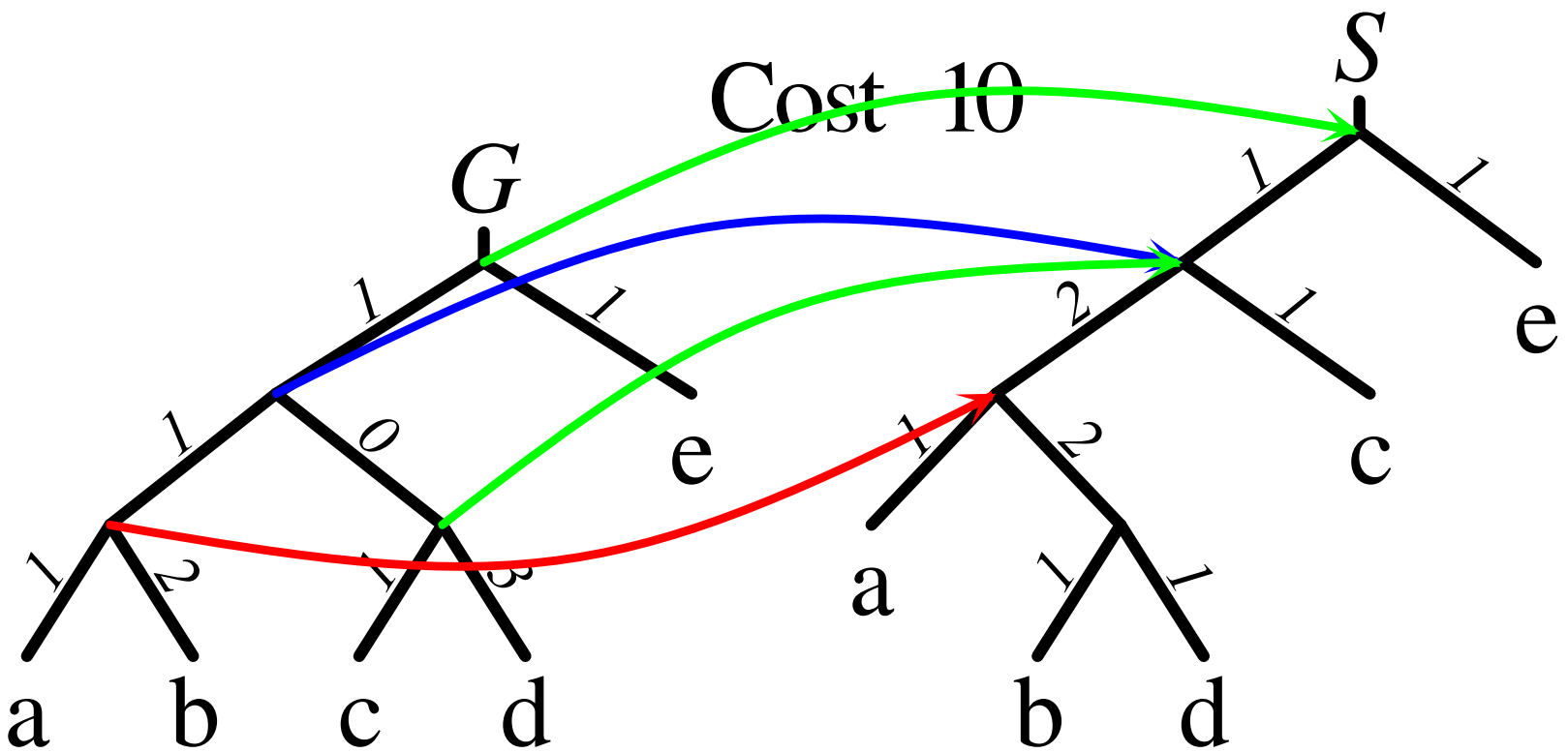
The Deep-Coalescence (DC) score between a gene tree G and a a species tree S is defined as follows

$$DC(T,S) := \sum_{g \in V_G \setminus \{\rho G\}} \|Mg, M\pi g\|, \tag{1}$$

where ρG is the root of G, πg is the parent of g, M is the lca-mapping, and $\|x, y\|$ denotes the number of edges on the path connecting x and y in S.

This definition is adapted from Maddison, 1997.
Notation: $M(g) = Mg$, $M\pi g = M(\pi g)$, etc.

DC example

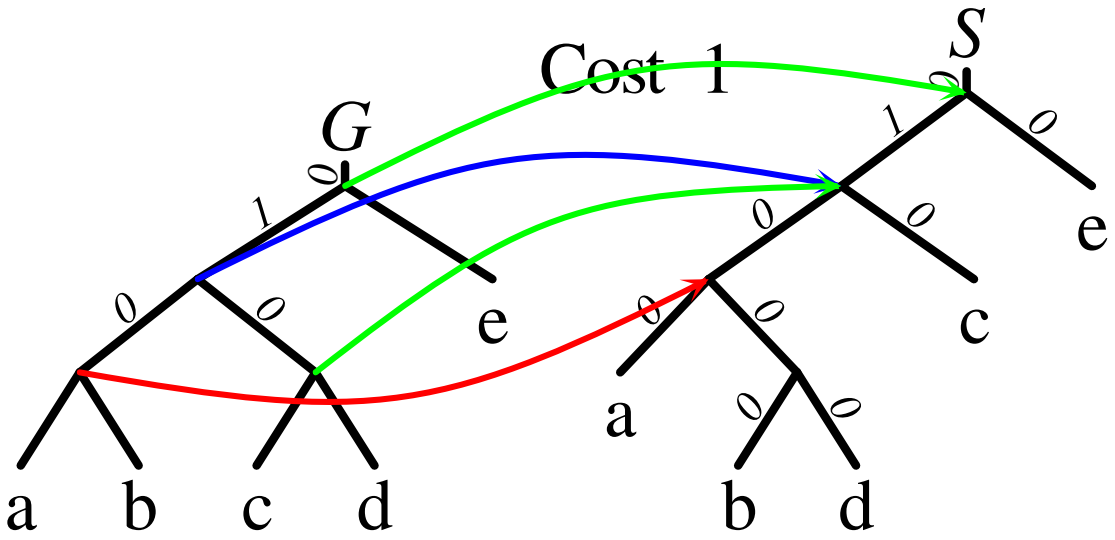


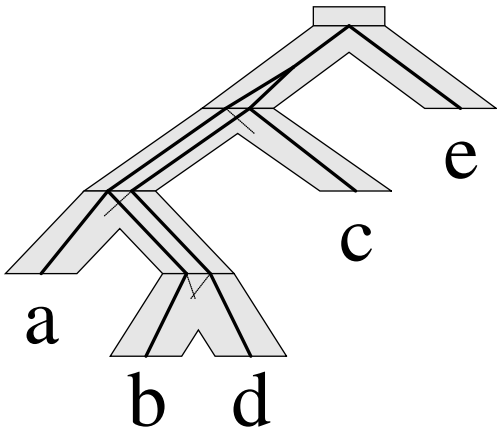
Duplication cost

Definition

An internal node g of a gene tree G over a species tree S is called a *duplication* if $Mg = Mc$, where c is a child of g.
The duplication cost (D) between a species tree S and a gene tree over S is the total number of duplication nodes present in G.

Duplication cost example





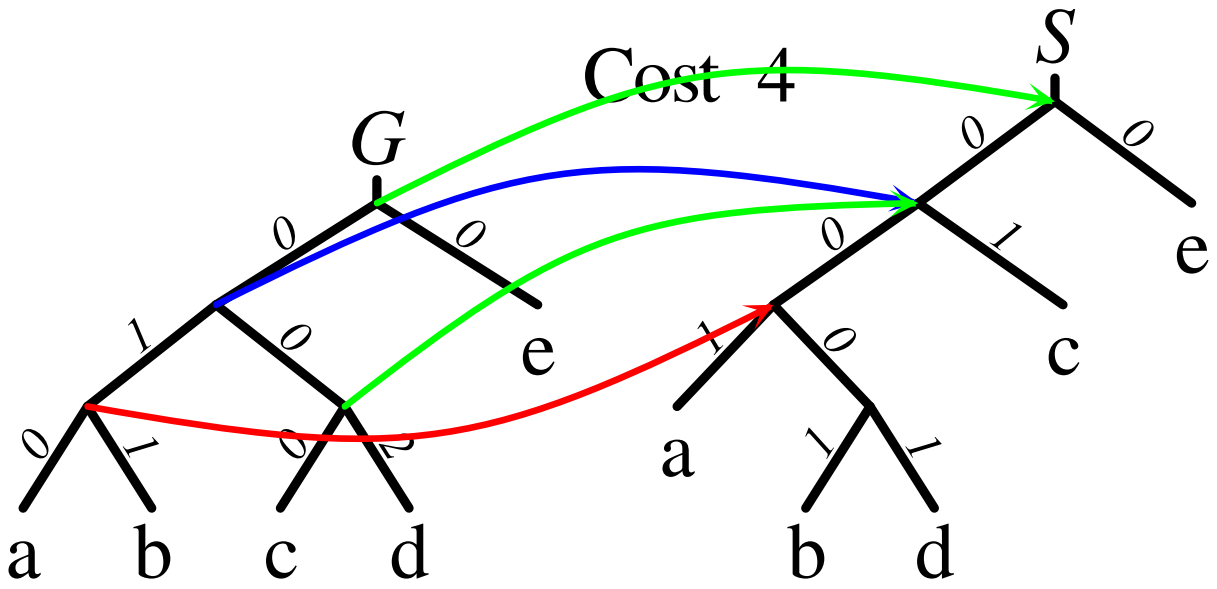
Gene loss cost

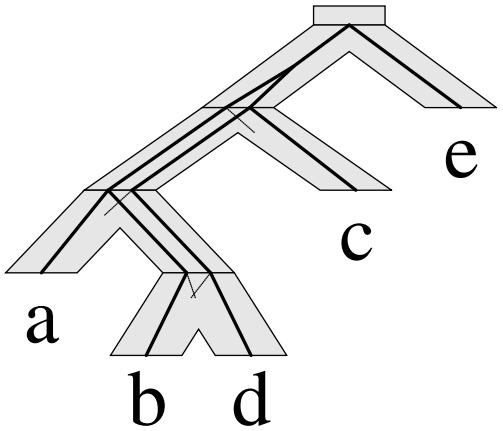
Definition

The number of gene losses required to reconcile G and S is given by:

$$L(G,S) := DC(G,S) + 2 * D(G,S) - |V_G| + 1.$$

Loss cost example

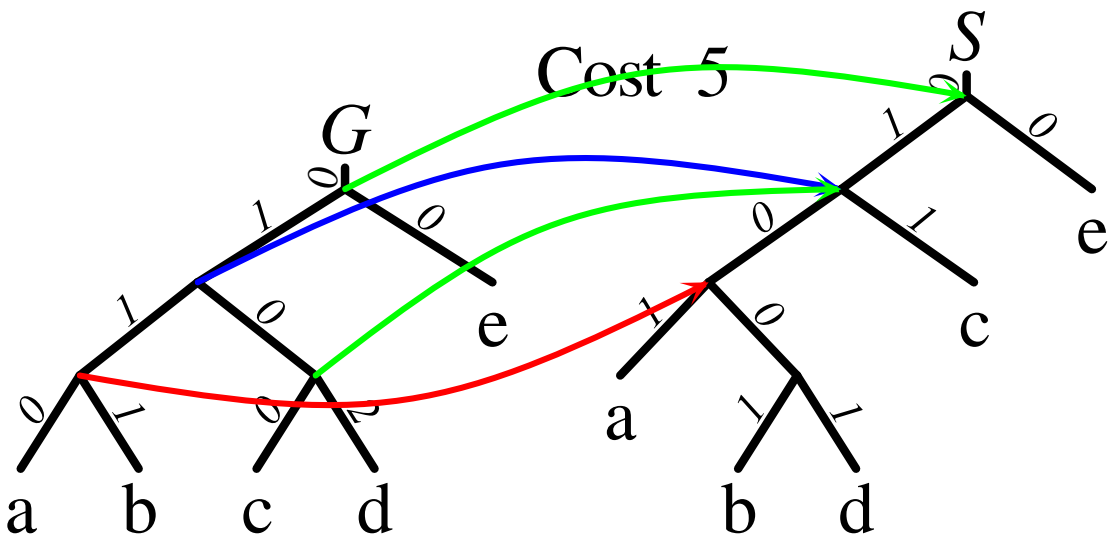


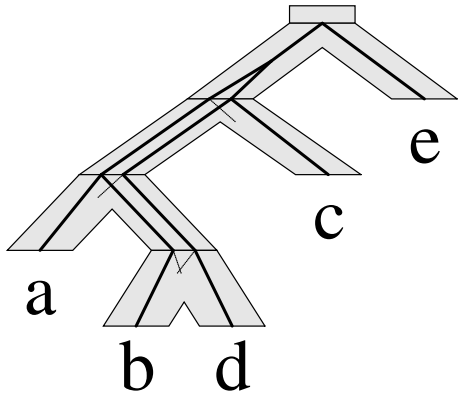


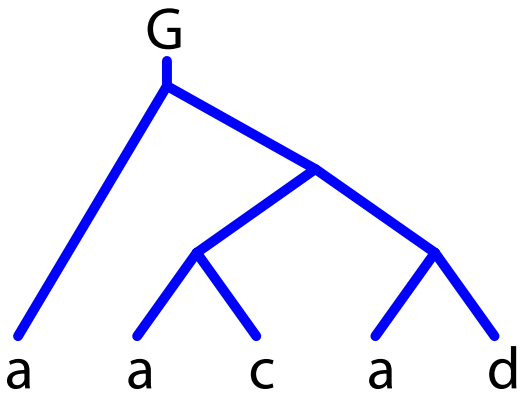
In our example: 10+2*1-9+1=4.

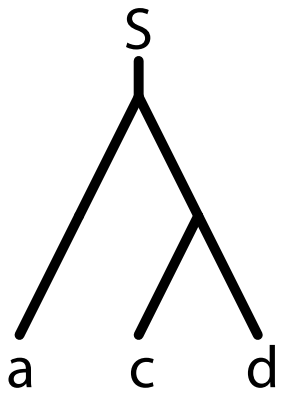
Definition

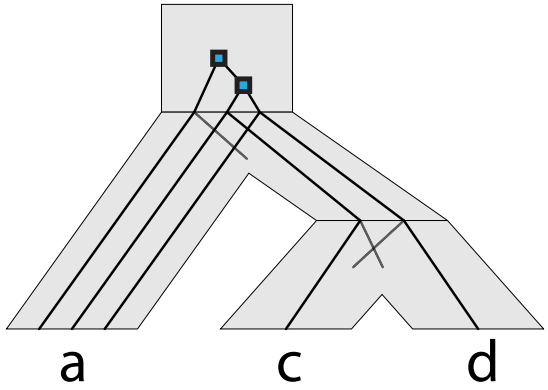
$$DL(G,S) := D(G,S) + L(G,S).$$

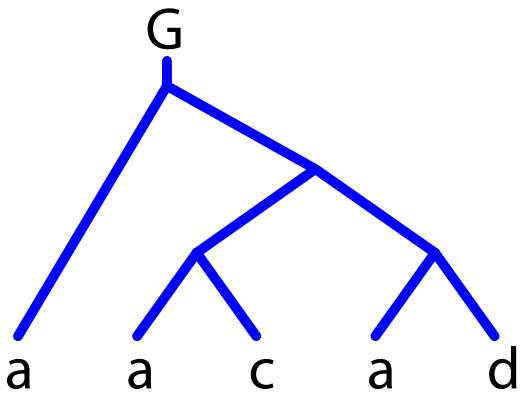


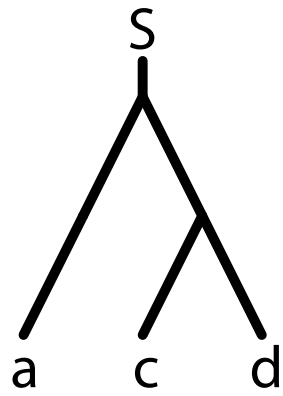


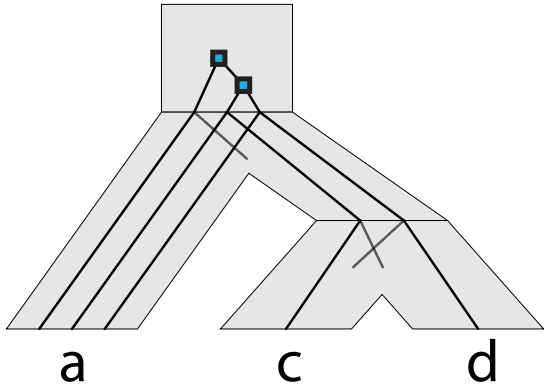

G
a
a
c
a
d

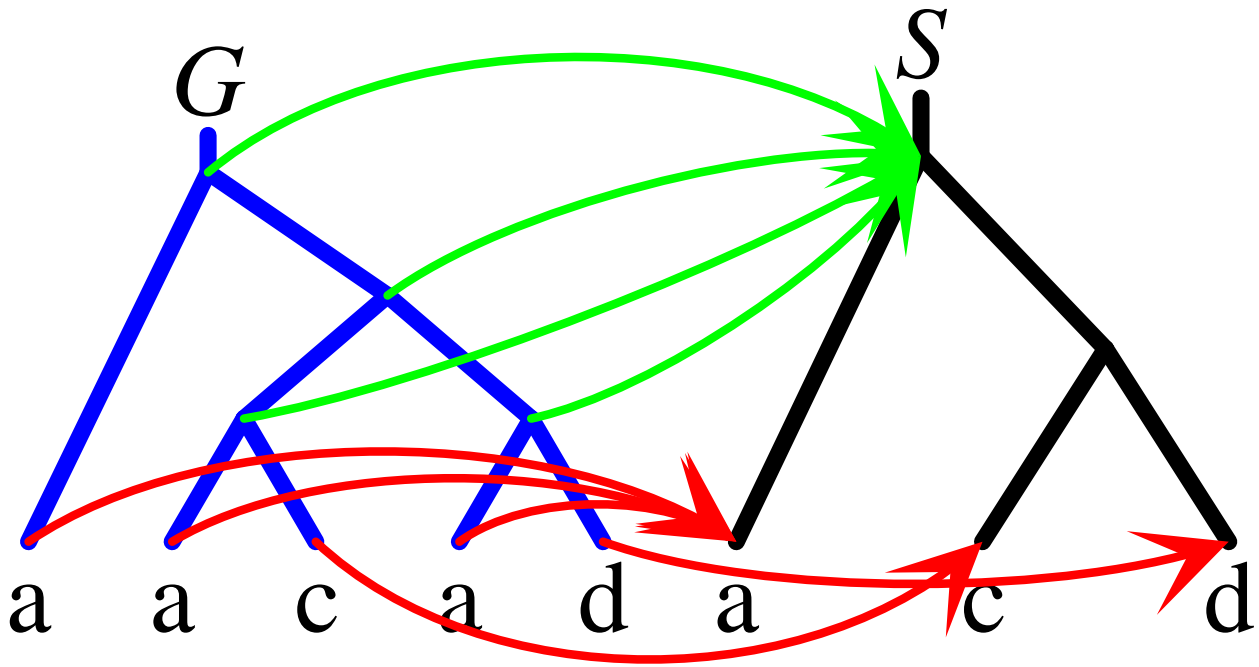

S
a
c
d

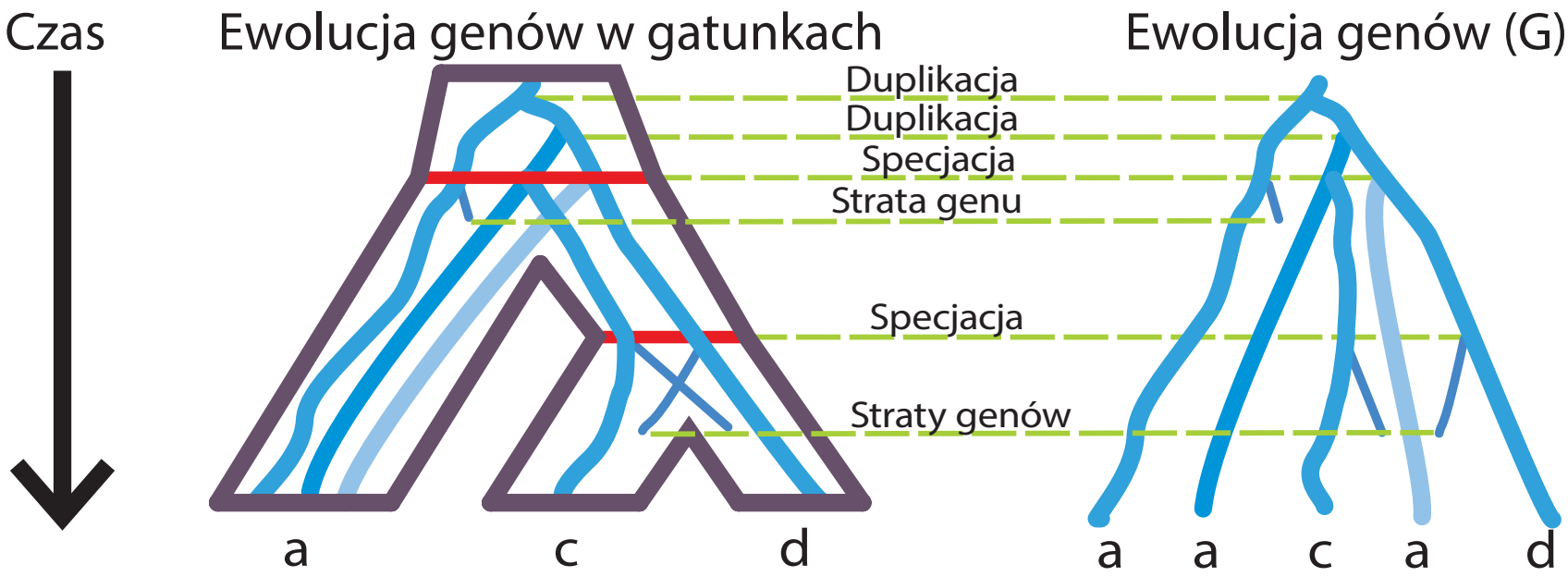

a
c
d

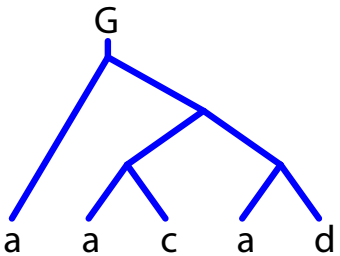

G
a
a
c
a
d

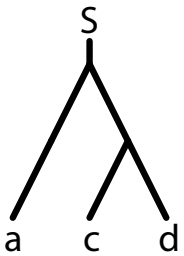

S
a
c
d

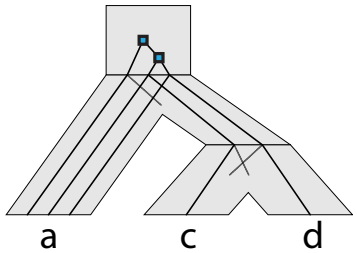

a
c
d

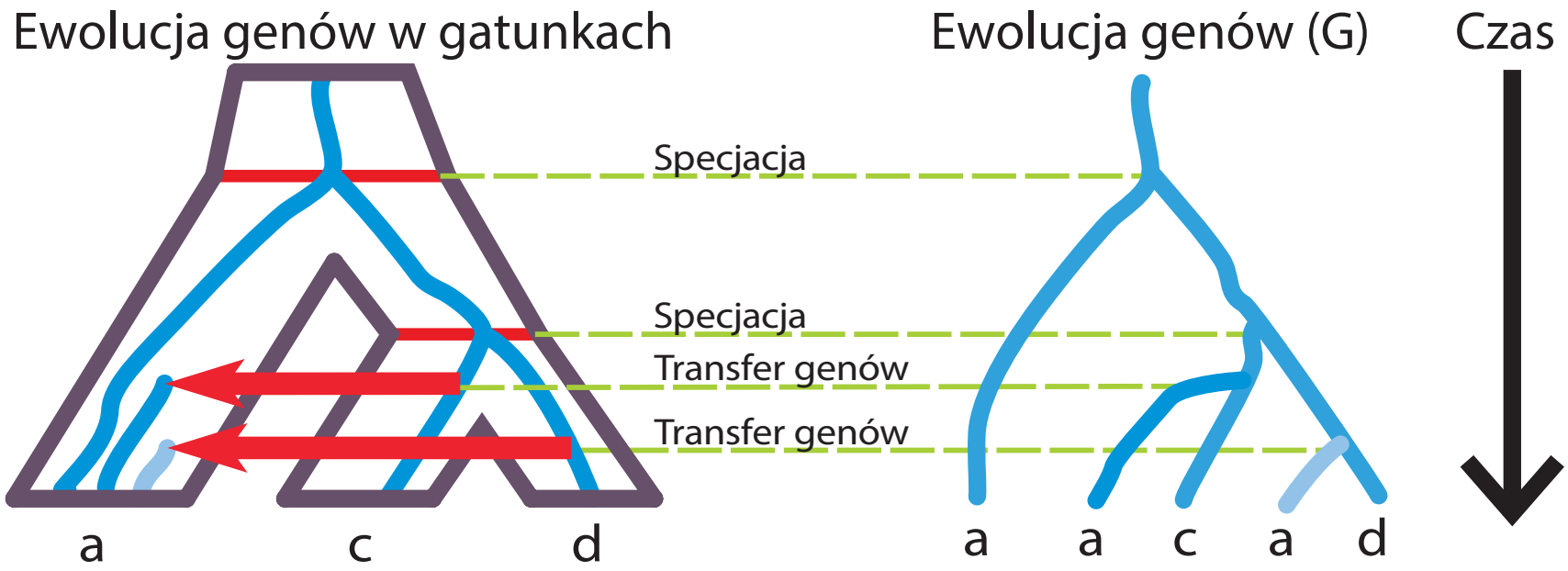

G
S
a
a
c
a
d
a
c
d

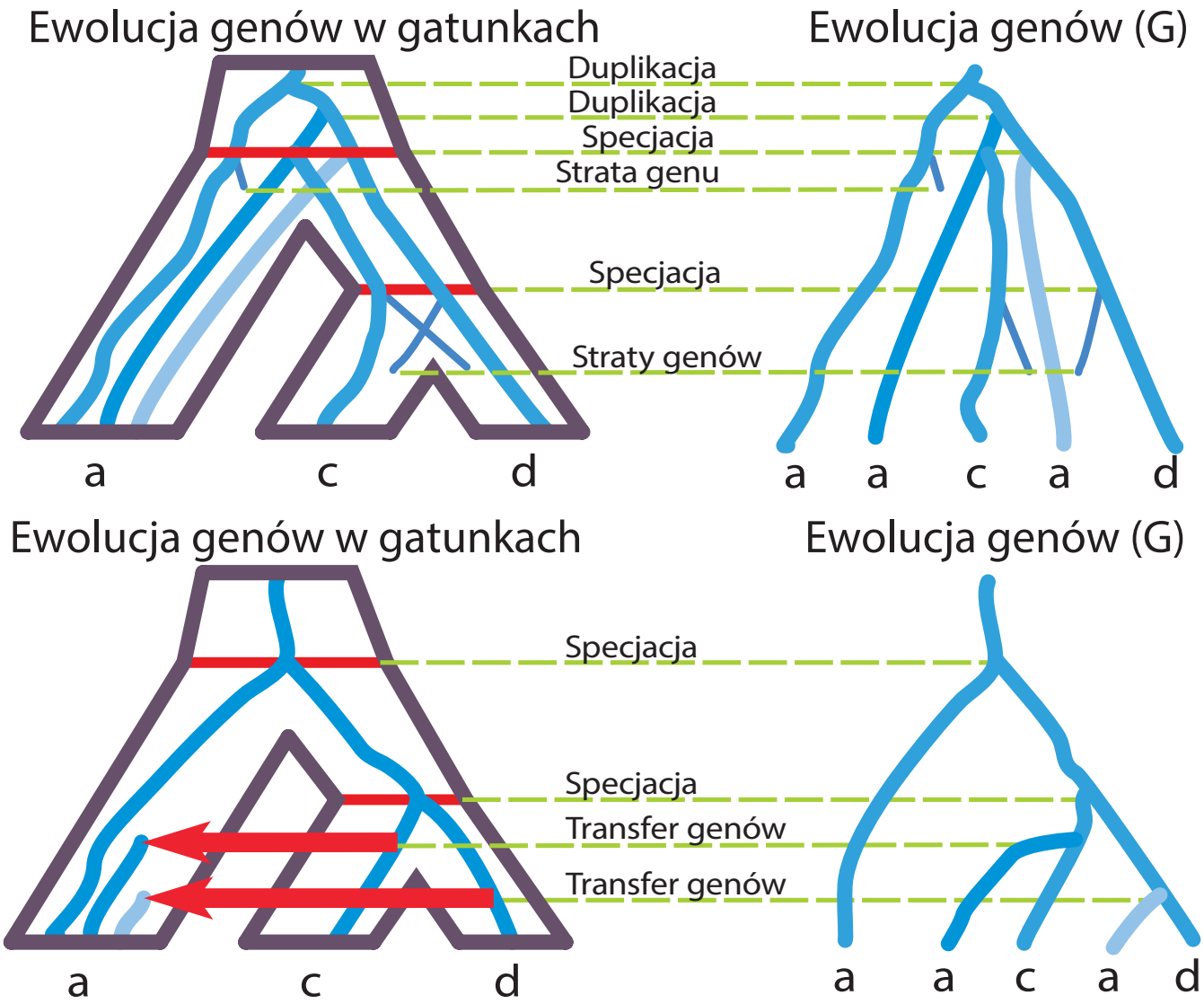

Czas
Ewolucja genów w gatunkach
Ewolucja genów (G)
Duplikacja
Duplikacja
Specjacja
Strata genu
Specjacja
Straty genów
a
c
d
a
a
c
a
d


G
a
a
c
a
d

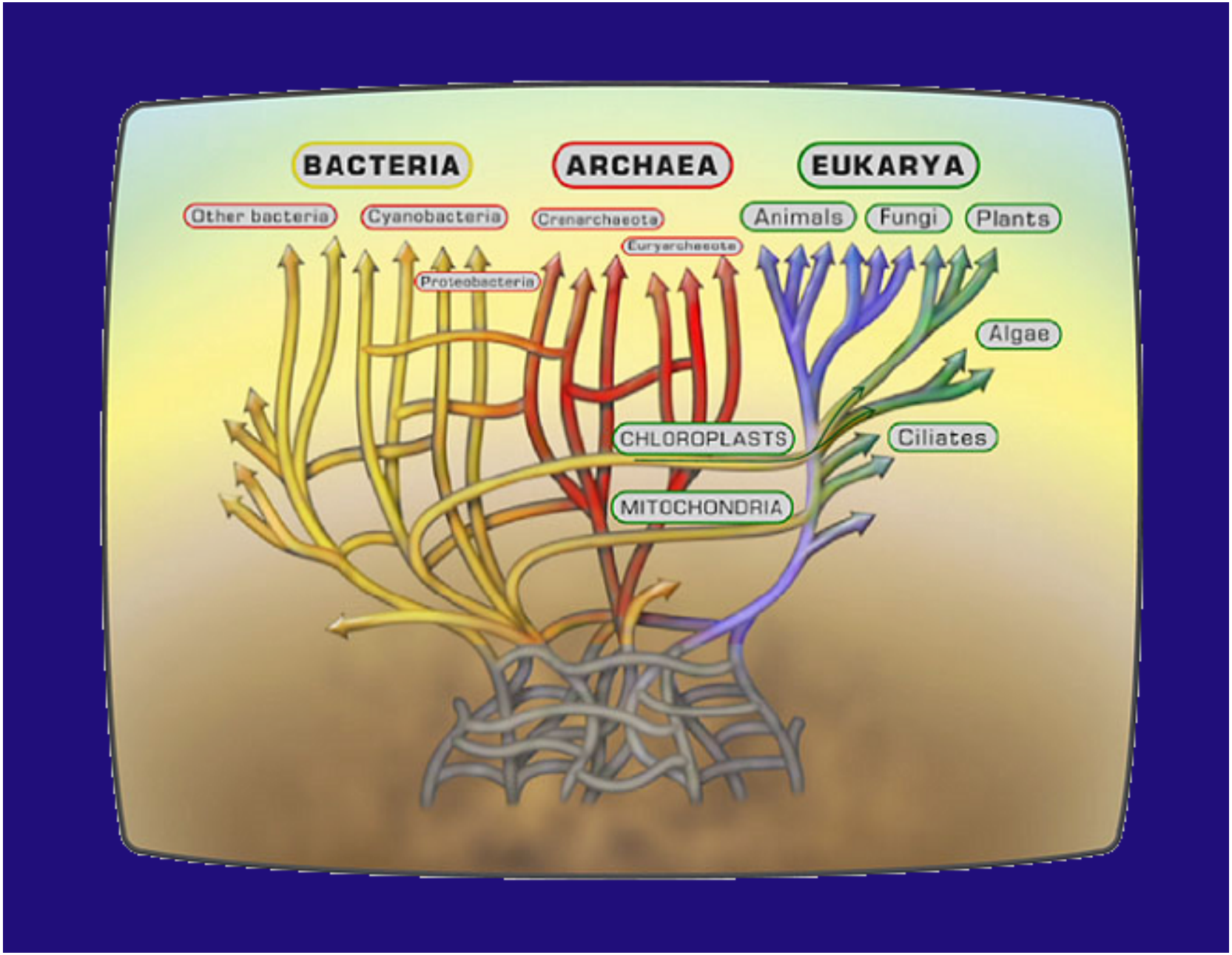

S
a
c
d


a
c
d

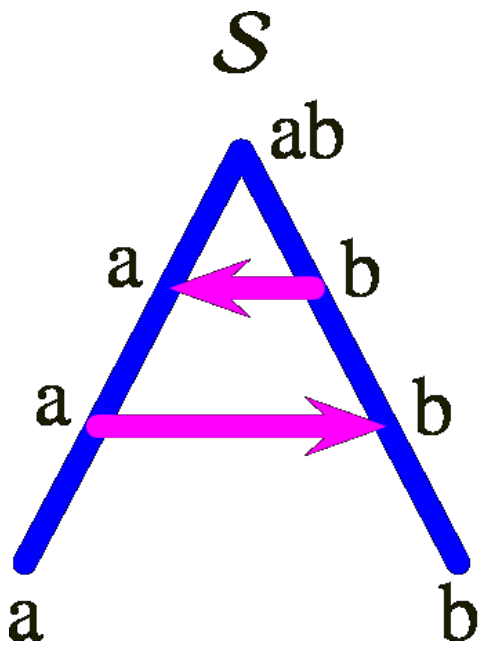

Ewolucja genów w gatunkach
Ewolucja genów (G)
Czas
Specjacja
Specjacja
Transfer genów
Transfer genów
a
c
d
a
a
c
a
d

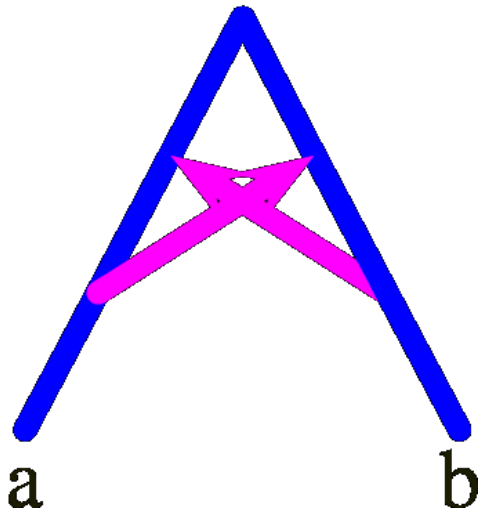


HGT może zmiejszyć koszt (2 transfery vs. 2 duplikacje + 3 straty).

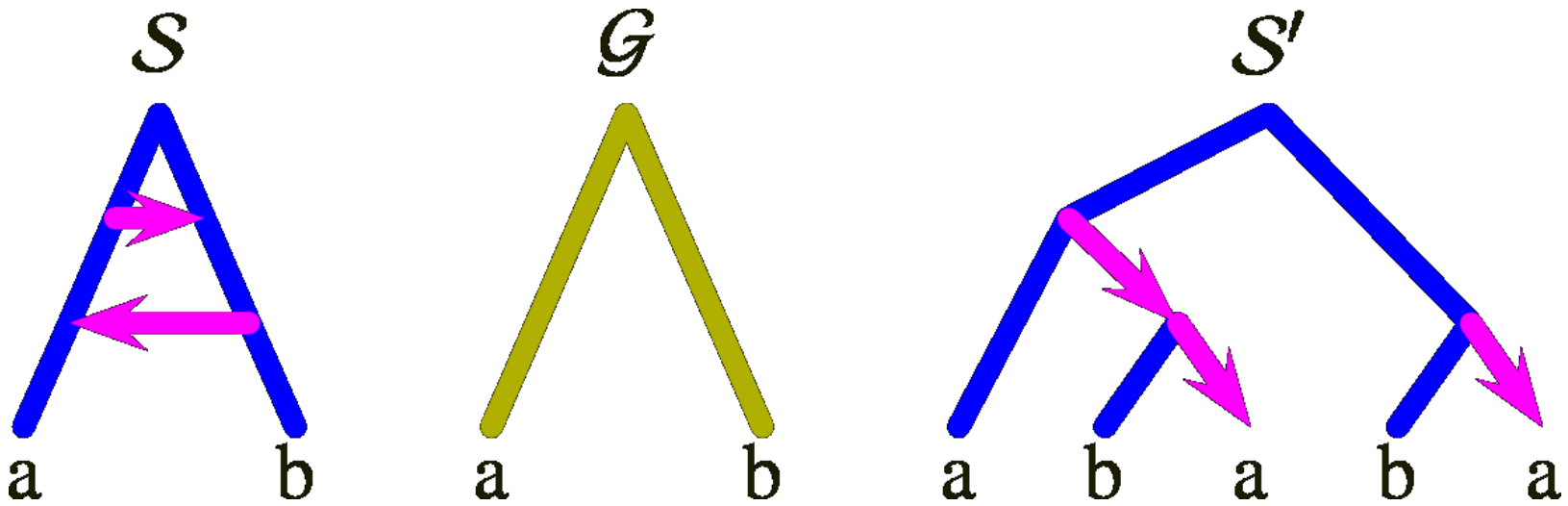

BACTERIA
ARCHAEA
EUKARYA
Other bacteria
Cyanobacteria
Crenarchaeota
Euryarchaeota
Animals
Fungi
Plants
Proteobacteria
Algae
CHLOROPLASTS
Ciliates
MITOCHONDRIA

Modelowanie transferów. Nie każdy układ jest poprawny.



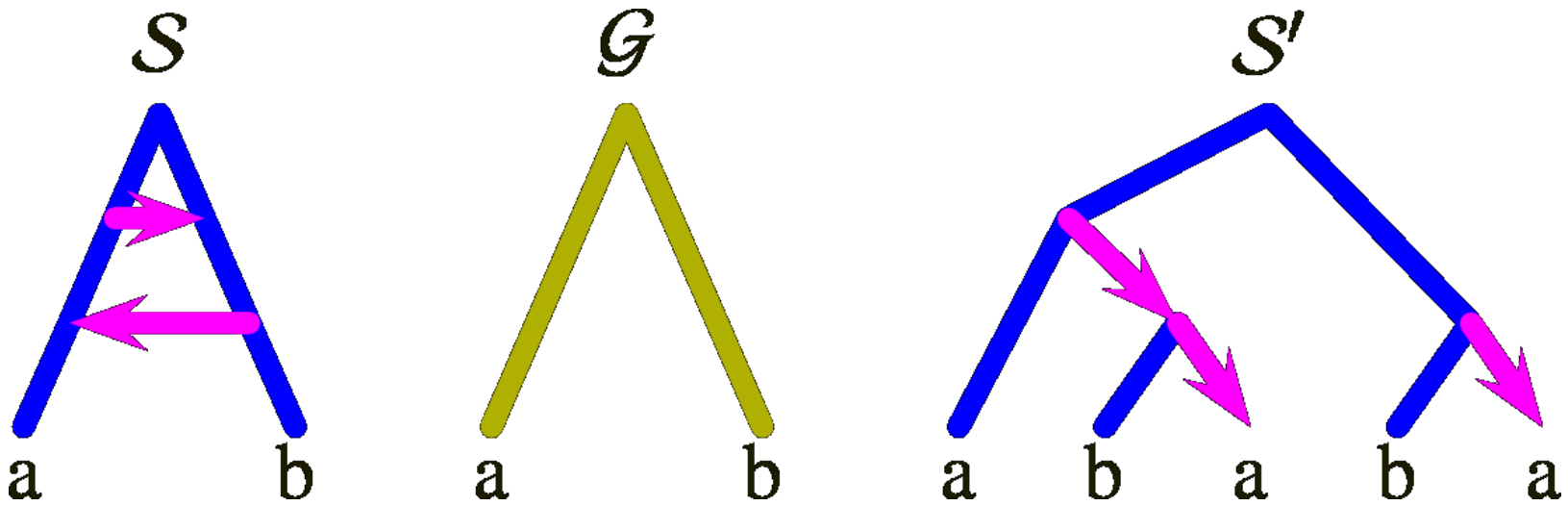


Jak znaleźć scenariusze (wbudowania)?



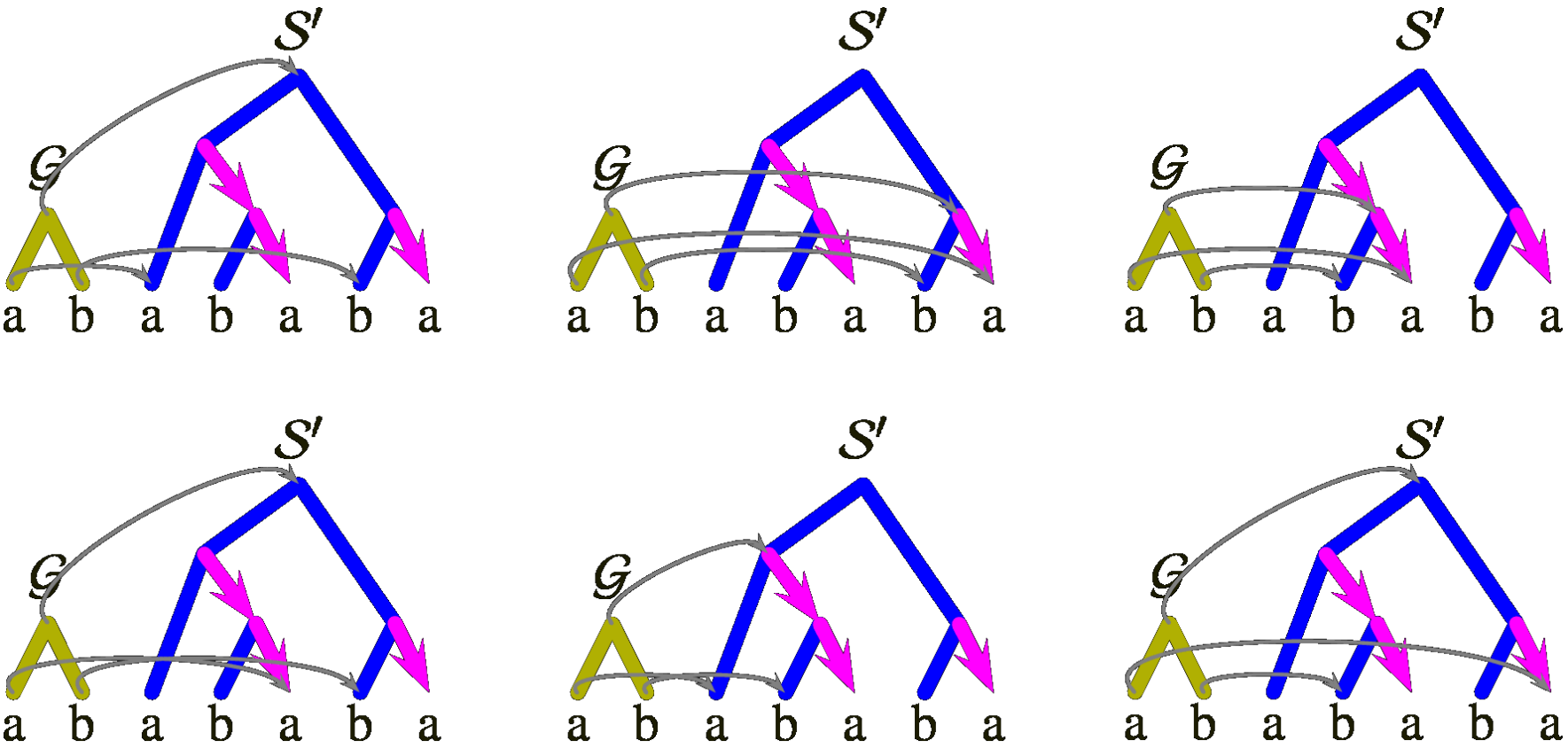
Zaczniemy od S' - rozwinięte drzewo gatunków z HGT.

Jak znaleźć scenariusze (wbudowania)?

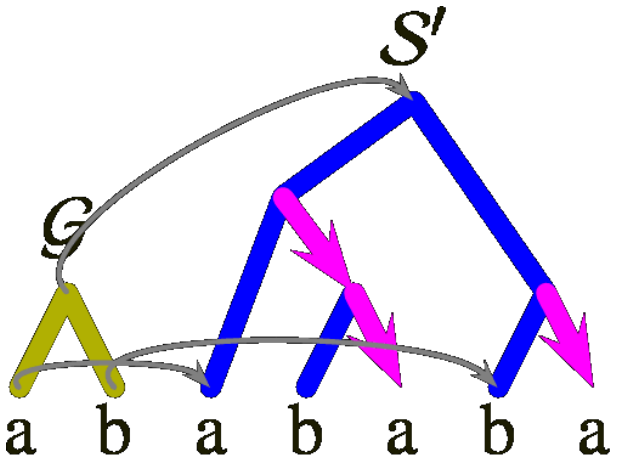


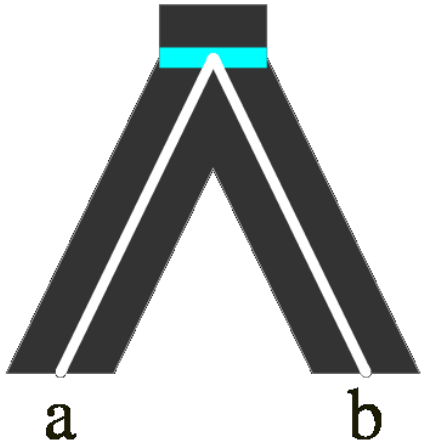
Zaczniemy od S' - rozwinięte drzewo gatunków z HGT.

Generujemy wszystkie możliwe mapowania liści z G do S'. Może być ich bardzo dużo. Tutaj: 3×2.

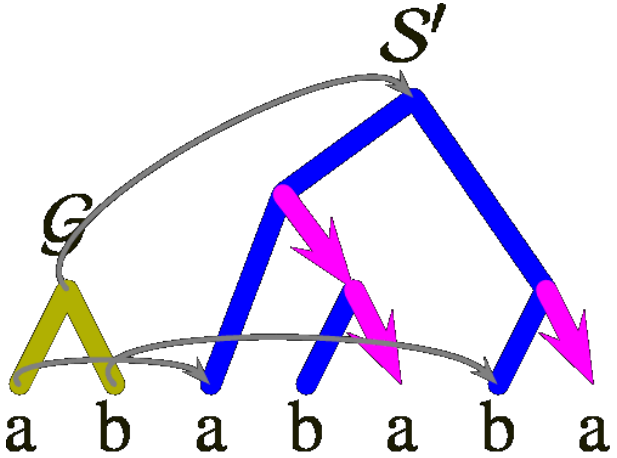


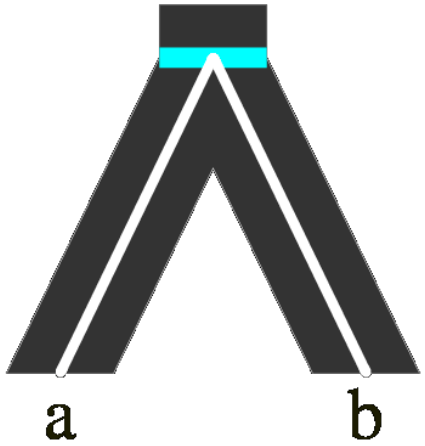
Z każdego tworzymy wbudowania (może być kilka).

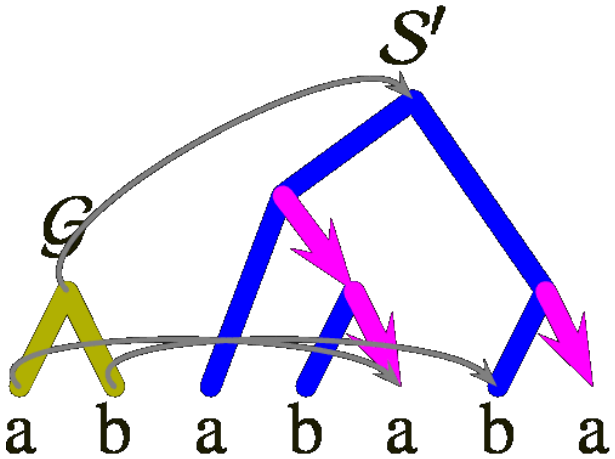


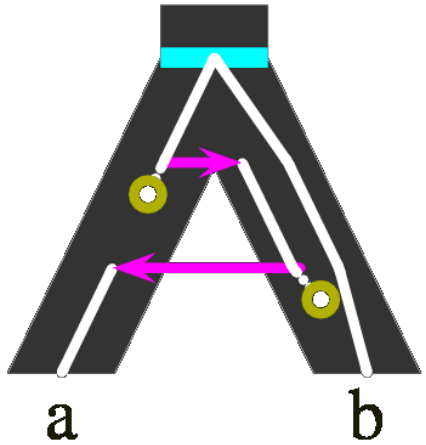


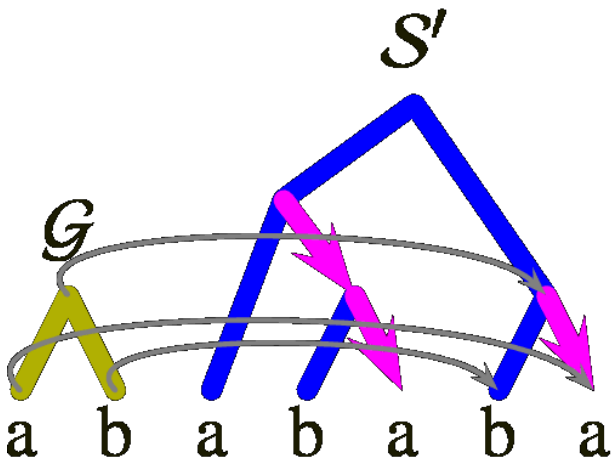
Z każdego tworzymy wbudowania (może być kilka).

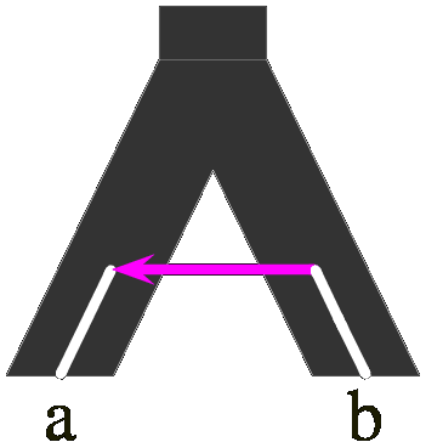


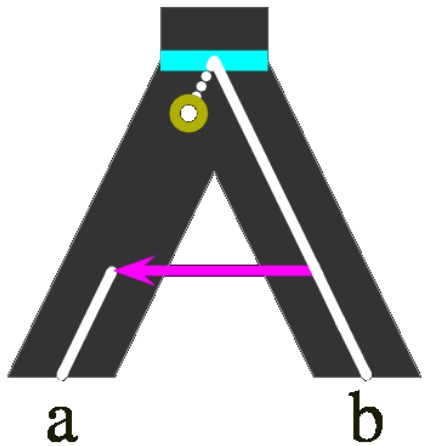


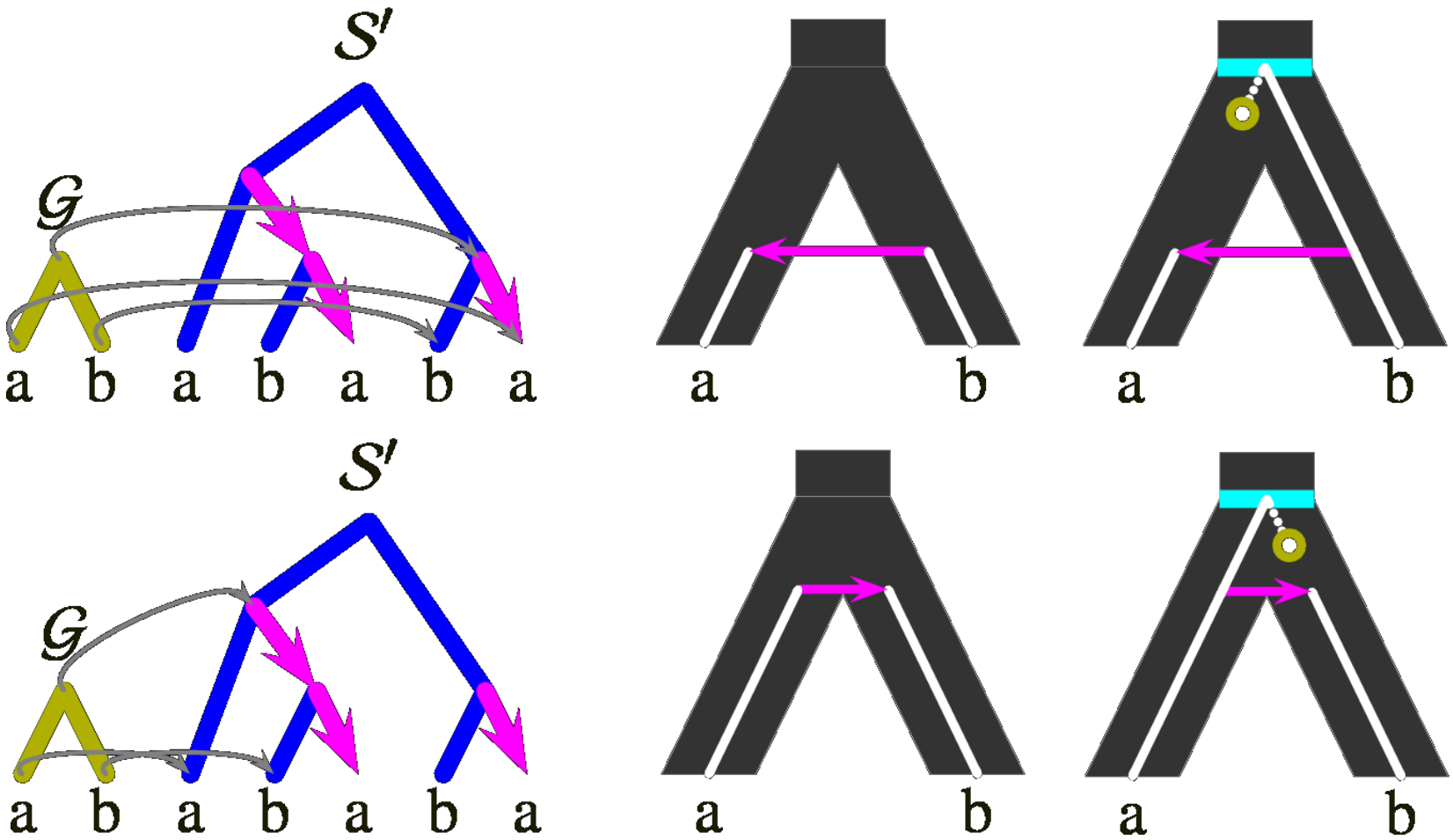


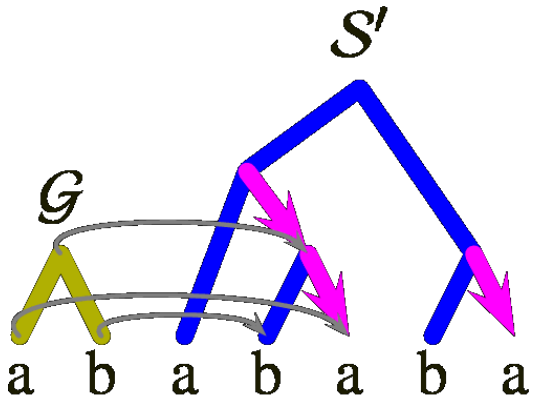


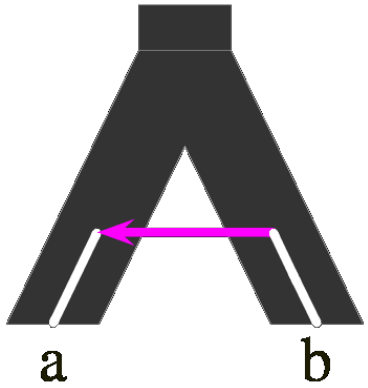

S'
G
a b a b a b a

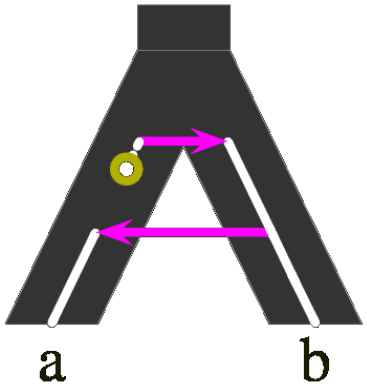

a
b

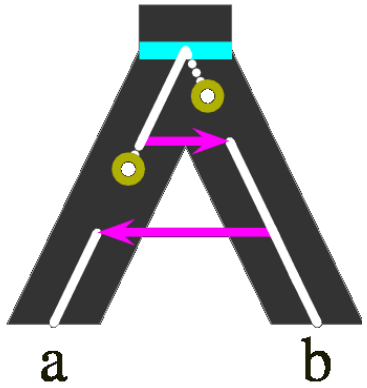

a
b

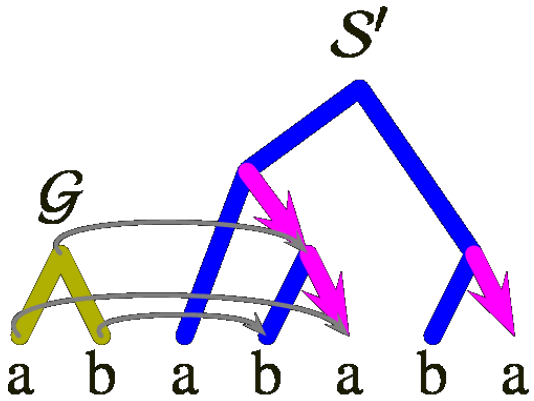

S'
G
a b a b a b a
a b
a b
S'
G
a b a b a b a
a b
a b

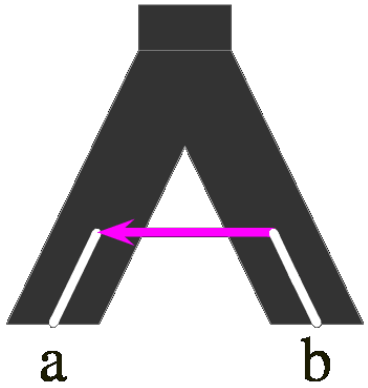

S'
G
a b a b a b a

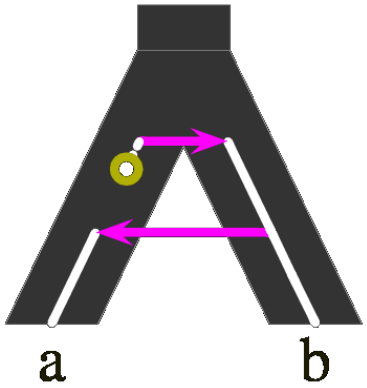

a b

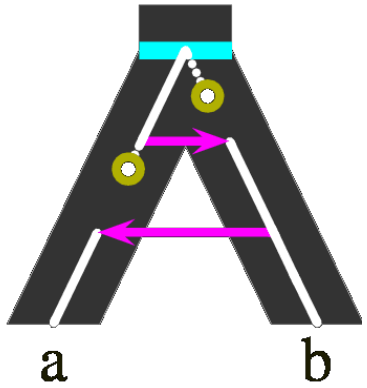

a b

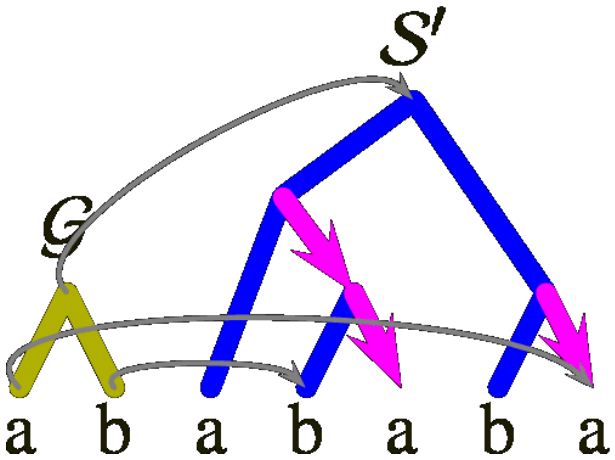

a b

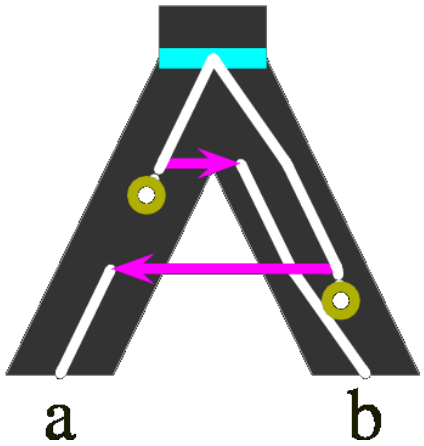

S'
G
a b a b a b a


a b


a b


a b


S'
G
a b a b a b a


a b

W ogólności nie stosuje się takiego rozwiązania. Problem:

Dane G oraz S + HGT. Znajdź scenariusz minimalizujący DLT koszt.

Rozwiązywany jest przez programowanie dynamiczne w czasie wielomianowym (Hallet'2004, Gorecki'2004).

W zastosowaniach nie znamy HGT:

Dane G oraz S (bez HGT). Znajdź scenariusz z HGT minimalizujący DLT koszt.

Jest to problem NP-zupełny (Hallet'2000, Tofigh'2011, Ovadia'2011). Powód to wymagania dot. transferów, które muszą zachować spójność czasu (nie mogą krzyżować się). Jeśli nie ma wymagania spójności, problem można rozwiązać w czasie kwadratowym (Bansal'2012, David'2012).

Lepszym podejściem jest:

Dane G oraz S bez HGT ale z czasami specjacji. Znajdź scenariusz z HGT minimalizujący DLT koszt.

Ten problem wielomianowy (Bansal'2012, Doyon'2010), najlepszy algorytm ma złożoność $O(n^2 \log n)$. Stosuje się metodę plasterkowania.