A hidden Markov model for CRLMM output

Robert Scharpf

January 12, 2010

See the copynumber.Rnw vignette in crlmm/inst/scripts for additional information on locus-level estimation of copy number.

```
> library(IRanges)
> library(VanillaICE)
> library(crlmm)
> celFiles <- list.celfiles("/thumper/ctsa/snpmicroarray/hapmap/raw/affy/1m",
+     full.names = TRUE, pattern = ".CEL")
> batch <- substr(basename(celFiles), 13, 13)
> celFiles <- celFiles[batch == "C" | batch == "Y"]
> batch <- batch[batch == "C" | batch == "Y"]
> cnOpts <- cnOptions(cdfName = "genomewidesnp6", outdir = "/thumper/ctsa/beaty/scharpf/crlmmOut/hapmap
+     batch = batch, chromosome = 21)
> if (!file.exists(file.path(cnOpts[["outdir"]], "cnSet_21.rda"))) {
+     message("Processing ...")
+     crlmmCopynumber(celFiles, cnOpts)
+ } else {
+     if (!exists("cnSet")) {
+         message("Loading ...")
+         load(file.path(cnOpts[["outdir"]], "cnSet_21.rda"))
+     }
+     cols <- grep("nuA_", fvarLabels(cnSet))
+ }
```

Remove loci for which there are missing values for the linear model parameters.

```
> cnSet <- cnSet[rowSums(is.na(fData(cnSet)[, cols])) == 0, ]
```

Compute transition and emission probabilities using the hmmOptions.

```
> initialPr <- c((1 - 0.99)/3, (1 - 0.99)/3, 0.99, (1 - 0.99)/3)
> if (!exists("hmmOpts")) {
+     hmmOpts <- hmmOptions(cnSet, copynumberStates = 0:3, log.initial = log(initialPr),
+         states = c("hom-del", "hem-del", "normal", "amp"), normalIndex = 3)
+ }
```

The R function hmm returns an object of class RangedData with start and end coordinates of the state path obtained from the Viterbi algorithm. The log likelihood ratio (LLR) compares the log likelihood of the predicted state sequence to the null (normal copy number).

```
> if (!any(is.na(hmmOpts[["log.emission"]]))) {
+     if (!file.exists(file.path(cnOpts[["outdir"]], "fit_hmm.rda"))) {
+         fit <- hmm(cnSet, hmmOpts)
+         save(fit, file = file.path(cnOpts[["outdir"]], "fit_hmm.rda"))
+     }
+     else {
+         if (!exists("fit")) {
+             load(file.path(cnOpts[["outdir"]], "fit_hmm.rda"))
+         }
+     }
+     fit
+ }
```

```
RangedData with 617 rows and 4 value columns across 1 space
         space               ranges |          sampleId     state numMarkers
   <character>            <IRanges> |       <character> <integer>  <integer>
1        chr21 [ 9758730, 10197771] | NA06985_GW6_C.CEL         3        127
2        chr21 [13267528, 46921373] | NA06985_GW6_C.CEL         3      24624
3        chr21 [ 9758730, 10197771] | NA06991_GW6_C.CEL         3        127
4        chr21 [13267528, 46921373] | NA06991_GW6_C.CEL         3      24624
5        chr21 [13267528, 46921373] | NA06993_GW6_C.CEL         3      24624
6        chr21 [ 9758730, 10179848] | NA06993_GW6_C.CEL         3        121
7        chr21 [10180148, 10197771] | NA06993_GW6_C.CEL         2          6
8        chr21 [ 9758730, 10197771] | NA06994_GW6_C.CEL         3        127
9        chr21 [13267528, 46921373] | NA06994_GW6_C.CEL         3      24624
10       chr21 [ 9758730, 10197771] | NA07000_GW6_C.CEL         3        127
         LLR
   <numeric>
1    0.00000
2    0.00000
3    0.00000
4    0.00000
5    0.00000
6    0.00000
7   10.67742
8    0.00000
9    0.00000
10   0.00000
...
<607 more rows>
```

1 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 2.11.0 Under development (unstable) (2009-11-22 r50541), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.iso885915, LC_NUMERIC=C, LC_TIME=en_US.iso885915, LC_COLLATE=en_US.iso885915, LC_MONETARY=C, LC_MESSAGES=en_US.iso885915, LC_PAPER=en_US.iso885915, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.iso885915, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: Biobase 2.7.2, crlmm 1.5.20, IRanges 1.5.21, oligoClasses 1.9.22, VanillaICE 1.9.1
- Loaded via a namespace (and not attached): affyio 1.15.1, annotate 1.25.0, AnnotationDbi 1.9.2, Biostrings 2.15.11, DBI 0.2-4, ellipse 0.3-5, genefilter 1.29.3, mvtnorm 0.9-8, preprocessCore 1.9.0, RSQLite 0.7-3, SNPchip 1.11.1, splines 2.11.0, survival 2.35-7, tools 2.11.0, xtable 1.5-6