Introduction

I checked whether the site likelihoods of an inferred tree from the language dataset, have the same distribution as the likelihoods from data that have been indeed produced by the tree itself. I did the same for the parsimony scores of the sites for the maximum likelihood tree.

The Code

```
library(phangorn)
rm(list=ls())
set.seed(8293) ## change this for further analyses

cog.tree=read.tree("IE2011_Cognates_rel_ANNOT.nwk")
data = read.table("46glossesfull.csv", header=TRUE, row.names=1, colClasses='character')
cog.tree$tip.label[cog.tree$tip.label == "Ptg-E"] = "PtgE"
datamat = as.matrix(data)

tab <- apply(datamat, 2, function(x){table(factor(x, levels=c(0,1)))})
indsToUse <- apply(tab, 2, function(x){x[1] > 1 && x[2] > 1})
datamat <- datamat[,indsToUse]
initialdatamat <- datamat

phy = as.phyDat(datamat, type='USER', levels=c("0","1"), ambiguity="?", names=rownames(data), n=nrow(data), return.index=TRUE)

dm  <- dist.hamming(phy)
treeNJ <- NJ(dm)
fit <- pml(treeNJ, data=phy)
fitJC <- optim.pml(fit, model="GTR", optInv=TRUE, optGamma=TRUE,rearrangement ="NNI")

pdf("siteMLValues.pdf")
plot(density(fitJC$siteLik), col="red", lwd=4, ylim=c(0, 0.15), main="ML tree/site likelihoods")
siteLikes=c()
for(i in 1:300){
    ## generate some simulated data for the tree
    d = simSeq(fitJC, l=219)
    ## calculate the site likelihoods
```

```
    fit <- pml(tree=fitJC$tree, data=d)
    ## store the site likelihoods
    siteLikes = c(siteLikes, fit$siteLik)
    ## make a plot
    points(density(fit$siteLik), col="gray", type='l')
}
points(density(fitJC$siteLik), col="red", lwd=4, type='l')
legend("topright", legend=c("real inferred", "simulated true
   tree"), col=c("red" ,"gray"), lwd=2)
dev.off()



dm  <- dist.hamming(phy)
treeNJ <- NJ(dm)
fit <- pml(treeNJ, data=phy)
fitJC <- optim.pml(fit, model="GTR", optInv=TRUE, optGamma=
   TRUE,rearrangement ="NNI")
fitpars = parsimony(fitJC$tree, data=phy, method = "fitch",
   cost = NULL, site = "site")
pdf("siteParsValues.pdf")
plot(density(fitpars), col="red", lwd=4, ylim=c(0, 0.4), main
   ="ML trees/Parsimony scores")
siteScores=c()
i=1
for(i in 1:300){
    ## generate some simulated data for the tree
    d = simSeq(fitJC, l=219)
    ## calculate the site likelihoods
    pars = parsimony(fitJC$tree, data=d, method = "fitch",
   cost = NULL, site = "site")
    ## store the site likelihoods
    siteScores = c(siteScores, pars)
    ## make a plot
    points(density(pars), col="gray", type='l')
}
points(density(fitpars), col="red", lwd=4, type='l')
legend("topright", legend=c("real inferred", "simulated true
   tree"), col=c("red" ,"gray"), lwd=2)
dev.off()
```

Results

The main results are presented in the two figures below

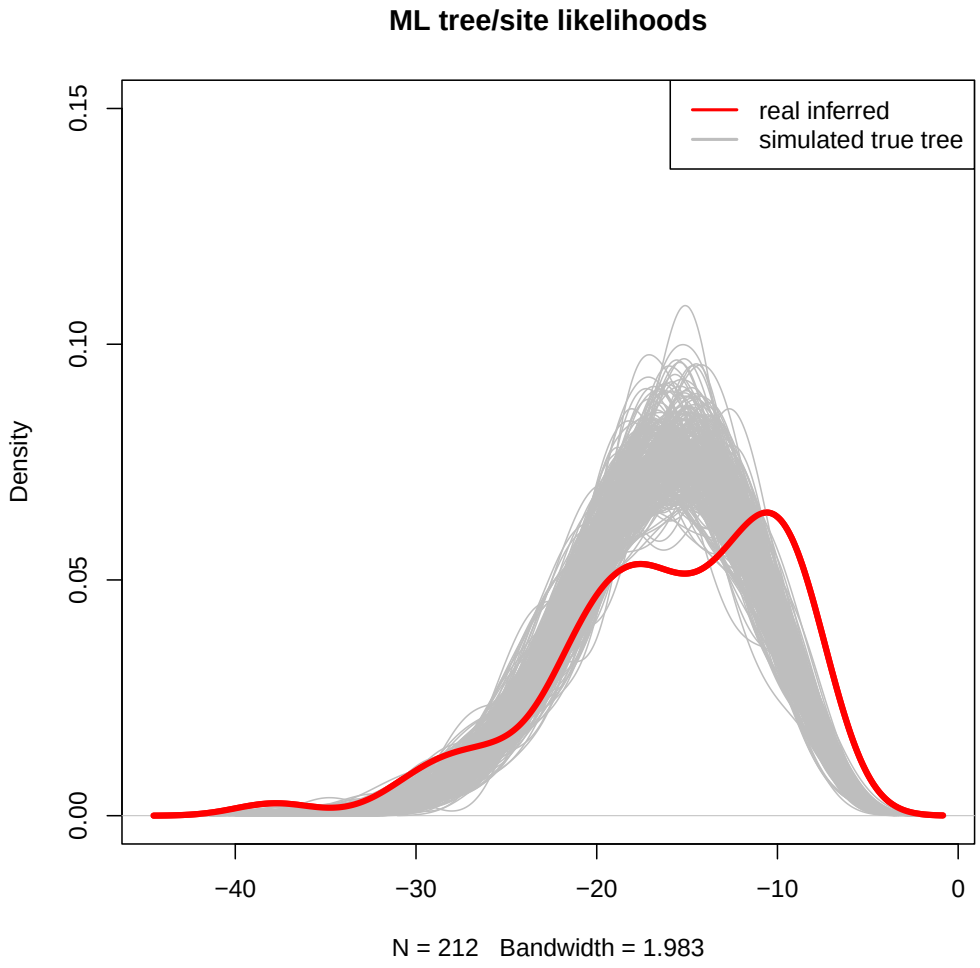

Figure 1: gray color: simulated data on the same ML tree as the inferred tree. Red color: real data on the inferred tree. Each line represents the density plot of the site likelihoods.

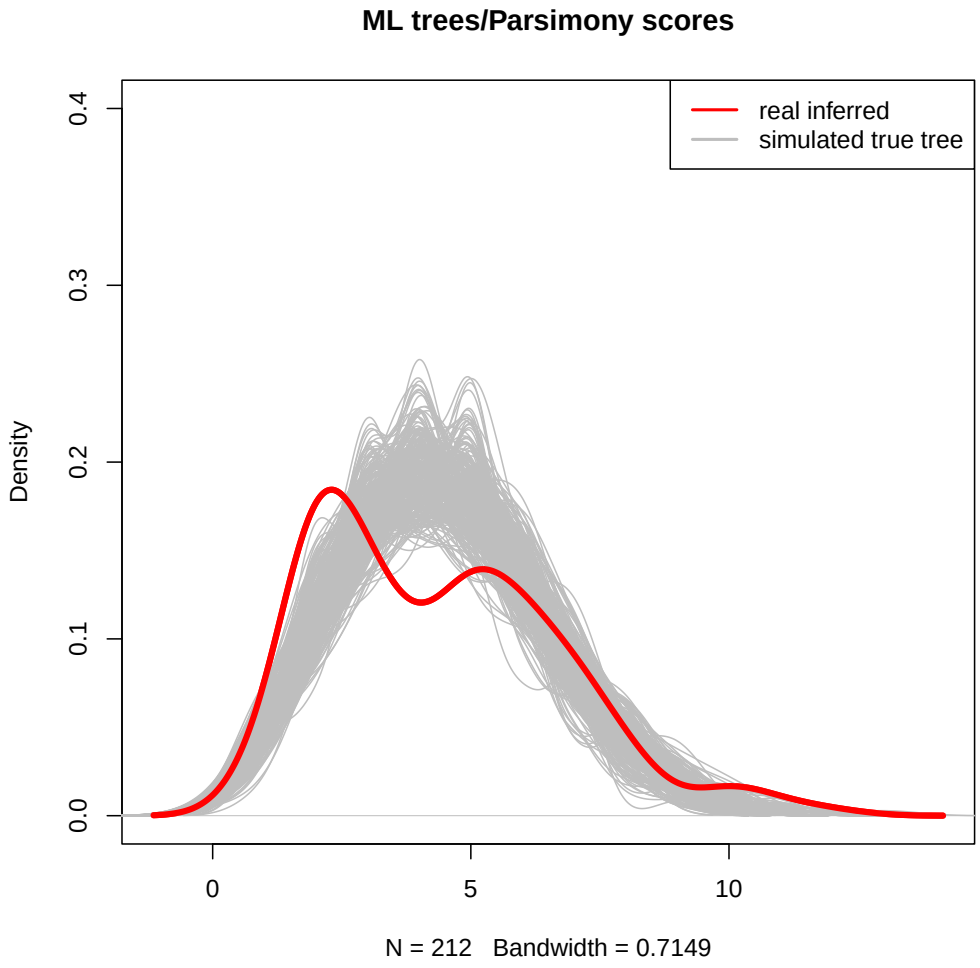

Figure 2: gray color: simulated data on the same ML tree as the inferred tree. Red color: real data on the inferred tree. Each line represents the density plot of the site parsimony scores.

Conclusion

It seems that the red distribution is qualitatively different than the gray ones. This may mean that the tree inferred from the language data, in fact, might be a non-correct tree. However, I understand that this is not a right way to show this. It's interesting that the distribution shows two peaks. Thus, it might be that in fact I have a mixture of two distributions, i.e., some sites produce a good score and some other a much worse score.