iSeq

March 24, 2012

`iSeq1` *Bayesian modeling of ChIP-seq data through hidden Ising models*

Description

iSeq1 implements the method that models the bin-based tag counts using Poisson-Gamma distribution and the hidden states of the bins using a standard 1D Ising model.

Usage

```
iSeq1(Y,gap=300,burnin=500,sampling=2000,ctcut=0.95,a0=1,b0=1,a1=5,b1=1,
  k0=3,mink=0,maxk=10,normsd=0.1,verbose=FALSE)
```

Arguments

`Y`	Y should be a data frame containing the first 4 columns of the data frame returned by function 'mergetag()'. The columns 1-4 of Y are chromosome IDs, start position of the bin, end position of the bin, tag counts in the bins. For one-sample analysis, the tag counts can be the number of forward and reverse tags falling in the bins. For two-sample analysis, tag counts are the adjusted counts of ChIP samples, which are obtained by subtracting the control tag counts from corresponding ChIP tag counts for each bin. If the user provides his/her own Y, Y must be firstly sorted by the chromosome ID, then by the start position, and then by the end position.
`gap`	gap is the average length of the sequenced DNA fragments. If the distance between two nearest bins is greater than 'gap', a bin with 0 tag count is inserted into the two neighboring bins for modeling.
`burnin`	The number of MCMC burn-in iterations.
`sampling`	The number of MCMC sampling iterations. The posterior probability of enriched and non-enriched state is calculated based on the samples generated in the sampling period.
`ctcut`	A value used to set the initial state for each window/bin. If tag count of a bin is greater than quantile(Y[,4],probs=ctcut), its state will be set to 1, otherwise -1. For typical ChIP-seq data, because the major regions are non-enriched, a good value for ctcut could be in the interval (0.9, 0.99).
`a0`	The scale hyper-parameter of the Gamma prior, alpha0.

b0	The rate hyper-parameter of the Gamma prior, beta0.
a1	The scale hyper-parameter of the Gamma prior, alpha1.
b1	The rate hyper-parameter of the Gamma prior, beta1.
k0	The initial parameter used to control the strength of interaction between neighboring bins, which must be a positive value (k0>0). A larger value of kappa represents a stronger interaction between neighboring bins.
mink	The minimum value of k(kappa) allowed.
maxk	The maximum value of k(kappa) allowed.
normsd	iSeq1 uses a Metropolis random walk proposal for sampling from the posterior distributions of the model parameter kappa. The proposal distribution is a normal distribution with mean 0 and standard deviation specified by normsd.
verbose	A logical variable. If TRUE, the number of completed MCMC iterations is reported.

Value

A list with the following elements.

pp	The posterior probabilities of bins in the enriched state.
kappa	The posterior samples of the interaction parameter of the Ising model.
lambda0	The posterior samples of the model parameter lambda0
lambda1	The posterior samples of the model parameter lambda1.

Author(s)

Qianxing Mo <moq@mskcc.org>

References

Qianxing Mo (2011). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, Advance Access published September 13, 2011. doi:10.1093/biostatistics/kxr029

See Also

iSeq2,peakreg,mergetag,plotreg

Examples

```
data(nrsf)
chip = rbind(nrsf$chipFC1592,nrsf$chipFC1862,nrsf$chipFC2002)
mock = rbind(nrsf$mockFC1592,nrsf$mockFC1862,nrsf$mockFC2002)
tagct = mergetag(chip=chip,control=mock,maxlen=80,minlen=10,ntagcut=10)
tagct22 = tagct[tagct[,1]=="chr22",]
res1 = iSeq1(Y=tagct22[,1:4],gap=200,burnin=200,sampling=500,
ctcut=0.95,a0=1,b0=1,a1=5,b1=1,k0=3,mink=0,maxk=10,normsd=0.1,verbose=FALSE)

reg1 = peakreg(tagct22[,1:3],tagct22[,5:6]-tagct22[,7:8],res1$pp,0.5,
       method="ppcut",maxgap=200)

reg2 = peakreg(tagct22[,1:3],tagct22[,5:6]-tagct22[,7:8],res1$pp,0.05,
        method="fdrcut",maxgap=200)

ID = (reg1[1,4]):(reg1[1,5])
plotreg(tagct22[ID,2:3],tagct22[ID,5:6],tagct22[ID,7:8],peak=reg1[1,6])
```

`iSeq2` *Bayesian hierarchical modeling of ChIP-seq data through hidden Ising models*

Description

iSeq2 implements the method that models the bin-based tag counts using Poisson-Gamma distribution and the hidden states of the bins using a hidden high-order Ising model.

Usage

```
iSeq2(Y,gap=300,burnin=500,sampling=2000,winsize=2,ctcut=0.95,
      a0=1,b0=1,a1=5,b1=1,k=3,verbose=FALSE)
```

Arguments

`Y`	Y should be a data frame containing the first 4 columns of the data frame returned by function 'mergetag()'. The columns 1-4 of Y are chromosome IDs, start positions of the bins, end positions of the bins, tag counts in the bins. For one-sample analysis, the tag counts can be the number of forward and reverse tags falling in the bins. For two-sample analysis, tag counts are the adjusted counts of ChIP samples, which are obtained by subtracting the control tag counts from corresponding ChIP tag counts for each bin. If the user provides his/her own Y, Y must be firstly sorted by the chromosome ID, then by the start position, and then by the end position.
`gap`	gap is the average length of the sequenced DNA fragments. If the distance between two nearest bins is greater than 'gap', a bin with 0 tag count is inserted into the two neighboring bins for modeling.
`burnin`	The number of MCMC burn-in iterations.
`sampling`	The number of MCMC sampling iterations. The posterior probability of enriched and non-enriched state is calculated based on the samples generated in the sampling period.
`winsize`	The parameter to control the order of interactions between genomic regions. For example, winsize = 2, means that genomic region i interacts with regions i-2,i-1,i+1 and i+2. A balance between high sensitivity and low FDR could be achieved by setting winsize = 2.
`ctcut`	A value used to set the initial state for each genomic bin. If tag count of a bin is greater than quantile(Y[,4],probs=ctcut), its state will be set to 1, otherwise -1. For typical ChIP-seq data, because the major regions are non-enriched, a good value for ctcut could be in the interval (0.9, 0.99).
`a0`	The scale hyper-parameter of the Gamma prior, alpha0.
`b0`	The rate hyper-parameter of the Gamma prior, beta0.
`a1`	The scale hyper-parameter of the Gamma prior, alpha1.
`b1`	The rate hyper-parameter of the Gamma prior, beta1.
`k`	The parameter used to control the strength of interaction between neighboring bins, which must be a positive value (k>0). The larger the value of k, the stronger iterations between neighboring bins. The value for k may not be too small (e.g. < 1.0). Otherwise, the Ising system may not be able to reach a super-paramagnetic state.

verbose — A logical variable. If TRUE, the number of completed MCMC iterations is reported.

Value

A list with the following elements.

pp	The posterior probabilities of the bins in the enriched state.
lambda0	The posterior samples of the model parameter lambda0
lambda1	The posterior samples of the model parameter lambda1.

Author(s)

Qianxing Mo <moq@mskcc.org>

References

Qianxing Mo, Faming Liang. (2010). Bayesian modeling of ChIP-chip data through a high-order Ising model. *Biometrics*, 66(4), 1284-94.

Qianxing Mo (2011). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, Advance Access published September 13, 2011. doi:10.1093/biostatistics/kxr029

See Also

iSeq1, peakreg, mergetag, plotreg

Examples

```
data(nrsf)
chip = rbind(nrsf$chipFC1592,nrsf$chipFC1862,nrsf$chipFC2002)
mock = rbind(nrsf$mockFC1592,nrsf$mockFC1862,nrsf$mockFC2002)
tagct = mergetag(chip=chip,control=mock,maxlen=80,minlen=10,ntagcut=10)
tagct22 = tagct[tagct[,1]=="chr22",]
res2 = iSeq2(Y=tagct22[,1:4],gap=200, burnin=100,sampling=500,winsize=2,ctcut=0.95,
  a0=1,b0=1,a1=5,b1=1,k=1.0,verbose=FALSE)
```

---

mergetag — *Aggregate sequence tags into dynamic genomic windows/bins and count the number of tags in the windows/bins.*

Description

A function to aggregate sequence tags into genomic windows/bins with dynamic length specified by the user and count the number of tags falling in the dynamic windows/bins.

Usage

```
mergetag(chip,control,maxlen=80,minlen=10,ntagcut=10)
```

Arguments

chip	A n by 3 matrix or data frame. The Rows correspond to sequence tags. chip[,1] contains chromosome IDs; chip[,2] contains the genomic positions of sequence tags matched to the reference genome. For each tag, in order to accurately infer the true binding sites, we suggest using the middle positions of the tags as the tags' positions on the chromosomes. Note a genomic position must be an integer. chip[,3] contains the direction indicators of the sequence tags. The user can basically use any symbols to represent the forward or reverse chains. Function 'mergetag' use integer 1 and 2 to represent the directions of the chains by doing as.numeric(as.factor(chip[,3])). Therefore, the user should know the directions referred by integer 1 and 2. For example, if the forward and reverse chains are represented by 'F' and 'R', respectively, then chains 1 and 2 will refer to the forward and reverse chain, respectively. In the output, the tag counts are summarized for chains 1 and 2, respectively (see the below for details).
control	A n by 3 matrix or data frame. The column names of control must be the same as the column names of chip.
maxlen	The maximum length of the genomic window/bin into which sequence tags are aggregated.
minlen	The minimum length of the genomic window/bin into which sequence tags are aggregated.
ntagcut	The tag count cutoff value for triggering bin size change. For example, suppose L_i and C_i are the length and tag count for bin i, respectively. If C_i >= ntagcut, the length for bin i+1 will be min(L_i/2,minlen); if C_i < ntagcut, the length for bin i+1 will be max(2*L_i, maxlen). Note, by default, the bin sizes decrease/increase by a factor of 2. Thus, the user should let maxlen = (2^n)*minlen.

Value

A data frame with rows corresponding to the bins and columns corresponding to the following:

chr	Chromosome IDs.
gstart	The start position of the bin.
gend	The end position of the bin.
ct12	For one-sample analysis, where only the ChIP data are available, ct12 = ipct1 + ipct2. For two-sample analysis, where both the ChIP and control data are available. ct12 = maximum(ipct1+ipct2-conct1-conct2,0).
ipct1	The number of sequence tags for the chain 1 of the ChIP data.
ipct2	The number of sequence tags for the chain 2 of the ChIP data.
conct1	The number of sequence tags for the chain 1 of the control data.
conct2	The number of sequence tags for the chain 2 of the control data.

Author(s)

Qianxing Mo <moq@mskcc.org>

References

Qianxing Mo (2011). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, Advance Access published September 13, 2011. doi:10.1093/biostatistics/kxr029

See Also

iSeq1, iSeq2, peakreg,plotreg

Examples

```
data(nrsf)
chip = rbind(nrsf$chipFC1592,nrsf$chipFC1862,nrsf$chipFC2002)
mock = rbind(nrsf$mockFC1592,nrsf$mockFC1862,nrsf$mockFC2002)

tagct = mergetag(chip=chip,control=mock,maxlen=80,minlen=10,ntagcut=10)
```

---

nrsf *nrsf data*

Description

This is a subset of the neuron-restrictive silencer factor (NRSF) data containing the information of the sequence tags that are uniquely mapped (up to two mismatches allowed) to chromosomes 22 and Y of human genome.

Usage

```
data(nrsf)
```

Source

Science 316, 1497-1502.

References

David S. Johnson, Ali Mortazavi, Richard M. Myers, Barbara Wold. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497-1502.

---

peakreg *Call and merge enriched genomic windows/bins.*

Description

A function used to call and merge enriched bins using the posterior probability calculated by iSeq1 or iSeq2 functions at certain posterior probability and false discovery rate (FDR) cutoffs.

Usage

```
peakreg(chrpos,count,pp,cutoff,method=c("ppcut","fdrcut"),maxgap=300)
```

Arguments

chrpos	A n by 3 matrix or data frame. The rows correspond to genomic bins. The first column contains chromosome IDs; the second and third columns contain the start and end positions of the bin, respectively.
count	A n by 2 matrix containing the number of sequence tags in the bins specified by chrpos. The first column contains the tag counts for chain 1 (usually the forward chain), and the second column contains the tag counts for chain 2 (usually the reverse chain). See the document of the function 'mergetag' for the definition of chain 1 and 2. The function uses the information in 'count' to find the center of the enriched regions, where the true binding sites are usually located.
pp	A vector containing the posterior probabilities of bins in the enriched state returned by functions iSeq1 or iSeq2.
cutoff	The cutoff value (a scalar) used to call enriched bins. If use posterior probability as a criterion (method="ppcut"), a bin is said to be enriched if its pp is greater than the cutoff. If use FDR as a criterion (method="fdrcut"), bins are said to be enriched if the bin-based FDR is less than the cutoff. The FDR is calculated using a direct posterior probability approach (Newton et al., 2004).
method	'ppcut' or 'fdrcut'.
maxgap	The criterion used to merge enriched bins. If the genomic distance of adjacent bins is less than maxgap, the bins will be merged into the same enriched region.

Value

A data frame with rows corresponding to enriched regions and columns corresponding to the following:

chr	Chromosome IDs.
gstart	The start genomic position of the enriched region.
gend	The end genomic position of the enriched region.
rstart	The row number for gstart in chrpos.
rend	The row number for gend in chrpos.
peakpos	The inferred center (peak) of the enriched region.
meanpp	The mean posterior probability of the merged regions/bins.
ct1	total tag counts for the region from gstart to gend for the chain corresponding to count[,1]; ct1=sum(count[rstart:rend,1])
ct2	total tag counts for the region from gstart to gend for the chain corresponding to count[,1]; ct2=sum(count[rstart:rend,2])
ct12	ct12 = ct1 + ct2
sym	A parameter used to measure if the forward and reverse tag counts are symmetrical (or balanced) in enriched regions. The values range from 0.5 (perfect symmetry) to 0 (complete asymmetry).

Author(s)

Qianxing Mo <moq@mskcc.org>

References

Qianxing Mo (2011). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, Advance Access published September 13, 2011. doi:10.1093/biostatistics/kxr029

Newton, M., Noueiry, A., Sarkar, D., Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5 , 155-176.

See Also

iSeq1, iSeq2, mergetag,plotreg

Examples

```
data(nrsf)
chip = rbind(nrsf$chipFC1592,nrsf$chipFC1862,nrsf$chipFC2002)
mock = rbind(nrsf$mockFC1592,nrsf$mockFC1862,nrsf$mockFC2002)
tagct = mergetag(chip=chip,control=mock,maxlen=80,minlen=10,ntagcut=20)
tagct22 = tagct[tagct[,1]=="chr22",]
res1 = iSeq1(Y=tagct22[,1:4],gap=200,burnin=200,sampling=500,ctcut=0.95,a0=1,b0=1,
 a1=5,b1=1, k0=3,mink=0,maxk=10,normsd=0.1,verbose=FALSE)

reg1 = peakreg(tagct22[,1:3],tagct22[,5:6]-tagct22[,7:8],res1$pp,0.5,
        method="ppcut",maxgap=200)

reg2 = peakreg(tagct22[,1:3],tagct22[,5:6]-tagct22[,7:8],res1$pp,0.05,
         method="fdrcut",maxgap=200)
```

`plotreg` *A function used to plot enriched genomic regions*

Description

A function used to plot enriched genomic regions.

Usage

```
plotreg(gpos,ipct,conct,peak,col=c("yellow","green","grey0","blue"))
```

Arguments

`gpos`	A n by 2 matrix or data frame. The rows correspond to genomic bins. The first and second columns contain the start and end positions of the genomic windows/bins, respectively.
`ipct`	A n by 2 matrix containing the ChIP tag counts corresponding to the bins in gpos. ipct[,1] contains the counts for the chain 1 (usually the forward chain); ipct[,2] contains the counts for the chain 2 (usually the reverse chain).
`conct`	A n by 2 matrix containing the control tag counts corresponding to the bins in gpos. ipct[,1] contains the counts for the chain 1 (usually the forward chain); ipct[,2] contains the counts for the chain 2 (usually the reverse chain).
`peak`	A vector containing the peak (center) positions of the genomic regions.
`col`	The colors used to fill the rectangles. col[1] is used for ipct[,1], col[2] for ipct[,2], col[3] for conct[,1] and col[4] for conct[,2], respectively.

Value

No value returned.

Author(s)

Qianxing Mo <moq@mskcc.org>

References

Qianxing Mo (2011). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, Advance Access published September 13, 2011. doi:10.1093/biostatistics/kxr029

See Also

iSeq1, iSeq2, peakreg,mergetag

Examples

```
#see the example in iSeq1
```

Index