Motif divergence between orthologous pairs of enhancer sequences

This vignette contains a brief example of how to use the `motifDiverge` package. The analysis focuses on a set of five human–mouse orthologous enhancer sequences (from []), and quantifies divergence in terms of the Nkx-2.5 motif.

Step 1: Obtain the sequence pairs, each as a `DNAStringSet` in `R`.

```
> require(motifDiverge)
> require(Biostrings)
> require(MotifDb)
> enh.hg.file = system.file( "extdata", "enh_human.fa",
+                            package="motifDiverge"   )
> enh.mm.file = system.file( "extdata", "enh_mouse.fa",
+                            package="motifDiverge"   )
> enh.hg = readDNAStringSet(enh.hg.file)
> enh.mm = readDNAStringSet(enh.mm.file)
> enh.hg

  A DNAStringSet instance of length 5
    width seq                                               names
[1]  2907 TTTAGCTTCCTGTCTAAGGGA...GTTAAGGACAGGCTGTGGGG hg18_chr5_5988
[2]  1451 AGTAGAGGCCTCCATGGGGTT...TTCCCAAAAGAGTGGAGAGC hg18_chr11_1298
[3]  3561 CAGTGGCCACAGGCCCTTCTG...AGCATTGTGAGGTGCCCTGA hg18_chr5_6257
[4]  1921 TTACCCCTCATTTACTCCTGC...TTCTACAAACCAGTTTTTTA hg18_chr11_1320
[5]  7341 CATCATTTAAAAAAACTAAAT...AGTAACTTGCCCCAAATCAA hg18_chr16_3051

> enh.mm

  A DNAStringSet instance of length 5
    width seq                                               names
[1]  4350 AGTAGGCTCCCCTCTAAAGTG...TTAGTCATTGCCACCAGCAT mm9_chr5_5988
[2]  1450 CGGGTGCTCTTACCCACTGAG...AGGACTAGAGAGTGGCTCCC mm9_chr11_1298
[3]  4000 TAAAAAGCTAAACAGACAAGG...TGTGGGTCCCTCCTACTGGC mm9_chr5_6257
[4]  2225 TTACCCCTGAGCCTCCCCCAA...CCTGGCAGTGGTGGCGCACG mm9_chr11_1320
[5]  8875 CAAGTTTATAAATTTTTTTTA...CATAACTCCCTCAAGGTCTT mm9_chr16_3051
```

Step 2: Obtain the motif for Nkx-2.5. First get the JASPAR [] position frequency matrix using `MotifDb`, and then use this as basis for a position specific score matrix. Also, the frequency matrix is regularized using a pseudocount.

```
> providerId = "MA0063.1" #- Jaspar NKX-2.5
> index      = grep(providerId,values(MotifDb)$providerId)
> pspm       = MotifDb[index][[1]]
> pssm       = pspmToPssm(pspm)
> pspm       = pspmToPssm(pspm,return.pspm=TRUE)$pspm
```

Step 3: Next, numerically calculate a score cutoff (for the `pssm` and its reverse complement, such that the Type I error rate is 1%. This example uses the observed sequence composition as a null model.

```
> bg  = colSums(alphabetFrequency(c(enh.hg,enh.mm))[,1:4])
> bg  = bg/sum(bg)
> cut = scoreCutFromErr(err=.01, pssm=pssm, pspm=pspm, bg=bg, type="type1")
```

Step 4a: For each of the sequence pair, calculate the model parameters specifying two correlated Bernoulli trials. This version does not assume an evolutionary model and would also be appropriate for non-homologous, independent sequences.

```
> pars.nomodel = cbernEstimateModelPars(  seqs.x = enh.mm,
+                                         seqs.y = enh.hg,
+                                         pssm   = pssm  ,
+                                         pspm   = pspm  ,
+                                         cut.fw = cut   ,
+                                         cut.rc = cut   )
```

Step 4b: This time estimate model parameters assuming an evolutionary model based on the UCSC conservation track (for instance `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates.mod`). The background frequencies are (again, see Step 3) adjusted to reflect the nucleotide composition of the sequences at hand.

```
> require(rphast)

Loading required package:  rphast
```

```
Attaching package:  'rphast'

The following object is masked from 'package:Biostrings':

    complement

> neutral.mod = get.neutralMod(evo.mod.file)       #- point ucscURL to model
> neutral.mod = mod.backgd.tm(neutral.mod,bg) #- adjust background
> pspm.mods   = get.modList(neutral.mod,pspm) #- models for pspm columns
> pars.model  = cbernEstimateModelPars(  seqs.x    = enh.mm   ,
+                                        seqs.y    = enh.hg   ,
+                                        pssm      = pssm     ,
+                                        pspm      = pspm     ,
+                                        cut.fw    = cut      ,
+                                        cut.rc    = cut      ,
+                                        indep     = FALSE    ,
+                                        useCounts = FALSE    ,
+                                        modList   = pspm.mods)

aligning sequences; keeping stand
seqence 1 of 5
seqence 2 of 5
seqence 3 of 5
seqence 4 of 5
seqence 5 of 5
xxxxx
```

Step 5: Calculate enrichment and depletion p–values according to the tail probabilities of the model with the estimated parameters:

```
> pvals.enr = apply(pars.model[,1:6],1,function(x) pcbern(x[1],x[2],x[3],
+                                                         x[4],x[5],x[6],
+                                                         lower.tail=F))
> pvals.dep = apply(pars.model[,1:6],1,function(x) pcbern(x[1],x[2],x[3],
+                                                         x[4],x[5],x[6],
+                                                         lower.tail=T))
```

The fifth sequence pair shows a significant depletion of Nkx-2.5 motifs in the mouse

sequence: Even thought the mouse sequence is longer (8,869bp vs. 7,335bp for human), it has fewer motif instances (53 compared to 66 in human).