Curated Reference Gene Trees Benchmark

Brigitte Böckmann, Adrian Altenhoff

April 16, 2013

Summary

This text briefly summarizes the Reference Gene Tree Test contributed by Brigitte Böckmann. It follows closely the procedure described in [**?**].

The test operates on labeled gene trees, i.e. gene trees where the internal nodes are annotated with the type of evolutionary event shaping them, produced by Brigitte Böckmann. These trees are available on `http://wiki.isb-sib.ch/swisstree/Main_Page`.

The benchmark server test procedure works in the following way:

1. extract all pairwise relations from each of the reference trees.
2. compare set of predicted orthologs with "true" relations.
3. estimate for each family true positive rate (TPR) and false positive rate (FPR) of orthology inference.
4. produce ROC plot with average TPR/FPR over all families.

In the following we explain these steps in more details.

Labeled Gene Tree to Pairwise Relations

A rooted gene tree with N leaves induce $\binom{N}{2}$ pairwise relations. Each of these relations has one evolutionary event associated to them, i.e. the event labeled in the node in the tree where the two genes coalesce. Generally, the gene trees are bi-furcating. Multi-furcating nodes are possible, but are internally expanded to bi-furcating events with very small branches between. Since all events will be of the same type, this does not change the induced pairwise relations at all.

Comparison of predictions to true relations

For each of the projects under scrutiny, we compare its predicted orthologs to the set of induced pairwise relations. This is done on a per-family bases.

Let $\mathcal{G} = \{g_i\}$ be the set of all genes in the gene tree and $\mathcal{R}_O = \{(g_i, g_j)\}, g_i \in \mathcal{G}, g_j \in \mathcal{G}, g_i \neq g_j$ be the set of true orthologous gene-pairs according to the extracted pairwise relations from the reference gene tree. Similarly, with $\mathcal{R}_P = \{(g_i, g_j)\}$ we denote the non-orthologous, i.e. paralogous, relations in the reference tree. These relations are compared to the orthologous predictions made by every inference method:

Let $P = \{(g_i, g_j)\}$ be the set of all orthologous predictions made by the method. Now, let $P_F = (g_i, g_j)|(g_i, g_j) \in P, g_i \in \mathcal{G}, g_j \in \mathcal{G}$ be the set of orthologous predictions where both genes are members of the reference gene tree. Now, $TP = P_F \cap \mathcal{R}_O$ are the true positive predictions, $FP = P_F \cap \mathcal{R}_P$ are the false positive predictions. The true and false negative predictions are the relations not reported by the method, i.e. $FN = \mathcal{R}_O - P_F$ and $TN = \mathcal{R}_P - P_F$.

Estimation of TPR and FPR

Point estimations for TPR and FPR rate are

$$TPR = \frac{|TP|}{|TP| + |FN|}\text{and}$$

$$FPR = \frac{|FP|}{|FP| + |TN|}.$$

To estimate the uncertainty of the TPR and the FPR we model these two measures as binomial distributed random variables. Hence, their sample variance is

$$\sigma^2_{TPR} = \frac{TPR(1 - TPR)}{|TP| + |FN|}\text{and}$$

$$\sigma^2_{FPR} = \frac{FPR(1 - FRP)}{|FP| + |TN|}$$

respectively. In the resulting table we report the 95% confidence interval, i.e. $FPR \pm 1.96\sigma_{FPR}$ and $TPR \pm 1.96\sigma_{TPR}$.

ROC plots over all families

To compute the overall TPR and FPR, we model the TPR and FPR of each family as an uncorrelated random variable X_i with the above means and variances. Assuming we have n gene families to combine, we then compute the average TPR and FPR as following:

$$\overline{TPR} = \frac{1}{n}\sum_{i=1}^{n} \frac{|TP_i|}{|TP_i| + |FN_i|}, \text{and}$$

$$\overline{FRP} = \frac{1}{n}\sum_{i=1}^{n} \frac{|FP_i|}{|FP_i| + |TN_i|}.$$

And the variance of these means is

$$Var(\overline{TPR}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(TPR_i)\text{and}$$

$$Var(\overline{FPR}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(FPR_i).$$

Again, for the plots, we report the 95% confidence interval, i.e. $\overline{FPR} \pm 1.96\sqrt{Var(\overline{FPR})}$ and $\overline{TPR} \pm 1.96\sqrt{Var(\overline{TPR})}$.

References

[1] Brigitte Boeckmann, Marc Robinson-Rechavi, Ioannis Xenarios, Christophe Dessimoz, *Conceptual Framework and Pilot Study to Benchmark Phylogenomic Databases Based on Reference Gene Trees.* Briefings in Bioinformatics, 12:5 (pp. 474-484), 2011.