A Red c on A go fo Co p ng e y dz on N e of o ee

Magnus Bordewich[1]**, Simone Linz**[2,3]**, Katherine St. John**[4]**, and Charles Semple**[2]

[1] Dep rt ent of Co puter Science Durh niversity Durh DH LE nited ingdo
[2] Bio the tics Rese rch Centre Dep rt ent of M the tics nd St tis tics niversity of C nter ury Christchurch Ne Ze nd
[3] Dep rt ent of Bioinfor tics Heinrich Heine niversity Dusse dorf Ger ny
[4] Dep rt ent of M the tics nd Co puter Science Leh n Co ege City niversity of Ne York SA

Correspondence: Si one Linz Bio the tics Rese rch Centre De p rt ent of M the tics nd St tistics niversity of C nter ury Pri v te B g Christchurch Ne Ze nd Te E i inz cs uni duesse dorf de

Running head: A Reduction A gorith for Hy ridiz tion

Key words: hy ridiz tion net or s reticu te evo ution gree ent forest

1 Abstract

Hy ridiz tion is n i port nt evo ution ry process for ny groups of species Thus con icting sign s in d t set y not e the resu t of s p ing or ode ing errors ut due to the f ct th t hy ridiz tion h s p yed signi c nt fief u t

When the initi co ection consists of two rooted in ry phy ogenetic trees Bordewich nd Se p . Consequent y s resu t of this co put tion di cu ty ost current rese rch considers the two tree pro e . There re now sever gorith s for ppro ching this tter pro e . However of these gorith s re either gorith s so ving restricted version of the pro e e.g. H ett nd L gergren Huson **et al.** N h eh **et al.** or po yno i ti e heuristics with no gu r ntee of the c ose ness of their so ution e.g. N h eh **et al.**

In this p per we descri e new nd recent y i p e ented e ct gorith for so ving the two tree pro e with no restrictions sed on three reductions th t preserve the ount of hy ridiz tion. A of these reductions e use of si i rities etween the two trees. It h s recent y een shown th t two of the reductions re enough to gu r ntee th t the gorith is ed p r eter tr ct where the p r eter is the s est nu er of hy ridiz tions to e p in the initi two trees Bordewich nd Se p . This e ns th t the gorith runs e cient y when this s est nu er is ounded. The re ining reduction ows

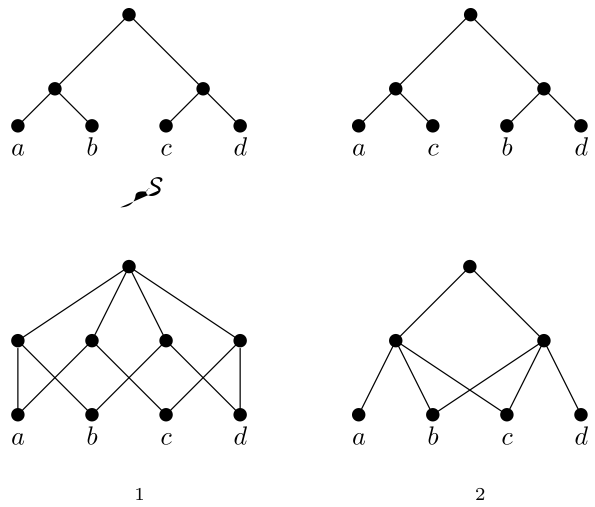


Figure 1. T o rooted in ry phy ogenetic trees nd nd t o hy ridiz tion net or s $_1$ nd $_2$ hich e p in oth trees

co in tion of the ed p r eter resu t descri ed in Borde ich nd Se p e hose proof of correctness is given y Proposition of th t p per nd the c uster reduction descri ed in B roni **et al.** hose proof of correctness is given y Theore in th t p per For si p icity in this p per e on y descri e the in ide s For the re der interested in the ner det i s e refer the to the origin p pers

3 Reduction Algorithm for Hybridization

e egin ith for description of the t o tree pro e A **rooted binary phylogenetic X-tree** is rooted tree th t h s e f set X nd hose root h s degree t o hi e other interior vertices h ve degree three A **cluster** of is su set of X th t cont ins precise y the e e ents th t re descend nts of so e verte of

A **rooted acyclic digraph** is digr ph ith no directed cyc es E ch such digr ph h s distinguished verte ρ hose in degree is zero nd h s the property th t there is directed p th fro ρ to every other verte For verte v in digr ph e denote the **in-degree** of v the nu er of edges directed into v y d^- v nd the **out-degree** of v the nu er of edges directed out of v y d^+ v A **hybridization network** on X is rooted cyc ic digr ph ith root ρ in hich

i X is the set of vertices of out degree zero

ii d^+ ρ nd

such net or the s er the size of the resu ting gree ent forest for S nd where the size of forest is the nu er of trees in the forest On the other h nd if e re given n gree ent forest for S nd then one c n reverse this process to construct hy ridiz tion net or th t e p ins S nd

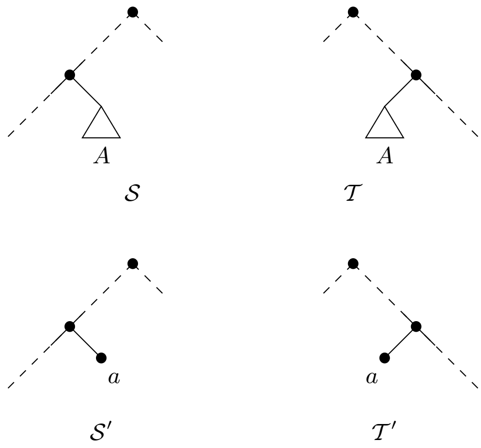


Figure 2. T o rooted in ry phy ogenetic trees $\mathcal{S}$ nd reduced under the su tree reduction ru e The tri ng e

Further ore the correctness of the ch in reduction ru e fo o s fro Proposition of Borde ich nd Se p e

ii Borde ich nd Se p e sho ed th t the su tree nd ch in reductions y the se ves re enough to erne ize the pro e nd give ed p r eter gorith for **Hybridization Number**. The c uster reduction provides n e tre e y usefu too for re ing the pro e into nu er of s er pro e s th t is required is th t the su trees h ve identic e f sets the topo ogies of the t o su trees c n e co p ete y di erent.

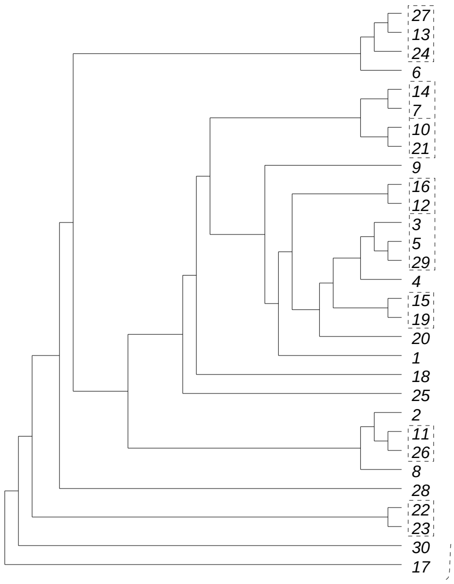
iii ithout going into det i s the c uster reduction h s si i r 7 vor to the Deco position Theore in Huson **et al.** This the ore descri es one to one correspondence et een the over pping cyc es of n unrooted net or $\mathcal{N}$ the connected co ponents of the inco p ti i ity gr ph of the sp its gener ted y $\mathcal{N}$ nd the netted co ponents of the sp its gr ph of the sp its gener ted y $\mathcal{N}$. Ho ever hi e this theore yie ds n gorith for ini izing the nu er of hy ridiz tion vertices ongst restricted c ss of net or s it is i port nt to note th t it does not give gener str tegy for ini izing this nu er ongst hy ridiz tion net or s s there is no gu r ntee th t such reduction e ds to n opti so ution. In con tr st B roni **et al.** sho ed th t such str tegy in p rticu r the c uster reduction or s for t o trees. It is n interesting open

Table 2. Resu ts for the **Poaceae** d t set.

pairwise combination		# taxa	hybridization number	run time[a]
nd	*yB*	**40**	**14**	**11 h**
nd	*cL*	**36**	**13**	**11.8 h**
nd	*oC*	**34**	**12**	**26.3 h**
nd	*ry*	**19**	**9**	**320 s**
nd	[illegible]	**46**	**at least 15**	**2 d**
yB	*cL*	**21**	**4**	**1 s**
yB	*oC*	**21**	**7**	**180 s**
yB	*ry*	**14**	**3**	**1 s**
yB	[illegible]	**30**	**8**	**19 s**
cL	*oC*	**26**	**13**	**29.5 h**
cL	*ry*	**12**	**7**	**230 s**
cL	[illegible]	**29**	**at least 9**	**2 d**
oC	*ry*	**10**	**1**	**1 s**
oC	[illegible]	**31**	**at least 10**	**2 d**
ry	[illegible]	**15**	**8**	**620 s**

[a]**run time on a 2000 MHz CPU, 2 GB RAM machine measured in seconds (s), hours (h), and days (d), respectively**

p st sequence phytochro e B **phyB** nd the nuc e r sequence of the inter n tr nscri ed sp cer of ri oso DNA **ITS** [illegible]ich h ve n over pping t set of present d y species see the ro[illegible]indic ted y the gr y c


27
13
24
6
14
7
10
21
9
16
12
3
5
29
4
15
19
20
1
18
25
2
11
26
8
28
22
23
30
17

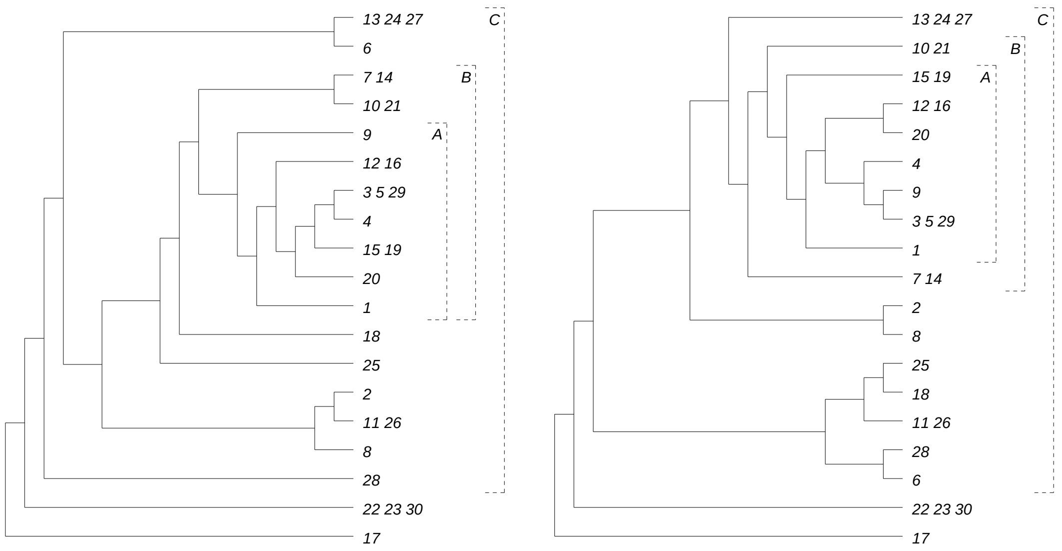
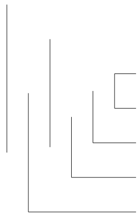

13 24 27
6
7 14
10 21
9
12 16
3 5 29
4
15 19
20
1
18
25
2
11 26
8
28
22 23 30
17
A
B
C
13 24 27
10 21
15 19
12 16
20
4
9
3 5 29
1
7 14
2
8
25
18
11 26
28
6
22 23 30
17
A
B
C

Figure 6.



su trees ch ins or c usters nich is i e y for ny io ogic e p es the ne gorith perfor s re r y e nd the hy ridiz tion nu er c n e found in re son e ti e

Note th t HybridNumber c cu tes er ound for the nu er of hy ridiz tion events to e p in the di erences et een t o phy ogenetic gene trees ssu ing th t hy ridiz tion is the on y c use of incongruence et een the t o trees It is possi e th t the re nu er of hy ridiz tion events th t h ppened during the evo ution of the co ection of present d y species under consider tion is underesti ted Indeed it is possi e th t so e hy ridiz tion events re never recognized Neverthe ess the gorith provides ni e rport nt esnts

N h eh L Ruths D nd ng LS RIATA HGT f st nd ccu r te heuristic for reconstructing horizont gene tr nsfer In **Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)** Lecture Notes in Co puter Science o Springer p

N h eh L rno T Linder CR et Reconstructing reticu te evo ution in species theory nd pr ctice **Journal of Computational Biology 12**

O sen G M tsud H H gstro R et f stDNA L too for con struction of phy ogenetic trees of DNA sequences using i u i e i hood **Comput Appl Biosci 10**

Riese erg LH R y ond O Rosenth DM et M jor eco ogi c tr nsitions in d sun ers f ci it ted y hy ridiz tion **Science 301**

Sch idt HA Phy ogenetic trees fro rge d t sets PhD thesis Heinrich Heine niversit t Dusse dorf

Se p e C nd Stee M **Phylogenetics** O ford niversity Press

Appendix

A Pseudocode

Here e present the pseudocode of HybridNumber For rooted in ry phy ogenetic X tree nd su set A of X e denote the ini su tree of connecting the e e ents in A y A Further e denote the tree for ed y rep cing c uster A ith the ne e f c y A c If B is su set of X e use B to denote the phy ogenetic tree o t ined fro y de eting e ch of the e e ents in B nd suppressing ny resu ting degree t o verte Fin y E denotes the forest o t ined fro the tree y de eting the edges in the set E Bec use of the ch in reduction ru e the input to HybridNumber inc udes eight function w on p irs of t this c n e t en to e zero for p irs in the initi input

Algorithm A.1: HybridNumber $(\mathcal{S}, w)$

$(\mathcal{S}, w) \leftarrow$ SubtreeReduction $(\mathcal{S}, w)$
$(\mathcal{S}, w) \leftarrow$ ChainReduction $(\mathcal{S}, w)$
if ini co on c uster C of $\mathcal{S}$ nd **and**
$< C <$ nu er of t of $\mathcal{S}$
do $(\mathcal{S}_1, {}_1, w_1, \mathcal{S}_2, {}_2, w_2) \leftarrow$ ClusterReduction $(\mathcal{S}, w)$
$h_1 \leftarrow$ Exhaustustustustustustu $\mathcal{S}$, w -4 (t) 9 (o) -4 f 0

Algorithm A.4: ClusterReduction $\mathcal{S}, w$

```
C ←  ini   co   on c uster of S nd
S_1 ← S C
S_2 ← S C   c
 1 ←  C
 2 ←  C   c
w_1 ← w restricted to p irs of t   in C
w_2 ← w restricted to p irs of t   not in C
return  S_1,  1, w_1, S_2,  2, w_2
```

Algorithm A.5: ExhaustiveSearch $\mathcal{S}, w$

```
if S      return
h ←  nu  er of  e ves of S
i ←
repeat
 for each E   su set of the edges of S such th t E   i
      ←    S, E
      if   is  n  cyc ic  gree  ent forest of S nd
  do        P ←    a, b   a, b  re iso  ted t    in
            h' ←  i   Σ_(a,b)∈P w a, b
       do   if h' < h
              do h ←  h'
 i ←  i
until i   h
return  h
```

Remarks

- The ctu   i p e ented   gorith s cont in v rious s   i prove ents co p red to the pseudocode in order to i prove running ti e hi st these ch nges do not   ect the theoretic   orst c se running ti e in pr ctice they re ene ci An e   p e is th t no gree ent forest h s n iso ted intern   verte hence in the e h ustive se rch e do not need to consider su sets of edges of size i to de ete fro S hich cont in the three edges incident ith   p rticu r verte

- In HybridNumber fo o ing   c   to the c uster reduction the c us ter re oved c nnot e reduced   ny further using the reductions in hich c se e i   edi te y c   ExhaustiveSearch Ho ever it   y no e possi e to further reduce the re   inder of the trees   nd so e c   HybridNumber