Block Gibbs Sampler for RNA Secondary Structure Prediction

Donglai Wei Charles Lawrence

2010.9.16

0.Welcome

1.Preparation:Packages to Install

1.1 Alignment Initialization: Probcons

http://probcons.stanford.edu/download.html

Command Line: *./probcons input_file >output_file*
Input Format: FASTA
Output Format: CLUSTALW/plain-text

1.2 Structure Sampler given Alignment: RNAalifold(Vienna Package)

http://www.tbi.univie.ac.at/RNA/

Command Line: *RNAalifold input_file >output_file*
Input Format: CLUSTALW
Output Format: Vienna

1.3 Alignment Sampler given Structure: Covariance Model(Inferno Package)

http://infernal.janelia.org/

a) Build CM model:
Command Line: *cmbuild -F output_file input_file*
Input Format: Stockholm
Output Format: CM

b) Sample the multiple alignment:
Command Line: *cmalign −−sample input_file1 input_file2 >output_file*
Input Format: 1:CM 2:FASTA
Output Format: Vienna

2.Block Gibbs Sampler:Python Code

2.1Main Function: B_GS_main.py

B_GS(seq_file,R_Dir,C_Dir,iteration,P_Dir)
Input:

1. seq_file: path of the sequence file
2. R_Dir: path of the RNAalifold function
3. C_Dir: path of the Infernal/src
4. iteration:number of iterations
5. P_Dir:path of probcons.exe

Output:

1. 00aln:#iteration+1 samples of the multiple alignment
2. 00str:#iteration samples of the consensus structure
3. project_i.str:project the consensus structures in 00str onto the ith sequence

2.2Util Function B_GS_util.py

Essentials:

2.2.1 Format Parser during Iteration

1. Ini_aln: Alignment Initialization with PROBCONS
2. Aln_aln: Translate the plain-text format of alignment from PROBCONS into ClustalW format
3. $Alifold\_sto$: Translate the plain-text format of structure from RNAalifold into Stockholm format
4. CM_aln: Translate the structural alignment from cmalign into ClustalW format

2.2.2 Data Analysis

1. $project\_strus$:project the consensus structures in 00str onto one certain sequence
2. sta: calculate sensitivity and PPV for the estimator
3. cal_roc: calculate the points on the ROC curve

2.3 Cluster Analysis $hier\_clus.m$

1. Bias-Variance: based on the hamming distance between sampled strutures and the reference one
2. Hierarchical Clustering:using CH-index to determine number of clusters
3. γ-Centroid Estimator: implement Nussinov-type DP algorithm

2.3 Paralellization

1. $para.py$: paralell computing on ccmb

3.Test Cases

3.1: Man-or-Boy test

In the test folder are:
1) two homologous squences from tRNA in fasta format
2) their reference structure from Rfam
3) the python script.

Firsly, You need to specify the path for other packages in the script.py.

Then cd into the test folder and type *python script.py* in the terminal.

3.2: Comprehensive test on Kiryu's dataset

In order to reproduce the result shown in the paper, we here provide the wrapper for doing RNAG on 85 subalignments of 10 homologous sequences from 17 RNA families. Everything is included in the "Initial.py"

4.Acknowledgement

Thanks Chip, who opened me the door to the amazing world of computational statistics, Bill who has always been helpful ever since I was a green hand on matlab and all my labmates who have shared pleasant years of lab meetings together.
Also, as usual, special thanks go to my forever loving family, especially Zoe:)
In the end, let me know if there is anything I can help(*donglai_wei@brown.edu*) and hope you are not the last one to read it :p