SUPPLEMENTAL MATERIALS FOR "AN ALGORITHM FOR COMPUTING THE GENE TREE PROBABILITY UNDER THE MULTISPECIES COALESCENT AND ITS APPLICATION IN THE INFERENCE OF POPULATION TREE" BY YUFENG WU

1 NOTATIONS OF THE COMPACTCH ALGORITHM

Table 5 lists the main notations of the CompactCH algorithm.

2 COMPARISON OF COMPACTCH AND STELLS

2.1 CompactCH runs in exponential time in general.

We first show that in the general case, the number of compact coalescent histories $\mathcal{N}(T_g, T_s)$ is exponential in terms of n and m for gene tree topology T_g and species tree T_s. Recall that n is the total number of gene lineages and m is the number of populations.

We consider the special case where T_g and T_s have the same topology, which is the rooted caterpillar shape. That is, the topology is in the Newick form $(\ldots(((1,2),3),4),\ldots,n)$. Note that this means there is a single gene lineage per taxon. In this case, $\mathcal{N}(T_g, T_s)$ is equal to the number of coalescent histories $\mathcal{H}(T_g, T_s)$. To see this, we only need to show that a compact coalescent history has only a single coalescent history associated with it. The important property of the caterpillar tree is that each coalescent occurs sequentially: if a coalescent c occurs, then its (single) direct descendant coalescent (if any) must have occurred already. Note that a compact coalescent history specifies the number of coalescents in each species tree branch. Suppose we examine each non-leaf species tree branch from low to high. Then, the set of coalescents along each non-leaf species tree branch can be fully determined. In other words, there is a single coalescent history corresponding to each compact coalescent history in this case. Since it is known in [2] that the number of coalescent histories is the Catalan number in n (which grows exponentially), $\mathcal{N}(T_g, T_s)$ is exponential in n. So we have:

PROPOSITION 2.1. *The gene tree probability computation in Equation 5 runs in exponential time for general T_g and T_s.*

2.2 The STELLS algorithm runs in exponential time for two populations.

We now show that the STELLS algorithm runs in exponential time in n even when there are only two species. That is, CompactCH runs in polynomial time while the original STELLS has exponential running time when the number of populations is two. Thus, when the size of T_s is small, our new algorithm outperforms the STELLS algorithm due to Theorem 3.2. We consider the case where T_g is maximally symmetric (i.e. fully balanced). For simplicity, we assume $n = 2^k$ for some k. The leaves in the left half of T_g are from the species A and the other leaves are from the species B. By a similar reasoning as in [19], the numbers of ancestral configurations ACs formed by lineages from one species just after the split of A and B are exponential in n (see the Appendix of [19]). Thus, the total number of ACs is exponential in n when T_g is fully balanced and gene lineages from two species are perfectly separated even when there are only two species.

%noindent

Table 5. Main notations of the CompactCH algorithm

Notation	Explanation
T_g	gene tree topology (with no branch lengths)
T_s	species (population) tree (with branch length)
$\boldsymbol{h}$	coalescent history
$\mathcal{H}(T_g\|T_s)$	the set of all coalescent histories for T_g and T_s
m	the number of species tree branches
b	a branch in T_s
λ_b	branch length of branch b
$u_b(\boldsymbol{h})$	number of gene lineages at the top (away from extant time) of branch b by coalescent history $\boldsymbol{h}$; also called upper lineage count
$v_b(\boldsymbol{h})$	number of gene lineages at the bottom (close to extant time) of branch b by coalescent history $\boldsymbol{h}$; also called lower lineage count
$p_{uv}(T)$	the probability of coalescing u lineages backward in time to v lineages within time T (in coalescent units)
$c_b(\boldsymbol{h})$	number of coalescents within branch b as specified by $\boldsymbol{h}$
$C_b(\boldsymbol{h})$	set of coalescents within b as specified by $\boldsymbol{h}$
$d_b(\boldsymbol{h})$	number of all possible ways of coalescing $c_b(\boldsymbol{h})$ events within b
$w_b(\boldsymbol{h})$	number of ways of coalescing $c_b(\boldsymbol{h})$ events within b s.t. the resulting gene tree topology matches T_g
$\boldsymbol{ch}$	compact coalescent history
$\mathcal{CH}(T_g\|T_s)$	The list of all compact coalescent histories for the trees T_s and T_g
$u_b(\boldsymbol{ch})$	upper lineage count specified by $\boldsymbol{ch}$
$v_b(\boldsymbol{ch})$	lower lineage count specified by $\boldsymbol{ch}$
$c_b(\boldsymbol{ch})$	number of coalescents within branch b as specified by $\boldsymbol{ch}$
$\mathcal{H}(\boldsymbol{ch})$	the set of coalescent histories combined in $\boldsymbol{ch}$
$Desc(b)$	the set of descendant branches of T_s
$C(\boldsymbol{ch})$	the coalescent coefficient of $\boldsymbol{ch}$
$C(\boldsymbol{ch})$	a term needed to compute $C(\boldsymbol{ch})$
$\boldsymbol{n}$	the numbers of coalescents within the subtree rooted at a coalescent c
$C_2(c, b_s, \boldsymbol{n})$	restricts C_1 to the subtree rooted at coalescent c where c occurs on b_s and the numbers of coalescents within the subtree rooted at c is $\boldsymbol{n}$
g_c	number of alleles in a population