Notes on PADRE algorithm

Graham Jones

2011-10-25

1 PADRE algorithm

From [1]:

> Input: A MUL-tree T on the taxa set X with n nodes and height h_{max}.
> Output: The network $N(T)$.
> Computation of $N(T)$
> 1. Assign to every node v in T a code $c(v)$ between 1 and n.
> 2. Initialise a list H of h_{max} lists, in which each list will contain
> nodes v ordered by their code $c(v)$, so that the last list contains $\rho(T)$
> and all remaining lists are empty.
> 3. for $h = h_{max}$ to 0 do
> 4. Choose the hth list $l_h \in H$ (so that, for
> all v in l_h the height of the subtree of T with root v is h).
> 5. while $|l_h| > 0$ do
> 6. Let T_1 be the subtree of T whose root $\rho(T_1)$ is the first node in l_h.
> 7. For each v a child of $\rho(T_1)$ having height i, add v into the ith list
> l_i of H.
> 8. Iterate through l_h to find the subtrees $T_2, .., , T_m$ of T
> equivalent to T_1 using node codes, stopping when an element with a
> non-equal code is found.
> 9. if $m \geq 2$ then
> subdivide the incoming edges to the subtrees $T_1, T_2, ..., T_m$,
> identify all newly created subdivision nodes,
> prune $T_2, ..., T_m$ and their incoming edges from the created network.
> 10. Remove $\rho(Ti)$, $1 \leq i \leq m$, from l_h.
> 11. end while
> 12. end for

Corrected l_i to l_h in step 8. $\rho(T_i)$ is the root of subtree T_i. $c(v)$ is explained like this:

From [1], on step 1:

> In particular, codes are assigned to nodes so that roots of subtrees of T are assigned the same code if and only if the subtrees are equivalent. Essentially, this is done by arbitrarily ordering the set $X = \{x_1, ..., x_{|X|}\}$, assigning code i to each leaf labelled by x_i, and then using a trie data structure (Knuth 1997) to recursively assign codes to the internal nodes of T in a bottom-up fashion based on the codes of their children.

If you only want a count of the minimum hybridizations, I think the algorithm can be simplified. There is no need to edit the network - step 9 can be skipped. Step 1 is key, and probably the slowest part. Once that is done, the nodes can be visited in order of decreasing height, constructing the lists for smallar heights (step 7) as you go. The count at each height is the number of nodes at that height minus the number of equivalence classes at that height. This is found at step 8.

So, skipping step 9 and summarising the other 'inner loop' steps 6,7,8,10:

Input: A MUL-tree T on the taxa set X with n nodes and height h_{max}.
Output: The minimum number of hybridizations ζ.
Computation of ζ
1. Assign to every node v in T a code $c(v)$ between 1 and n.
2. Initialise a list H of h_{max} lists, so that the last list contains $\rho(T)$ and all remaining lists are empty. (Each list will contain nodes v of the same height ordered by their code $c(v)$. The hth list in H, denoted l_h, contains nodes of height h).
3. $\zeta \leftarrow 0$
4. for $h = h_{max}$ to 0 do
    5. Find the equivalence classes $q_1, ...q_m$ of subtrees in l_h
    6. $\zeta \leftarrow \zeta + |l_h| - m$
    7. For one representative T_j^* of each q_j, add each child of $\rho(T_j^*)$ to the ith list l_i of H, where i is the child's height.
8. end for

In the case of diploids and allotetraploids only, leaf labels only occur once or twice. So equivalence classes $q_1, ...q_m$ in step 5 will only have sizes one or two. Furthermore, once two equivalent subtrees have been found, each tree must contain only singly-labelled leaves. So in step 7, it is only necessary to add children of subtrees that are not equivalent to any other, ie, those in q_j with $|q_j| = 1$.

To do step 1, could use a list like H where $l_h \in H$ contains all the nodes of height h. Assign codes to leaf nodes as described above.

1. $r \leftarrow |X| + 1$ (r is the next code to use)
2. for $h = 1$ to h_{max} do
    3. for $v \in l_h$ do
        4. if (v is uncoded)
            5. $c(v) \leftarrow r$
            6. for $w \in l_h$ do
                7. if (codes of w's and v's children match) $c(w) \leftarrow r$
            8. end for
            9. $r \leftarrow r + 1$
        10. end if
    11. end for
12. end for

References

[1] Katharina T Huber and Bengt Oxelman and Martin Lott and Vincent Moulton, *Reconstructing the Evolutionary History of Polyploids from Multilabeled Tree*, Mol. Biol. Evol. 23(9):17841791. 2006 doi:10.1093/molbev/msl045