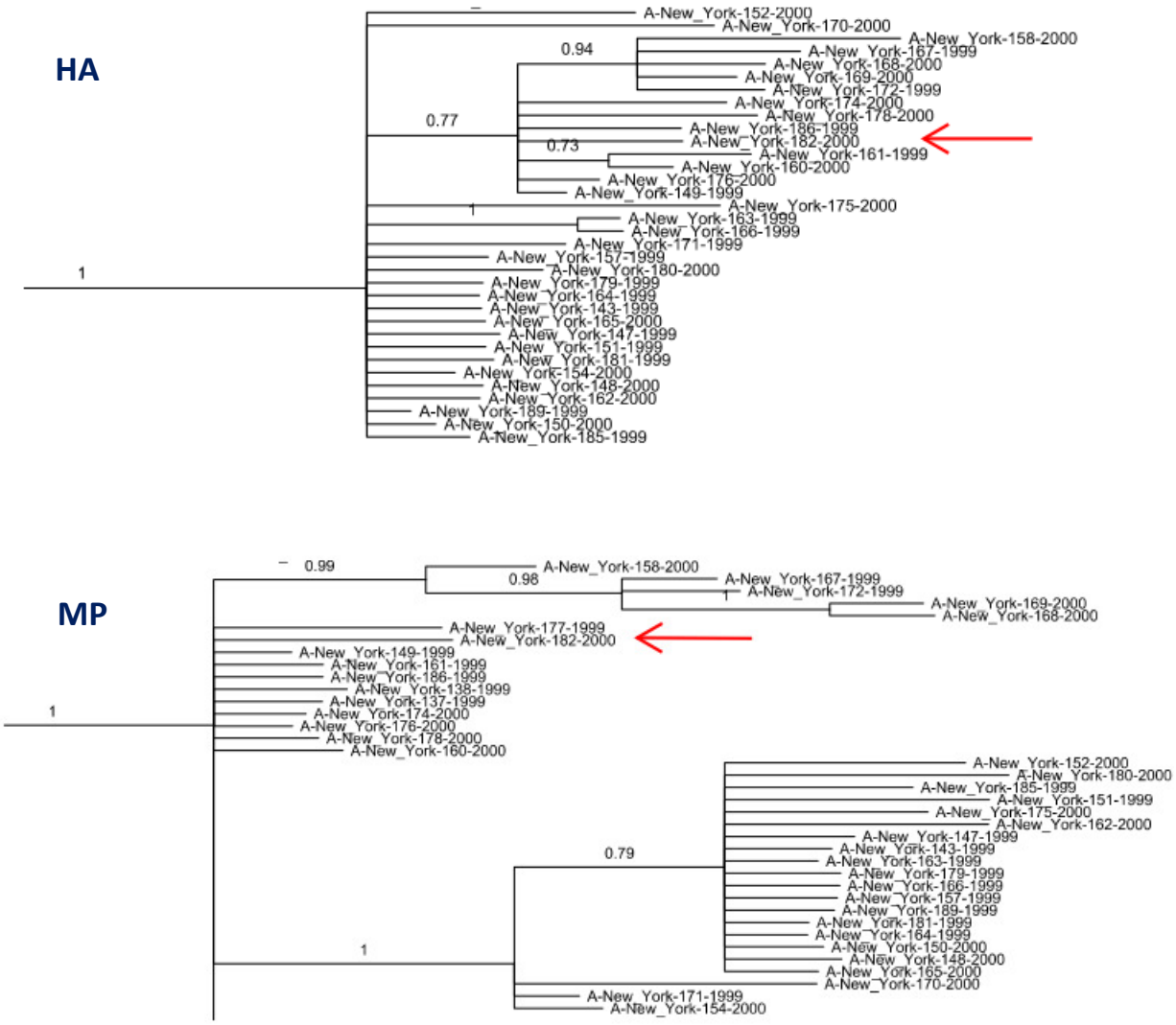
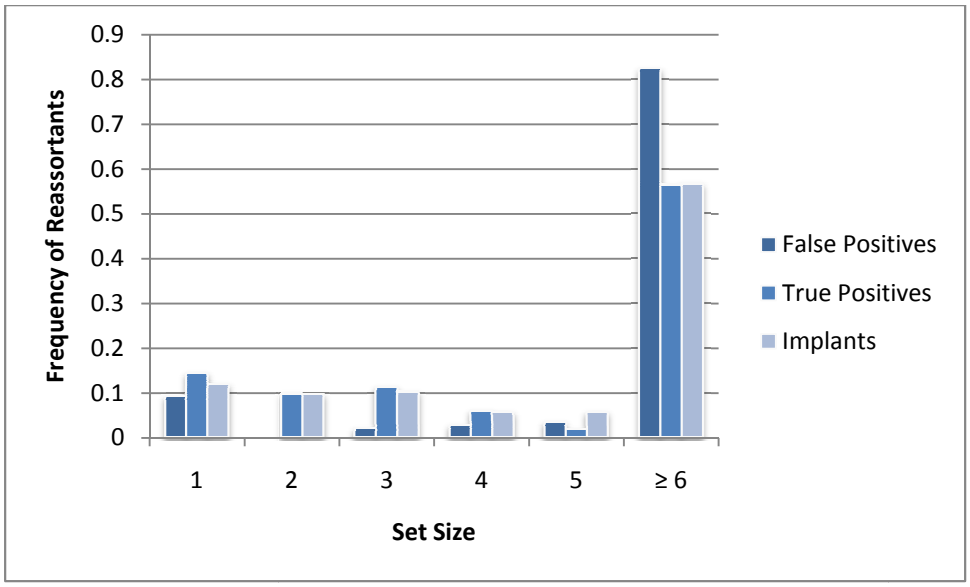
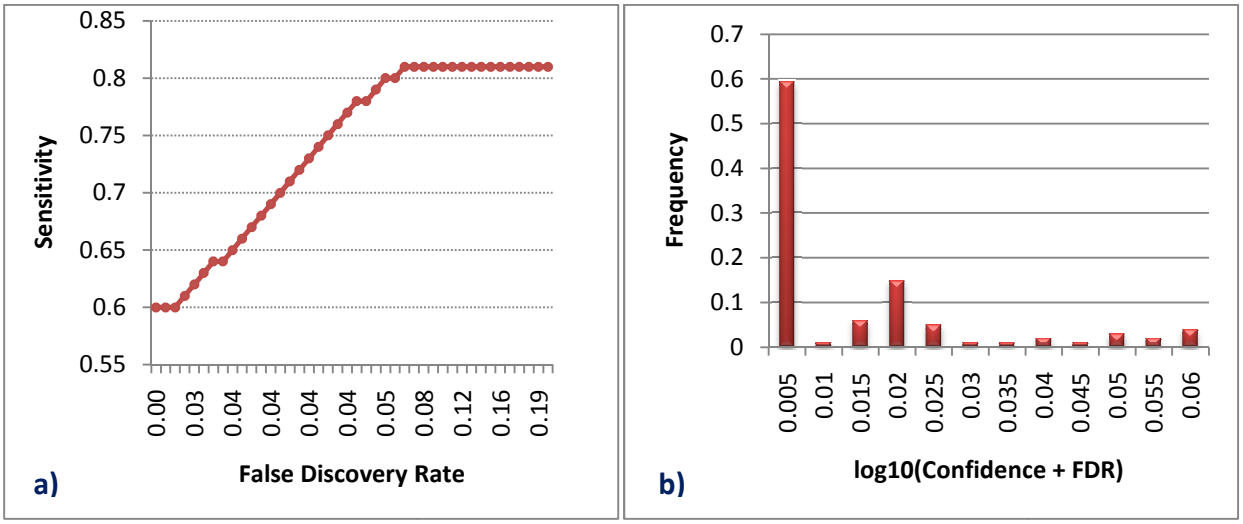
Supplementary Material



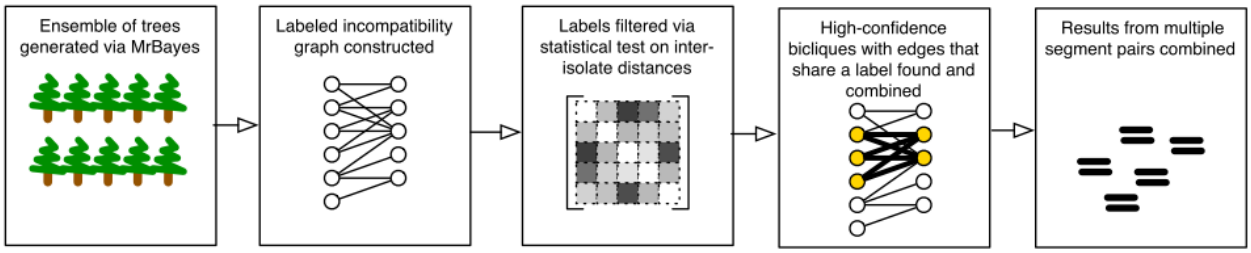
Supplementary Figure 1: HA and MP segment phylogenies (partial) for isolates from Holmes et al. The set of strains in the phylogenetic neighbourhood of A/New York/182/2000 seem to be consistent between the two segment phylogenies suggesting that it is unlikely to be a reassortant (consensus trees from GiRaF, drawn using the program FigTree, http://tree.bio.ed.ac.uk/software/figtree/).



Supplementary Figure 2: Size distribution of implanted and candidate reassortments. The graph shows the frequency of reassortants as a function of the size of true positive, false positive and implanted reassortments for the "All Events" experiment described in **Table 1**. Interestingly, while the size distribution of true positives matches that of the implants, false positives are more often large sets (≥ 6 taxa) and hence amenable to manual filtering. The trend of false positives being large sets is also more pronounced with multiple reassortments (data not shown).



Supplementary Figure 3: Confidence values for candidate reassortments. a) Tradeoffs using various confidence value thresholds for the "All Events" experiment described in **Table 1**, b) Calibration of confidence values for the same experiment – in the ideal case the histogram shown would have a single peak (for the range 0-0.005) indicating that *false-discovery-rate = 1 – confidence-value* in all cases.



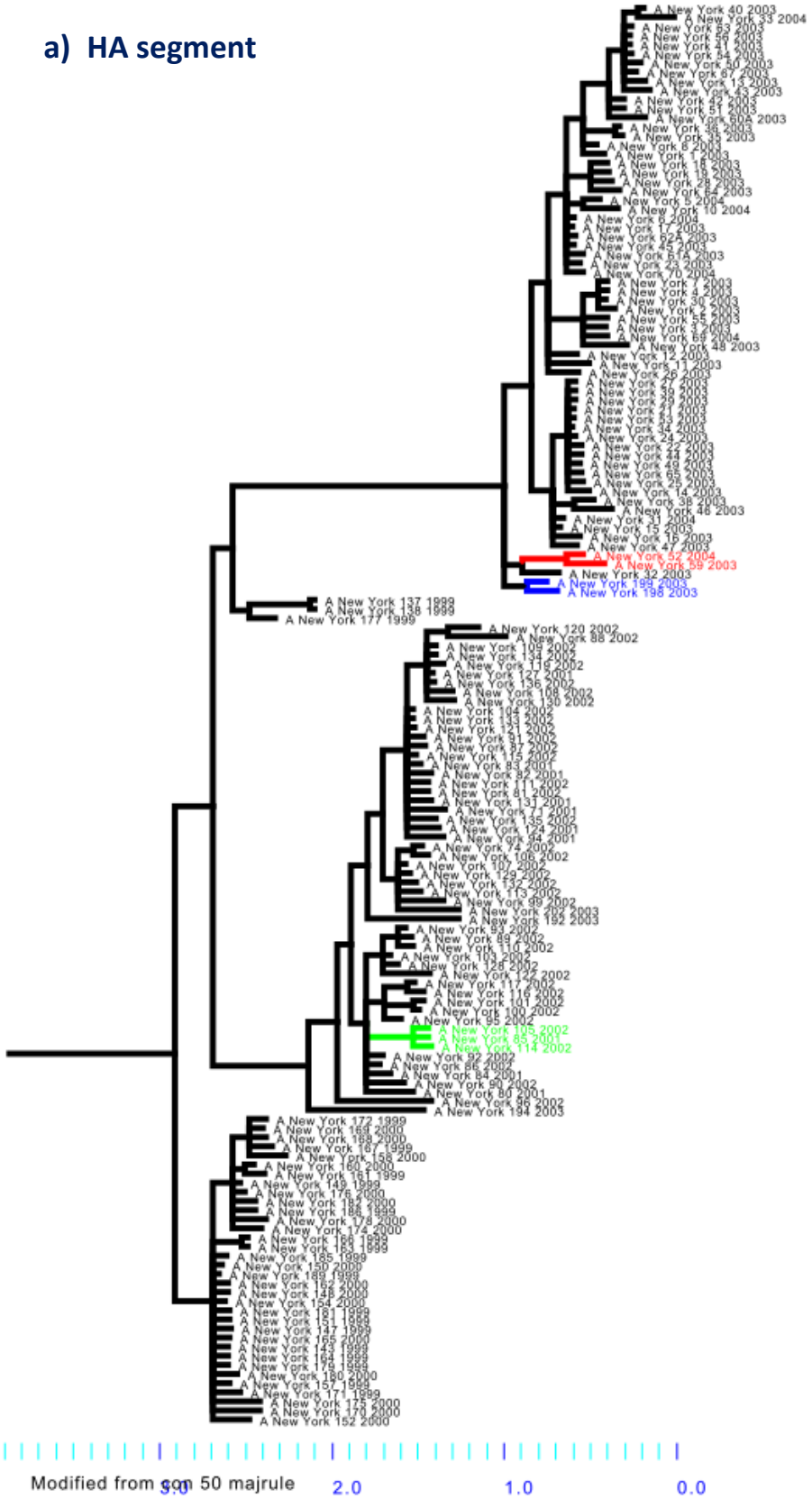
Supplementary Figure 4: Schematic for computational steps in GiRaF.

Supplementary Figure 5: Reassortment events in human influenza A (H3N2) isolates. Detailed consensus trees, in following pages, for a) HA segment and b) NA segment for the corresponding condensed trees in **Figure 1**.

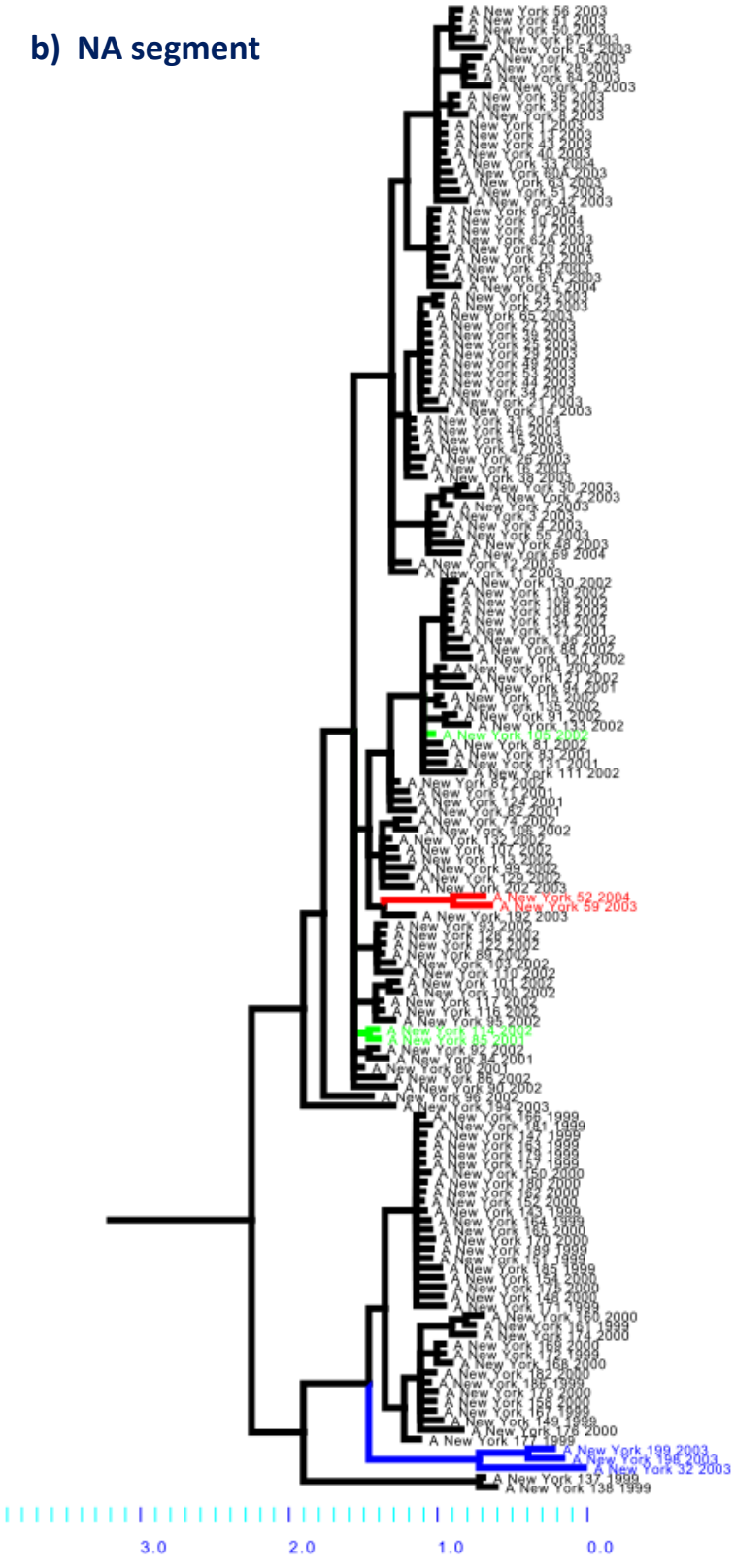
Supplementary Table 1: Catalog of reassortant H5N1 Strains. The table (additional excel file) contains a list of 18 candidate reassortments and their reassortment architectures that were automatically identified by GiRaF based on all pairwise segment comparisons for a set of H5N1 genomes.

Supplementary Table 2: Catalog of reassortant Swine Influenza & S-OIV strains. The table (additional excel file) contains a list of 37 candidate reassortments among swine influenza strains as identified by GiRaF through a fully automated analysis. Representatives of the S-OIV strains involved in the recent "swine flu" outbreak were also included in the dataset and identified correctly as reassortants by GiRaF.

a) HA segment


A New York 40 2003
A New York 33 2004
A New York 53 2003
A New York 50 2003
A New York 41 2003
A New York 54 2003
A New York 30 2003
A New York 67 2003
A New York 13 2003
A New York 43 2003
A New York 42 2003
A New York 51 2003
A New York 60A 2003
A New York 35 2003
A New York 36 2003
A New York 8 2003
A New York 1 2003
A New York 18 2003
A New York 19 2003
A New York 28 2003
A New York 64 2003
A New York 5 2004
A New York 10 2004
A New York 6 2004
A New York 17 2003
A New York 62A 2003
A New York 45 2003
A New York 61A 2003
A New York 23 2003
A New York 70 2004
A New York 7 2003
A New York 4 2003
A New York 30 2003
A New York 52 2003
A New York 57 2003
A New York 69 2004
A New York 48 2003
A New York 12 2003
A New York 11 2003
A New York 26 2003
A New York 27 2003
A New York 36 2003
A New York 39 2003
A New York 29 2003
A New York 21 2003
A New York 53 2003
A New York 34 2003
A New York 24 2003
A New York 22 2003
A New York 44 2003
A New York 49 2003
A New York 65 2003
A New York 25 2003
A New York 14 2003
A New York 38 2003
A New York 46 2003
A New York 31 2004
A New York 15 2003
A New York 16 2003
A New York 47 2003
A New York 52 2004
A New York 59 2003
A New York 42 2003
A New York 199 2003
A New York 198 2003
A New York 137 1999
A New York 138 1999
A New York 177 1999
A New York 120 2002
A New York 88 2002
A New York 109 2002
A New York 134 2002
A New York 116 2002
A New York 127 2001
A New York 136 2002
A New York 108 2002
A New York 130 2002
A New York 104 2002
A New York 133 2002
A New York 121 2002
A New York 91 2002
A New York 87 2002
A New York 115 2002
A New York 83 2001
A New York 82 2001
A New York 111 2002
A New York 81 2002
A New York 131 2001
A New York 71 2001
A New York 135 2002
A New York 124 2001
A New York 94 2001
A New York 74 2002
A New York 106 2002
A New York 107 2002
A New York 129 2002
A New York 132 2002
A New York 113 2002
A New York 99 2002
A New York 197 2003
A New York 192 2003
A New York 93 2002
A New York 89 2002
A New York 110 2002
A New York 103 2002
A New York 128 2002
A New York 122 2002
A New York 117 2002
A New York 116 2002
A New York 101 2002
A New York 100 2002
A New York 95 2002
A New York 105 2002
A New York 85 2001
A New York 114 2002
A New York 92 2002
A New York 86 2002
A New York 84 2001
A New York 90 2002
A New York 80 2001
A New York 96 2002
A New York 194 2003
A New York 172 1999
A New York 169 2000
A New York 158 2000
A New York 167 1999
A New York 160 2000
A New York 161 1999
A New York 149 1999
A New York 176 2000
A New York 187 2000
A New York 188 1999
A New York 174 2000
A New York 166 1999
A New York 163 1999
A New York 185 1999
A New York 150 2000
A New York 189 1999
A New York 162 2000
A New York 148 2000
A New York 154 2000
A New York 181 1999
A New York 151 1999
A New York 147 1999
A New York 155 2000
A New York 143 1999
A New York 191 1999
A New York 179 1999
A New York 180 2000
A New York 157 1999
A New York 171 1999
A New York 175 2000
A New York 170 2000
A New York 152 2000
3.0
2.0
1.0
0.0
Modified from con 50 majrule

b) NA segment



Modified from Untitled Tree

Alternative Tree Topology

To investigate the effect of tree topology on the results from our simulated datasets, we also generated datasets using an alternative topology. Specifically, we used a neighbor-joining tree for the HA segment of isolates from the S-OIV and swine influenza sequences and repeated the other steps of our experiment as described in the text. While the number of taxa in this set is similar to that in the original set (140 as compared to 156), the topology is more clearly divergent. In general, we obtained similarly good results using GiRaF on these datasets (for the "All Events" set, sensitivity was 94% and PPV was 90%) suggesting that the results reported in the text are a reasonable measure of GiRaF's performance in general.

Performance of Distance Test in Isolation

To assess the utility of a distance test independent of information from tree topology in predicting reassortments we experimented with a heuristic approach based on Rabadan et al. For this, we used the test described in Rabadan et al. (only steps 1 & 2) to compare all pairs of taxa and identify those that have diverged with respect to each other (bonferroni-corrected E-value threshold of 0.01). Taxa that had identical profiles of divergence were then clustered into putative reassortments and sets of size less than 20 were reported. In general, this approach performed poorly and compared to GiRaF on the "All Events" set had low sensitivity (35%) and PPV (20%) values.

Identifying Potential Parents

For reassortments identified by GiRaF, potential parents in each segment can be computed using the script "get_parents.pl" that is provided with it. This script scans the set of splits for each segment and identifies those splits that contain the reassortment set entirely on one side. The corresponding sets (that contain the reassortment set) are considered *potential parents* with a confidence value given by the split. The script then selects the smallest sets with confidence value greater than 0.5 (the least common ancestors) and among these the most confident set is reported as the potential parent in each segment.

Biclique Finding

The biclique-finding algorithm used in GiRaF is based on the idea of using the *consensus* operation to enumerate through the space of high-confidence bicliques in the incompatibility graph for two segments (say A and B). Given two bicliques (L_1, R_1) and (L_2, R_2) (where L_i and R_i are the vertex sets on the "left" and "right" of the biclique), the consensus of the bicliques is given by $(L_1 \cup L_2, R_1 \cap R_2)$. Given two *maximal* bicliques (vertex sets are not contained in another bilcique), the consensus operation can be shown to always produce another maximal biclique. The algorithm used in GiRaF then uses the following steps to enumerate all maximal bicliques (L_i, R_i) which have high confidence (i.e. $C_A(L_i) > T$ and $C_B(R_i) > T$, where $C_X(Y)$ is the confidence value associated with a set of splits Y based on trees for segment X and T is a confidence threshold parameter, that is set at 0.7 in GiRaF):

1. Let **S** be the set of all star bicliques (L_i, R_i) ($|L_i| = 1$ and R_i contains all neighbors of the node in L_i) s.t. $C_B(R_i) > T$.
2. Set **F** to **S**

3. For each biclique B_i in **S** and B_j in **F**, let D = (L, R) be the consensus of B_i and B_j. If $C_B(R) > T$, add D to **F**.
4. If new sets were added to **F** in 3, repeat 3, considering only the new additions in **F**.
5. Filter out all bicliques (L_i, R_i) in **F** where $C_A(L_i) \leq T$. Report **F** as the set of all high-confidence maximal bicliques in the graph.

While the runtime of this algorithm is linear in the number of maximal bicliques (which can be exponential in the size of the graph), in practice, the restriction to high-confidence bicliques reduces the search space drastically and makes the runtime feasible even for large graphs.