

## USING THE ONCOTREE PACKAGE

ANIKO SZABO, KENNETH BOUCHER AND LISA PAPPAS

ABSTRACT. This paper shows a short example of building and exploring oncogenetic trees using the Oncotree package. A detailed description of the theory of oncogenetic trees can be found in

- Desper, R.; Jiang, F.; Kallioniemi, O.; Moch, H.; Papadimitriou, C. and Schäffer, A. A. “Inferring tree models for oncogenesis from comparative genome hybridization data.” *Journal of Computational Biology*, 1999, 6, 37-51.
- Szabo, A. and Boucher, K. “Estimating an oncogenetic tree when false negatives and positives are present.” *Mathematical Biosciences*, 2002, 176, 219-236.
- Szabo, A. and Boucher, K. “Oncogenetic trees” in *Handbook of cancer models with applications* Tan, Hanin (ed.) World Scientific, 2008.

A short introduction is given in doc/Oncotree.pdf.

We start by loading a dataset. The package contains an example dataset:

```
> library(Oncotree)
> data(ov.cgh)
> str(ov.cgh)
'data.frame':      87 obs. of  7 variables:
 $ 8q+: int  0 0 1 1 0 1 1 0 0 1 ...
 $ 3q+: int  0 0 1 0 0 1 1 1 1 0 ...
 $ 5q-: int  0 0 1 0 0 1 1 1 0 1 ...
 $ 4q-: int  0 1 1 0 0 1 1 0 0 1 ...
 $ 8p-: int  0 0 0 0 0 1 1 0 0 1 ...
 $ 1q+: int  1 1 0 0 0 0 0 0 0 1 ...
 $ Xp-: int  0 0 0 0 0 0 1 0 1 1 ...
```

Based on these data, we construct the oncogenetic tree using the default  $\ell_2$ -distance error function to estimate the false-positive and false-negative error rates.

```
> ov.tree <- oncotree.fit(ov.cgh)
```

The fitted tree can be examined several ways: `printing` it produces a quick summary, but the result of `plotting` is easier to interpret (the plots are shown in Figure 1).

```
> ov.tree
```

Oncogenetic tree from 7 events

Parent function:

```
8q+ <- Root
3q+ <- 8q+
5q- <- Root
4q- <- 5q-
8p- <- 5q-
1q+ <- Root
Xp- <- 8p-
```

Estimated error rates: epos= 0.2084556 , eneg= 0.0267696

```
> plot(ov.tree, edge.weights="est")
> pstree.oncotree(ov.tree, edge.weights="est", shape="oval")
```

We can compare the observed and fitted marginal occurrence frequencies of the mutations (the distance between these two was minimized for the error-rate estimation). The plot is shown in Figure 2.

```
> print(obs <- colMeans(ov.tree$data))
      Root      8q+      3q+      5q-      4q-      8p-      1q+      Xp-
1.0000000 0.7011494 0.5517241 0.5287356 0.5057471 0.4712644 0.4367816 0.4252874
> print(est <- marginal.distr(ov.tree, with.errors=TRUE))
```

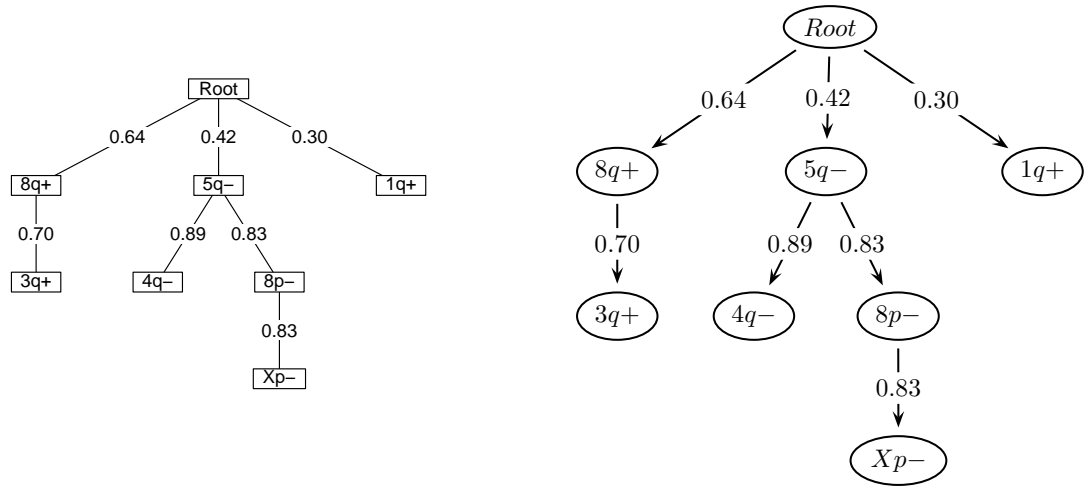
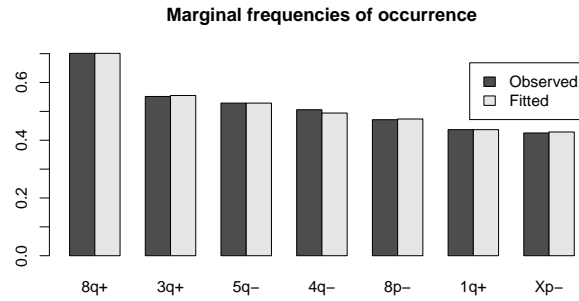
FIGURE 1. Fitted oncogenetic tree for the `ov.cgh` data set.

FIGURE 2. Observed and fitted frequencies of occurrence of each event.

```

      Root      8q+      3q+      5q-      4q-      8p-      1q+      Xp-
1.0000000 0.7011494 0.5550202 0.5287356 0.4943105 0.4736503 0.4367816 0.4286807
> #plot is in Figure 2
> barplot(rbind(obs[-1],est[-1]), beside=T, legend.text=c("Observed","Fitted"),
+         main="Marginal frequencies of occurrence")

```

In addition to the marginal frequencies, it is possible to estimate the entire joint distribution generated by the tree:

```

> dd <- distribution.oncotree(ov.tree, with.errors=TRUE)
> head(dd)

```

	Root	8q+	3q+	5q-	4q-	8p-	1q+	Xp-	Prob
1	1	0	0	0	0	0	0	0.029222901	
2	1	1	0	0	0	0	0	0.027992097	
3	1	0	1	0	0	0	0	0.009202964	
4	1	1	1	0	0	0	0	0.062160896	
5	1	0	0	1	0	0	0	0.008323722	
6	1	1	0	1	0	0	0	0.007973145	

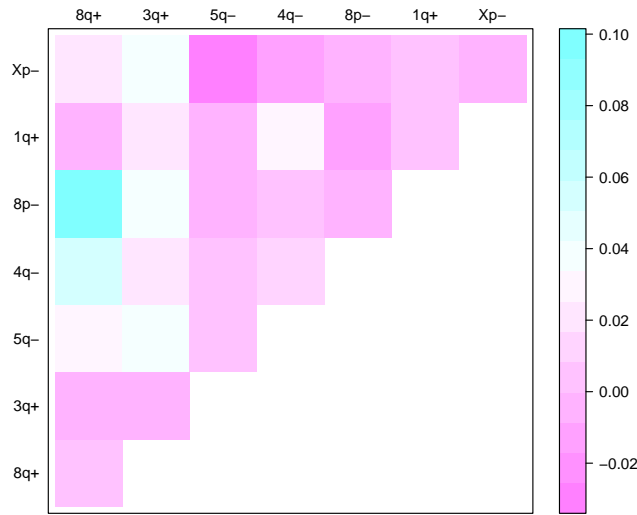
**Observed – Expected probabilities of co-occurrence of events**

FIGURE 3. Goodness-of-fit plot: difference between observed and expected probabilities of two events being observed.

Using the overall joint distribution, it is straightforward to obtain marginal joint distributions (2- or higher way) if needed (the plot is shown in Figure 3).

```
> #estimated probabilities of 2 events
> print(est2way <- t(data.matrix(dd[2:8])) %*% diag(dd$Prob) %*% data.matrix(dd[2:8]))
```

	8q+	3q+	5q-	4q-	8p-	1q+	Xp-
8q+	0.7011494	0.4834457	0.3707227	0.3465855	0.3320996	0.3062492	0.3005693
3q+	0.4834457	0.5550202	0.2934589	0.2743523	0.2628855	0.2424226	0.2379265
5q-	0.3707227	0.2934589	0.5287356	0.3884206	0.3683135	0.2309420	0.3245477
4q-	0.3465855	0.2743523	0.3884206	0.4943105	0.3393380	0.2159057	0.2992688
8p-	0.3320996	0.2628855	0.3683135	0.3393380	0.4736503	0.2068817	0.3130649
1q+	0.3062492	0.2424226	0.2309420	0.2159057	0.2068817	0.4367816	0.1872399
Xp-	0.3005693	0.2379265	0.3245477	0.2992688	0.3130649	0.1872399	0.4286807

```
> #observed probabilities of 2 events
> print(obs2way <- t(ov.tree$data[, -1]) %*% ov.tree$data[, -1]/nrow(ov.tree$data))
```

	8q+	3q+	5q-	4q-	8p-	1q+	Xp-
8q+	0.7011494	0.4827586	0.4022989	0.4022989	0.4252874	0.2988506	0.3218391
3q+	0.4827586	0.5517241	0.3333333	0.2988506	0.2988506	0.2643678	0.2758621
5q-	0.4022989	0.3333333	0.5287356	0.3908046	0.3678161	0.2298851	0.2988506
4q-	0.4022989	0.2988506	0.3908046	0.5057471	0.3448276	0.2413793	0.2873563
8p-	0.4252874	0.2988506	0.3678161	0.3448276	0.4712644	0.1954023	0.3103448
1q+	0.2988506	0.2643678	0.2298851	0.2413793	0.1954023	0.4367816	0.1954023
Xp-	0.3218391	0.2758621	0.2988506	0.2873563	0.3103448	0.1954023	0.4252874

```
> oe.diff <- obs2way-est2way
> oe.diff[lower.tri(oe.diff)] <- NA #clear half of symmetric matrix for plotting
> require(lattice) #the plot is in Figure 3
> levelplot(oe.diff, xlab="", ylab="", scales=list(x=list(alternating=2), tck=0),
+           main="Observed – Expected probabilities of co-occurrence of events")
```

Another way to evaluate goodness-of-fit is through bootstrap resampling of the data. Two approaches are implemented: a parametric bootstrap that assumes that the model is correct and a non-parametric bootstrap. The plot is shown in Figure 4.

```
> set.seed(43636)
> ov.boot <- bootstrap.oncotree(ov.tree, type="nonparam", R=1000)
> ov.boot
```

Out of the 1000 replicates there are 309 unique trees with frequencies from 83 down to 1  
The bootstrap process found the original tree 83 times

```
> opar <- par(mfrow=c(3,2)) #the plot is in Figure 4
> plot(ov.boot, minfreq=45)
> par(opar)
```

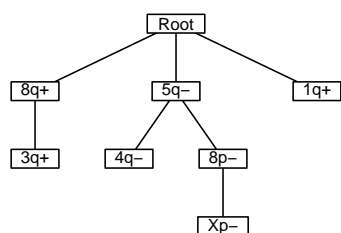
The non-parametric bootstrap gives an estimate of the reconstruction confidence: the original tree was obtained 83 times out of 1000 resamples, so the estimated confidence is 8.3%.

We can look at the frequency of edge occurrences in the bootstrapped trees:

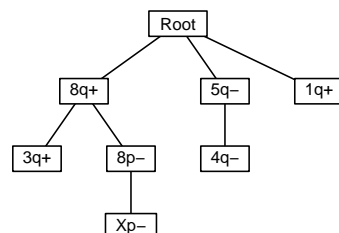
```
> ov.boot$parent.freq
```

	Child								
Parent	Root	8q+	3q+	5q-	4q-	8p-	1q+	Xp-	
	1000	0	0	0	0	0	0	0	0
Root	0	997	69	519	225	4	807	67	
8q+	0	0	929	89	162	409	24	7	
3q+	0	2	0	44	0	0	94	42	
5q-	0	1	2	0	522	399	9	169	
4q-	0	0	0	275	0	143	50	116	
8p-	0	0	0	70	84	0	4	599	
1q+	0	0	0	0	4	0	0	0	
Xp-	0	0	0	3	3	45	12	0	

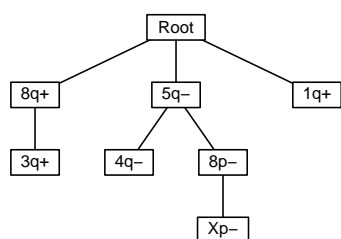
It is clear that some edges are really stable: Root  $\rightarrow$  8q+, 8q+  $\rightarrow$  3q+, root  $\rightarrow$  1q+, all with confidence  $> 80\%$ , while other edges are less stable (for example, 8p- is the child of 8q+ about as often as of 5q-).



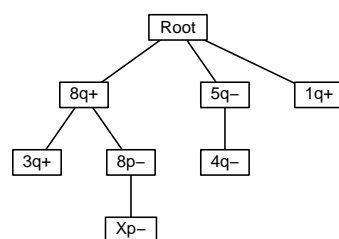
Original Tree



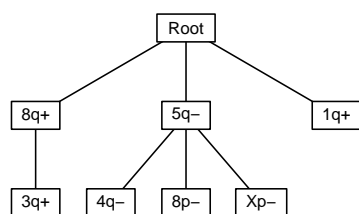
Tree based on most frequent parent



Observed Frequency = 83



Observed Frequency = 63



Observed Frequency = 49

FIGURE 4. The most frequently occurring bootstrap trees.