BIOS 731
Advanced Statistical Computing
Fall 2022

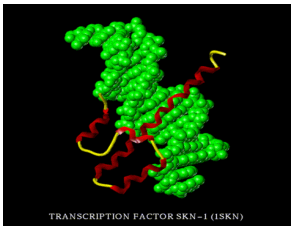
Lecture 13
Applications of MCMC and SMC

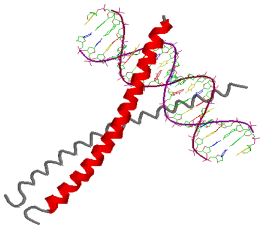
Steve Qin

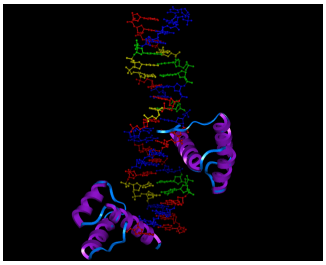
Review

- Gibbs sampler
- Grouping and collapsing
- Convergence check
- Sequential Monte Carlo
  - Acceptance rejection method
  - Importance sampling

Appliation: Transcription Factor Binding Sites Discovery







3

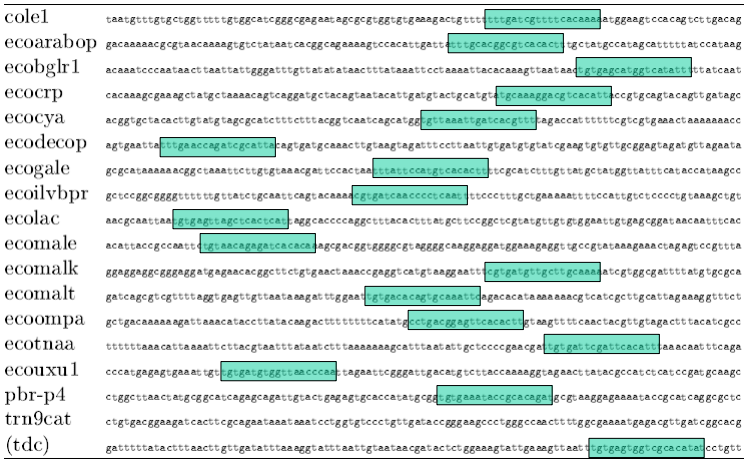
Example: cyclic receptor protein (CRP)

cole1	taatgtttgtgctggttttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
ecobglr1	acaaatcccaataacttaattattgggatttgttatatataactttataaatttcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgtttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
ecodecop	agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale	gcgcataaaaaacggctaaaattcttgtgtaaacgattccactaatttattccatgtcacactttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggttttttgttatctgcaattcagtacaaaacgtgatcaacccctcaattttccctttgctgaaaaattttccattgtctccccgtaaagctgt
ecolac	aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale	acattaccgccaattctgtaacagagatcacacaaagcgacggtgggggcgtagggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttta
ecomalk	ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagattggaattgtgacacagtgcaaattcagacacatataaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacatacctttatacaagactttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagacttttacatcgcc
ecotnaa	ttttttaaacatttaaaattctttacgtaatttataatctttaaaaaaagcatttaatattgctcccgaacgattgtgattcgattcacatttaaacaatttcaga
ecouxu1	cccatgagagtgaaattgttgtgatgtggttcaccccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatcccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
trn9cat	ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
(tdc)	gatttttatacttttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt

4

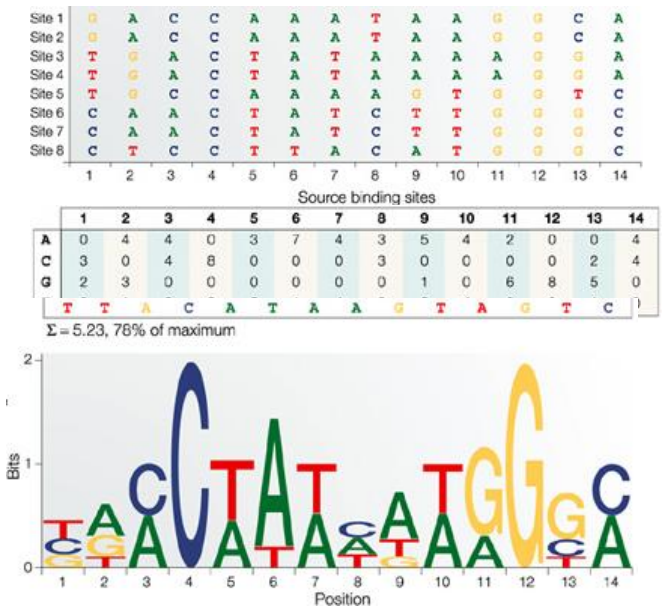
Stormo and Hartzell, 1989

Example: cyclic receptor protein (CRP)



Stormo and Hartzell, 1989

Transcription factor binding site (TFBS)



Motif identification model

a_1
aaaggtcgagtagctactcgatcgatactagcaatcgttaccctagctcgatcgaaa
a_2
acgtgagatcagctatgaccgatagctactcgataaccg
a_3
gaatagctactcgatcgatactagcaatcgttaccctagctcgatcgagatggaaag
• • •
a_J
acgtgagatcagctatcgatcgattgataactactcgtacgtat

Alignment variable $A = \{a_1, a_2 ..., a_J\}$

7

Posterior distributions

- The posterior conditional distribution for alignment variable $\boldsymbol{A}$

$$p(a_j = l \mid \boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A}_{-j}) \propto \prod_{k=1}^{4} \theta_{0k}^{h_k(\boldsymbol{R_j})} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}} \right)^{h_k(r_{j,l+i-1})} \propto \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}} \right)^{h_k(r_{j,l+i-1})}$$

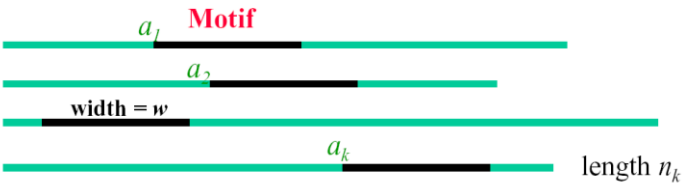
DNA sequence data

$$\boldsymbol{R} = (\boldsymbol{R_1}, ..., \boldsymbol{R_J})$$

Lawrence *et al. Science* 1993, Liu *et al. JASA* 1995

8

Motif Alignment Model



The missing data: Alignment variable: $A=\{a_1, a_2, \ldots, a_k\}$

- Every **non-site positions** follows a common multinomial with $\boldsymbol{p_0}=(p_{0,1},\ldots, p_{0,20})$
- Every position i in the motif element follows probability distribution $\boldsymbol{p_i}=(p_{i,1},\ldots, p_{i,20})$

9

Statistical Model

- Objects:
  - Seq: sequence data to search for motif
  - θ_0: non-motif (genome background) probability
  - θ: motif probability matrix parameter
  - π: site locations
- Problem: $P(\theta, \pi \mid seq, \theta_0)$
- Approach: alternately estimate
  - π by $P(\pi \mid \theta, seq, \theta_0)$
  - θ by $P(\theta \mid \pi, seq, \theta_0)$

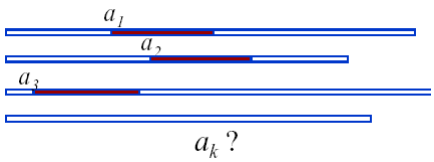
10

The Algorithm

- Initialize by choosing random starting positions
- Iterate the following steps many times;
  - Randomly or systematically choose a sequence to exclude
  - Carry out the predictive-updating step to update the starting position
  - Stop when no more observable changes in likelihood.

11

The Predictive Updating Step



- Compute predictive frequencies of each position *i* in motif

  c_{ij} = count of amino acid type ***j*** at position ***i.***

  c_{0j} = count of amino acid type j in all non-site positions.

  $q_{ij} = (c_{ij}+b_j)/(K\text{-}1+B)$, $B=b_1+ \bullet\bullet\bullet + b_K$ *"pseudo-counts"*

- Sample from the predictive distribution of a_k

$$P(a_k = l+1) \propto \prod_{i=1}^{w} \frac{q_{i,R_k(l+i)}}{q_{0,R_k(l+i)}}$$

12

References

- Lawrence et al. (1993) *Science.*
- Liu, Neuwald and Lawrence (1995) *JASA.*
- Liu and Lawrence (1999) *Bioinformatics.*

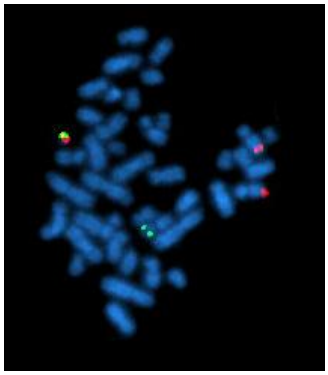
13

Infer the 3D shape of chromosomes

Slides from Ming Hu

Microscopic Methods

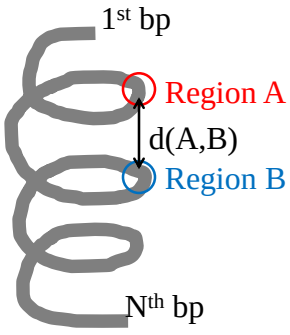
- Fluorescent *in situ* hybridization (FISH)



http://en.wikipedia.org/wiki/Cytogenetics

15

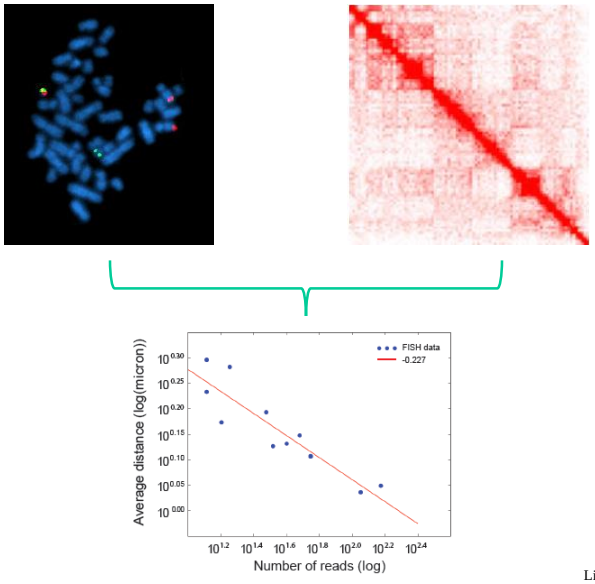
FISH Data Representation



3D chromosomal structure

16

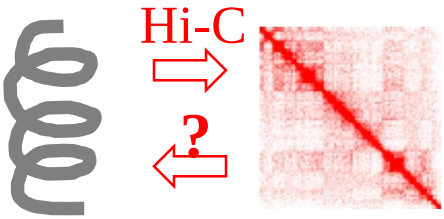
Contact Frequency vs. Spatial Distance



17

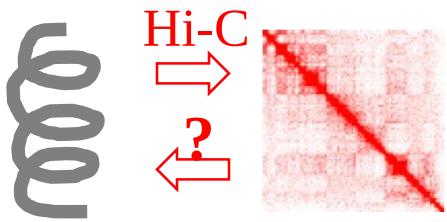
Lieberman-Aiden, et al, 2009

Problem setting



18

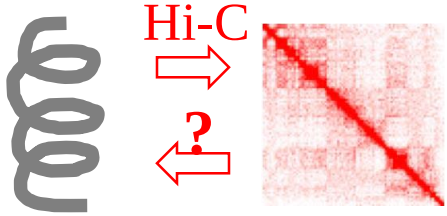
Problem setting



- Challenges:
  - Sequencing uncertainties
  - Biases: enzyme, GC content, mappability

19

Problem setting



- Challenges:
  - Sequencing uncertainties
  - Biases: enzyme, GC content, mappability

20

Yaffe and Tanay, 2011

Beads-on-a-string Representation

ACGTAGCTAGATACTGTAGTGTAGTTTGGAACCTGAGGG

21

Beads-on-a-string Representation

ACGTAGCTAGATACTGTAGTGTAGTTTGGAACCTGAGGG

22

Beads-on-a-string Representation

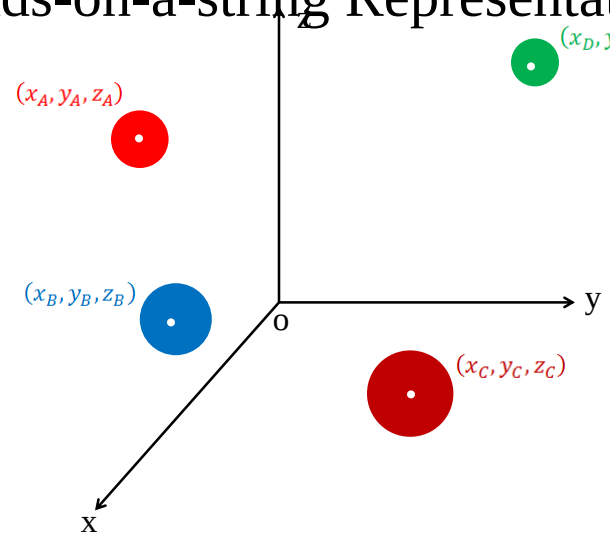
ACGTAGCTAG ATACTGTAGT GTAGTTTGGA ACCTGAGGG

23

Beads-on-a-string Representation

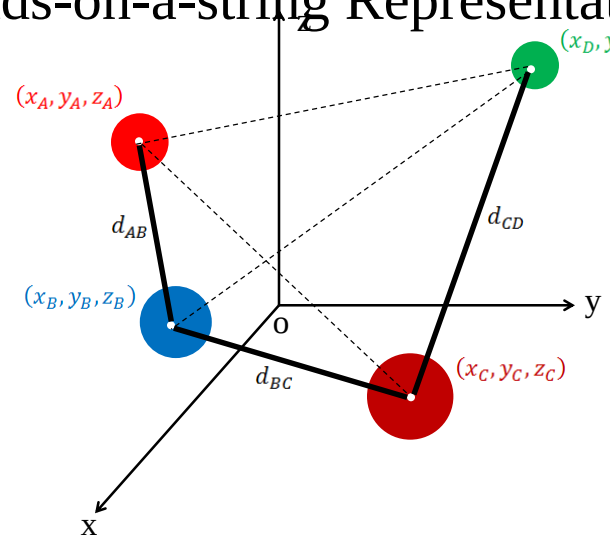
24

Beads-on-a-string Representation



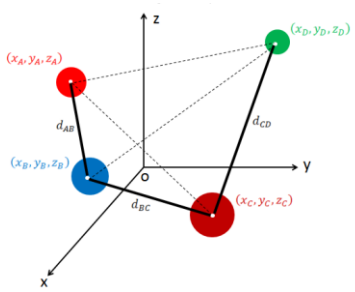
25

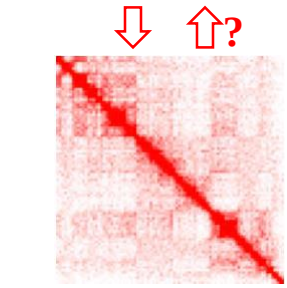
Beads-on-a-string Representation



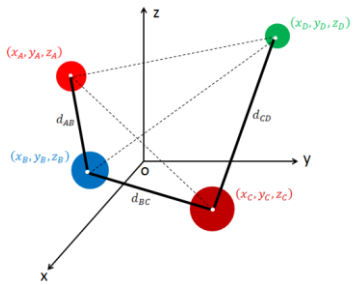
26

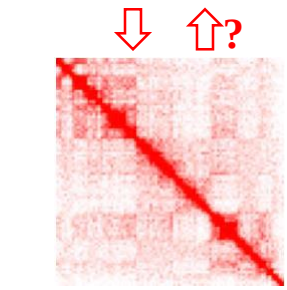
Bayesian Statistical Model





Bayesian Statistical Model



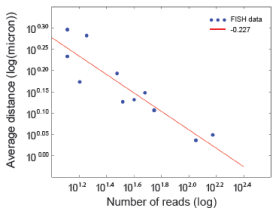


u_{ij} : # of reads between loci i and j
(x_i, y_i, z_i) : Euclidian coordinates of locus i
d_{ij} : spatial distance between loci i and j
e_i : # of enzyme cut site in locus i
g_i : GC content of locus i
m_i : mappability of locus i

Hi-C read counts: population summation

$$u_{ij} \sim Poisson(\theta_{ij})$$

Hi-C read counts vs. spatial distance: log-log linear



$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log(d_{ij}) + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

Lieberman-Aiden, et al, 2009

Bayesian Statistical Model

- Likelihood:

$$L(u_{ij}, 1 \le i < j \le N | x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) = \prod_{1 \le i < j \le N} \frac{e^{-\theta_{ij}} \theta_{ij}^{u_{ij}}}{u_{ij}!}$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\right) + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

29

Bayesian Statistical Model

- Likelihood: $\binom{N}{2}$ data points, $3N + 5$ parameters

$$L(u_{ij}, 1 \le i < j \le N | x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) = \prod_{1 \le i < j \le N} \frac{e^{-\theta_{ij}} \theta_{ij}^{u_{ij}}}{u_{ij}!}$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\right) + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

30

Bayesian Statistical Model

- Likelihood: $\binom{N}{2}$ data points, $3N + 5$ parameters

$$L(u_{ij}, 1 \le i < j \le N | x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) = \prod_{1 \le i < j \le N} \frac{e^{-\theta_{ij}} \theta_{ij}^{u_{ij}}}{u_{ij}!}$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\right) + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

- Posterior distribution

$$\pi(x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m | u_{ij}, 1 \le i < j \le N) \propto L(u_{ij}, 1 \le i < j \le N | x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) prior$$

31

Statistical Inference

- Algorithm: Bayesian 3D constructor for Hi-C data (BACH)

$$\pi(x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m | u_{ij}, 1 \le i < j \le N)$$

32

Statistical Inference

- Algorithm: Bayesian 3D constructor for Hi-C data (BACH)

$$\pi(x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m | u_{ij}, 1 \le i < j \le N)$$

➢ Initialization 1: use Poisson regression to obtain the initial values for $\beta_0, \beta_e, \beta_g, \beta_m$. Set $\beta_1 = -1$.

$$u_{ij} \sim Poisson(\theta_{ij}) \quad \log(\theta_{ij}) = \beta_0 + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

33

Statistical Inference

- Algorithm: Bayesian 3D constructor for Hi-C data (BACH)

$$\pi(x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m | u_{ij}, 1 \le i < j \le N)$$

➢ Initialization 1: use Poisson regression to obtain the initial values for $\beta_0, \beta_e, \beta_g, \beta_m$. Set $\beta_1 = -1$.

$$u_{ij} \sim Poisson(\theta_{ij}) \quad \log(\theta_{ij}) = \beta_0 + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

➢ Initialization 2: use sequential important sampling to get the initial 3D chromosomal structure $\{x_i, y_i, z_i, 1 \le i \le N\}$.

34

Statistical Inference

- Algorithm: Bayesian 3D constructor for Hi-C data (BACH)

$$\pi(x_i, y_i, z_i, 1 \le i \le N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m | u_{ij}, 1 \le i < j \le N)$$

- Initialization 1: use Poisson regression to obtain the initial values for $\beta_0, \beta_e, \beta_g, \beta_m$. Set $\beta_1 = -1$.

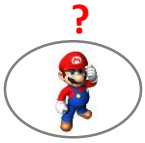
$$u_{ij} \sim Poisson(\theta_{ij}) \quad \log(\theta_{ij}) = \beta_0 + \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

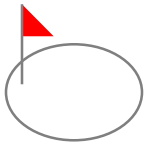
- Initialization 2: use sequential important sampling to get the initial 3D chromosomal structure $\{x_i, y_i, z_i, 1 \le i \le N\}$.

- Refinement: use Gibbs sampler with hybrid Monte Carlo to refine the initial values for parameters.

35

Sequential Importance Sampling

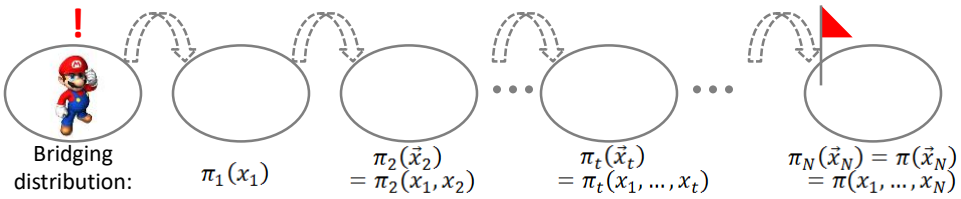




$$\pi(\vec{x}_N) = \pi(x_1, \ldots, x_N)$$

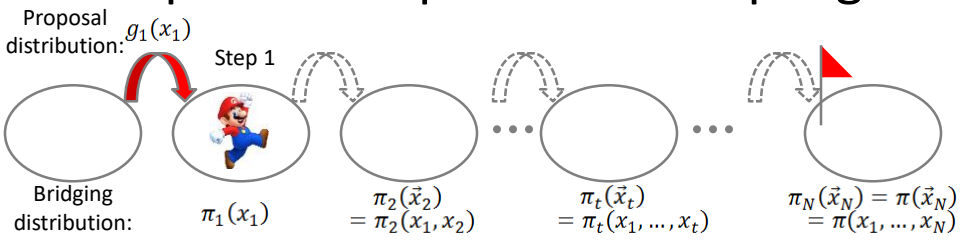
36

Sequential Importance Sampling



37

Sequential Importance Sampling



Weighted samples:

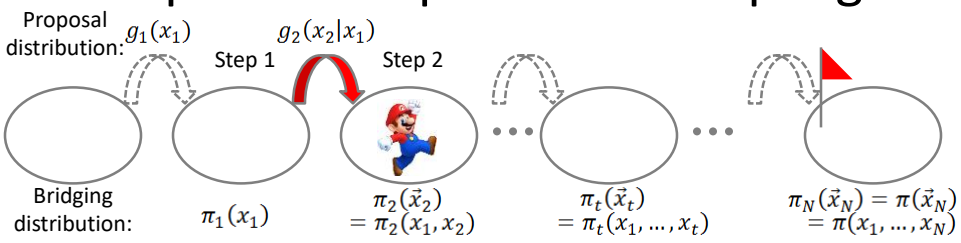
$x_1 \sim g_1(x_1)$

$w_1 = \frac{\pi_1(x_1)}{g_1(x_1)}$

(x_1, w_1)

38

Sequential Importance Sampling



Weighted samples:

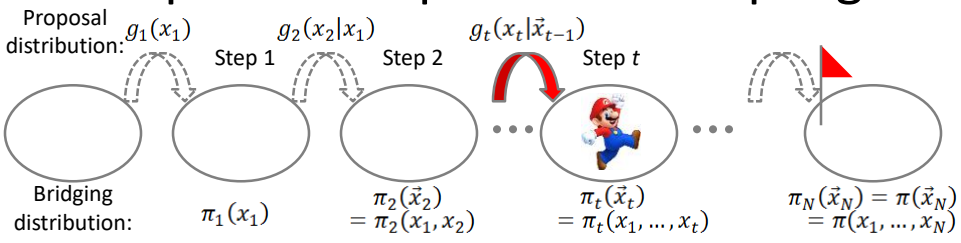
(x_1, w_1)

$x_2 \sim g_2(x_2|x_1)$

$$w_2 = \frac{w_1 \pi_2(\vec{x}_2)}{\pi_1(x_1) g_2(x_2|x_1)}$$

$(\vec{x}_2, w_2)$

Sequential Importance Sampling



Weighted samples:

(x_1, w_1)

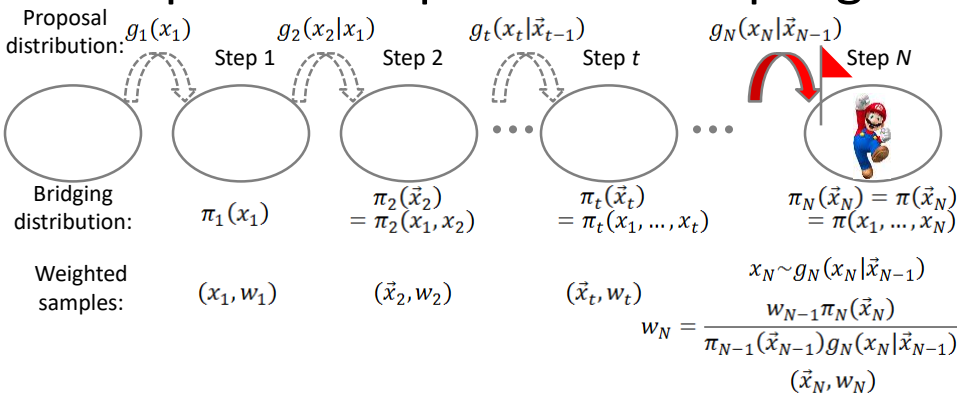
$(\vec{x}_2, w_2)$

$x_t \sim g_t(x_t|\vec{x}_{t-1})$

$$w_t = \frac{w_{t-1} \pi_t(\vec{x}_t)}{\pi_{t-1}(\vec{x}_{t-1}) g_t(x_t|\vec{x}_{t-1})}$$

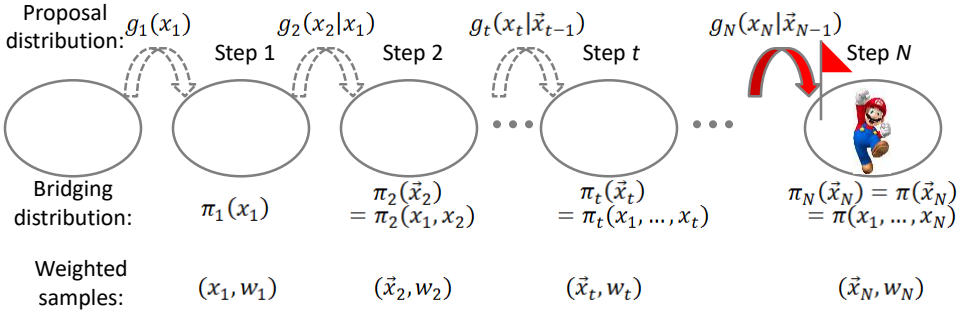
$(\vec{x}_t, w_t)$

Sequential Importance Sampling



41

Sequential Importance Sampling



Sequential Importance Sampling (SIS) Algorithm:

(1) Design bridging distributions $\pi_t(\vec{x}_t)$ and proposal distributions $g_t(x_t|\vec{x}_{t-1})$

(2) Sequentially draw weighted samples $x_t \sim g_t(x_t|\vec{x}_{t-1})$, and update weight

$$w_t = \frac{w_{t-1}\pi_t(\vec{x}_t)}{\pi_{t-1}(\vec{x}_{t-1})g_t(x_t|\vec{x}_{t-1})}$$

42

SIS in BACH: Outline

- Goal: use sequential importance sampling to sequentially put *N* loci into 3D space, i.e. sample from:

$$\pi(x_i, y_i, z_i, 1 \le i \le N | u_{ij}, 1 \le i < j \le N)$$

43

SIS in BACH: Outline

- Goal: use sequential importance sampling to sequentially put *N* loci into 3D space, i.e. sample from:

$$\pi(x_i, y_i, z_i, 1 \le i \le N | u_{ij}, 1 \le i < j \le N)$$

- Bridging distributions:

$$\pi_t(x_i, y_i, z_i, 1 \le i \le t | u_{ij}, 1 \le i < j \le t)$$

44

SIS in BACH: Outline

- Goal: use sequential importance sampling to sequentially put *N* loci into 3D space, i.e. sample from:

$$\pi(x_i, y_i, z_i, 1 \le i \le N | u_{ij}, 1 \le i < j \le N)$$

- Bridging distributions:

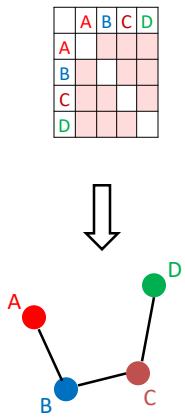
$$\pi_t(x_i, y_i, z_i, 1 \le i \le t | u_{ij}, 1 \le i < j \le t)$$

- Proposal distributions (given the first t-1 loci, put the t th locus in to 3D space):

$$g_t(x_t, y_t, z_t | x_i, y_i, z_i, 1 \le i \le t-1, u_{ij}, 1 \le i < j \le t)$$

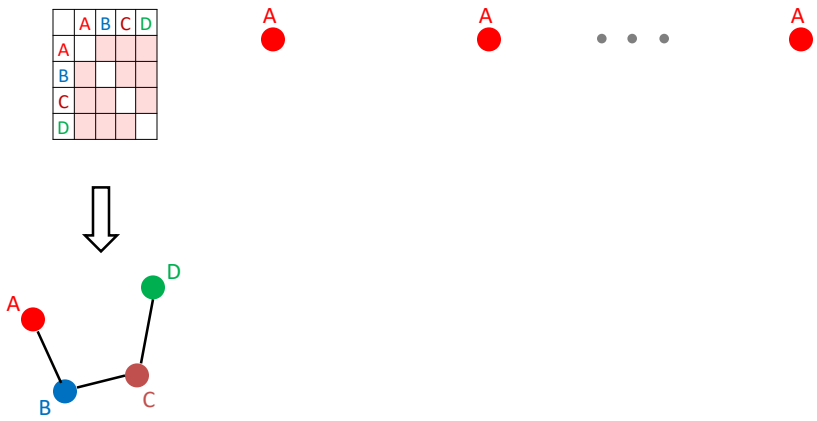
45

SIS in BACH: Illustration



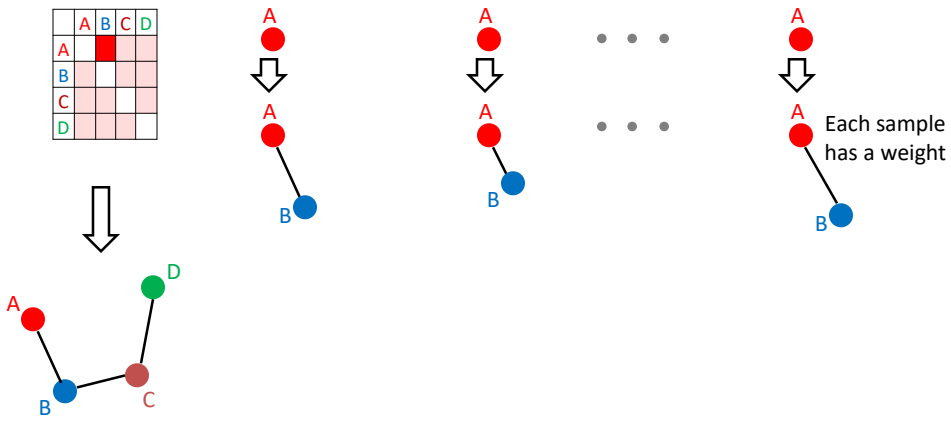
46

SIS in BACH: Illustration



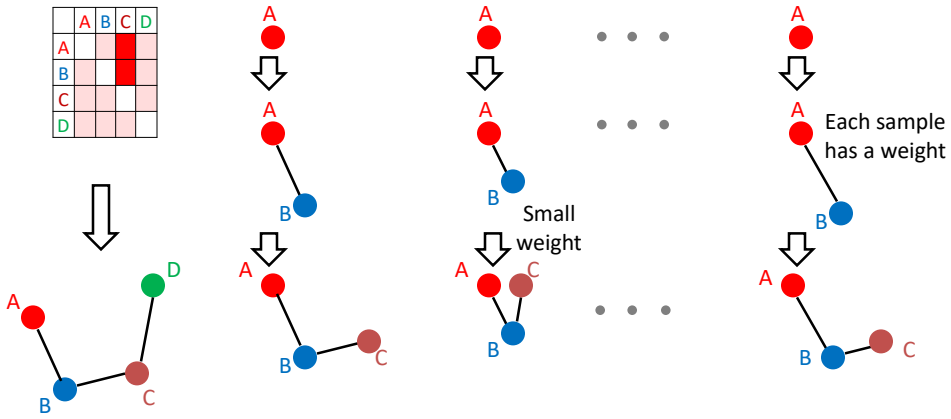
47

SIS in BACH: Illustration



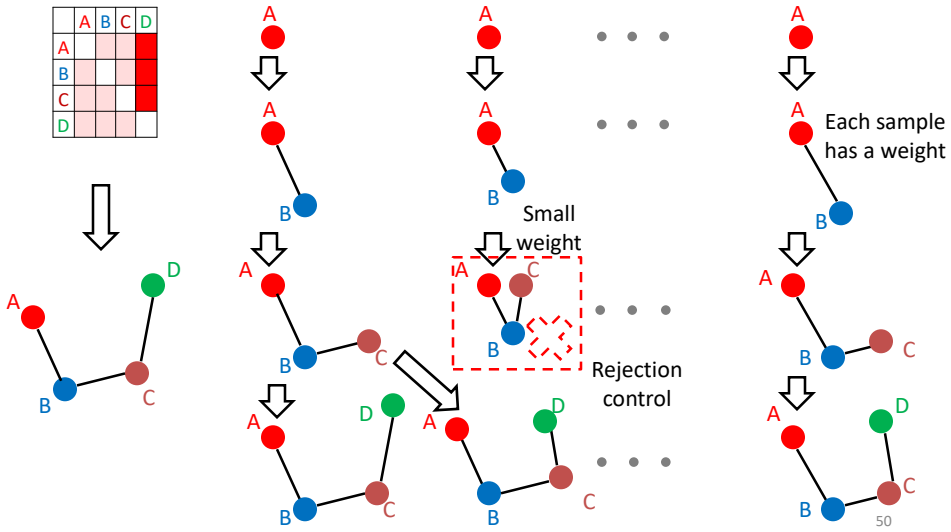
48

SIS in BACH: Illustration



49

SIS in BACH: Illustration



50

Hybrid Monte Carlo

- Goal: do efficient group move to refine initial 3D chromosomal structure, since local 3D coordinates are highly correlated.
- Combine molecular dynamics with Metropolis acceptance-rejection rule.

51
Duane, et al, 1987

Hybrid Monte Carlo in BACH

- Goal: sampling from

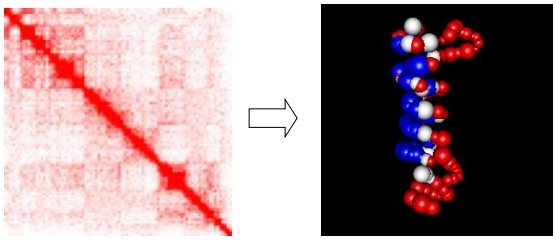
$$\pi(x_i, y_i, z_i, 1 \le i \le N | u_{ij}, 1 \le i < j \le N)$$

- Take partial derivate of log likelihood over 3D coordinates $(x_i, y_i, z_i, 1 \le i \le N)$.
- Run the leap-frog algorithm, adaptively tune the time interval to achieve acceptance rate ~ 90%.

52

Conclusions

- BACH: reconstruct chromosome 3D structures from Hi-C data
- Remove systematic biases
- Predicted spatial distances are consistent with FISH data
- Elongation of chromatin is highly associated with genetic/epigenetic features.
- Separation of compartments of A and B can be visualized.



53

References

- Hu M, Deng K, Qin ZS, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS. (2013) Bayesian inference of three-dimensional chromosomal organization. *PLoS Comput Biol.* **9** e1002893.

  http://www.people.fas.harvard.edu/~junliu/BACH/

54