

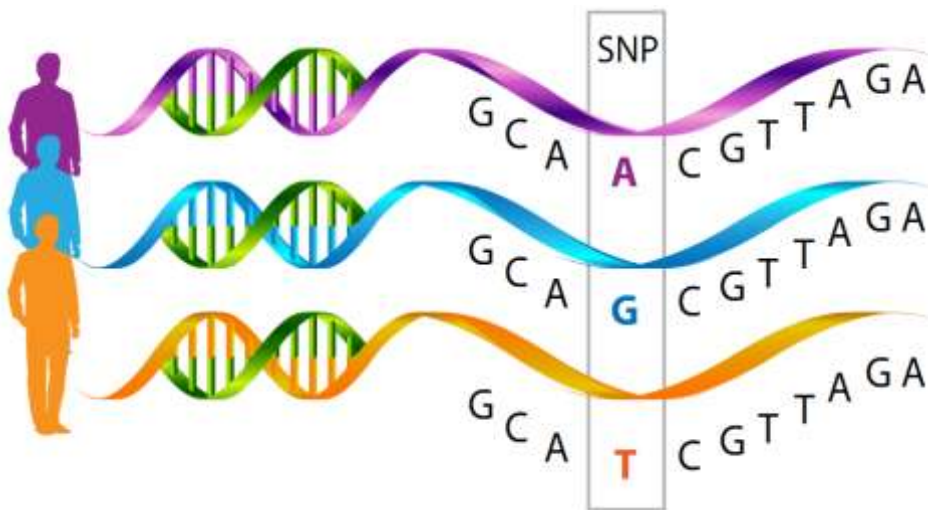
تعیین ساختار جمعیت در مسائل بیوانفرماتیک با روش‌های استنتاج بیزی



وحید حیدری
دانشگاه تهران
رشته‌ی الگوریتم‌ها و محاسبات

خلاصه واژه‌ها و مفاهیم

- دگره (Allele)
- چند ریختی تک نکلئوتیدی (SNP)
- ژن‌نمود (Genotype)
- رخ‌نمود (Phenotype)
- تعادل هاردی-وینبرگ (HWE)
- عدم تعادل پیوستگی (LD)

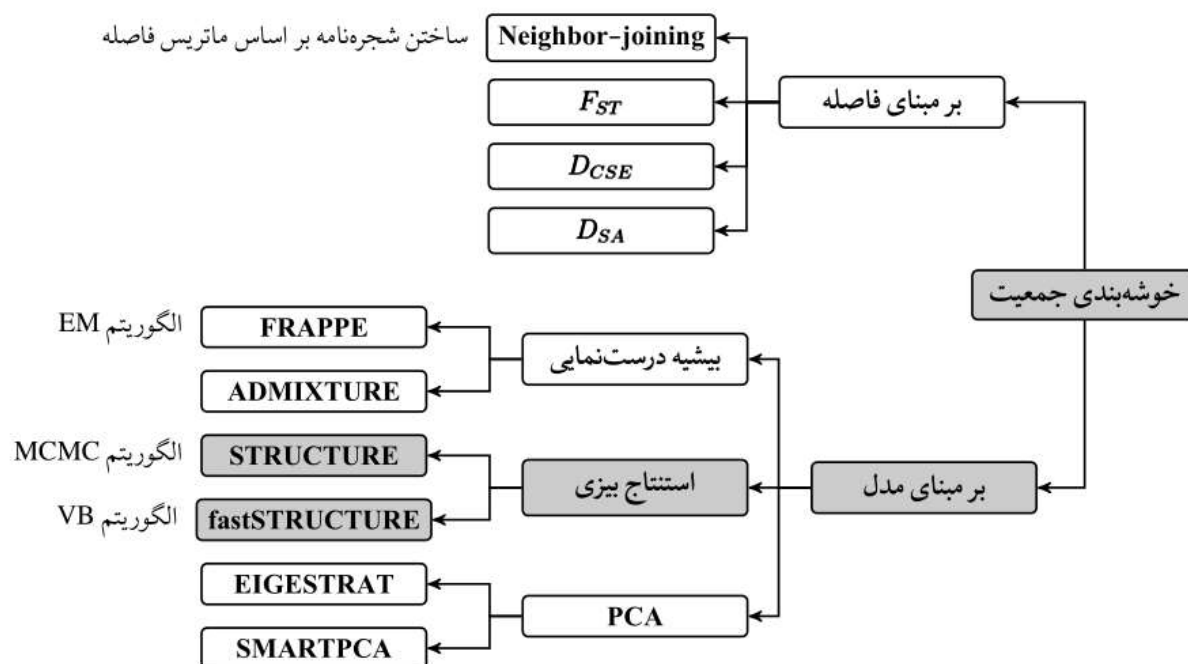




- افرادی با منشأ نامعلوم که در زیر جمعیت‌هایی خوشه بندی می‌شوند
 - تفاوت نژاد افراد جمعیت قابل تشخیص با ویژگی‌های ظاهری نیست
 - افراد هر زیر جمعیت دارای شباهت ژنتیکی درون جمعیتی هستند
 - افراد هر زیر جمعیت دارای تفاوت ژنتیکی بیرون جمعیتی هستند
- هر زیر جمعیت علل بیماری ژنتیکی خاصی دارد که با بقیه متفاوت است
 - عواملی مانند فاصله جغرافیایی باعث جدایی جمعیت‌ها می‌شود
 - هر زیر جمعیت مسیر تکاملی متفاوتی طی می‌کند
 - علائم بیماری یکسان ولی علل ژنتیکی متفاوت



- ساختار جمعیت
- مبتنی بر فاصله
- مبتنی بر مدل

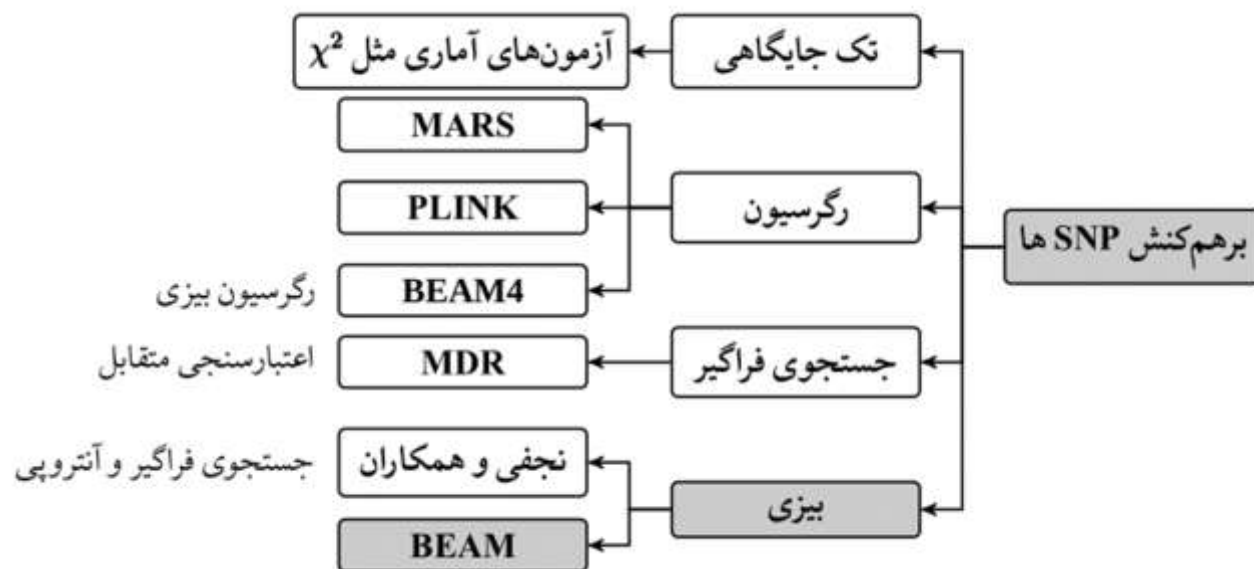




پیشینه تحقیق

• تشخیص برهم کنش

- تک جایگاهی
- رگرسیون
- جستجوی فراگیر
- بیزی



- مخفف Bayesian Epistasis Association Mapping

- BEAM1 نشانگر ۳ حالتی ساده
- BEAM2 بلوک بندی ژنوم
- BEAM3 گراف بیماری بجای بلوک بندی
- BEAM4 رگرسیون بیزی



مدل پیشنهادی

- ساختار جمعیت

- بیز گوناگونی




- تشخیص بیماری

- تشخیص روایستایی با BEAM3



مدل پیشنهادی (ورودی‌ها)

- ژن‌نمود دو دگرهای از L جایگاه برای N فرد
- Y برچسب «بیمار» یا «سالم»
- تعداد خوشه‌ها K از قبل می‌دانیم

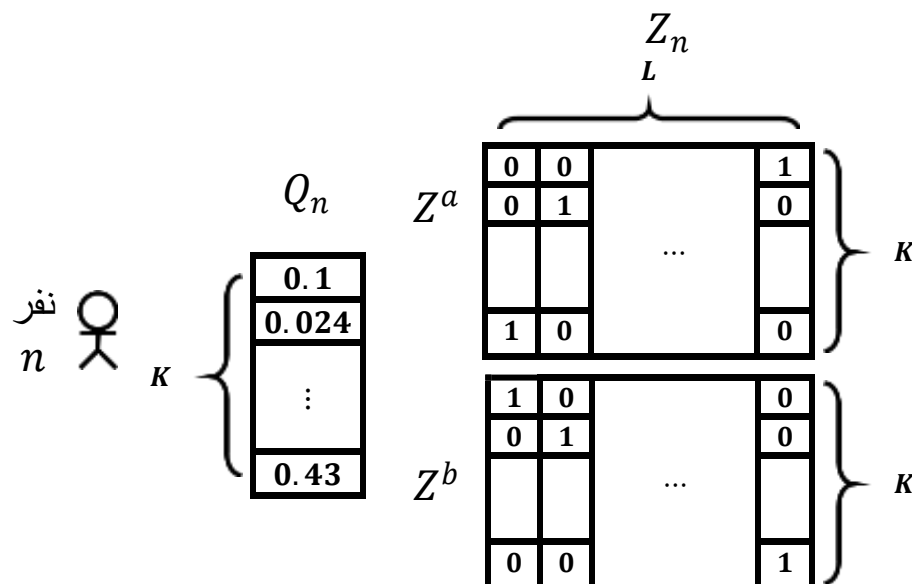
		Y	X			
			1	2	...	L
1		0	a b	a a	...	b b
2		1	b b	a b	...	b a
	\vdots	\vdots			\vdots	
N		0	a b	a b	...	a a



مدل پیشنهادی (پارامترها)

• پارامترهای هر فرد

- تعلق یک دگره به یک خوشه Z
- تعلق یک فرد به یک خوشه Q





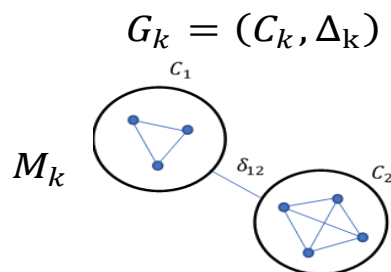
مدل پیشنهادی (پارامترها)

- پارامترهای هر خوشه
- بسامد (فراوانی) یک دگره در یک خوشه P
- مدل بیماری M
- جایگاه‌های بیماری I
- گراف بیماری $G = (C, \Delta)$

$$P_k$$

L			
p_a	p_a	...	p_a
p_b	p_b		p_b

زیر جمعیت
 k



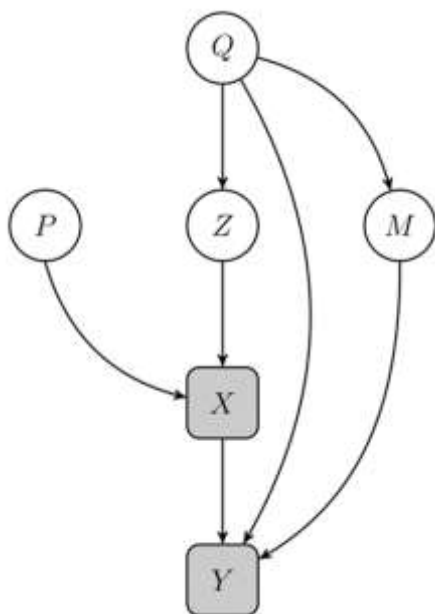
$$I_k$$

L			
0	1		0



مدل پیشنهادی (مدل گرافیکی)

- احتمال توأم پارامترها

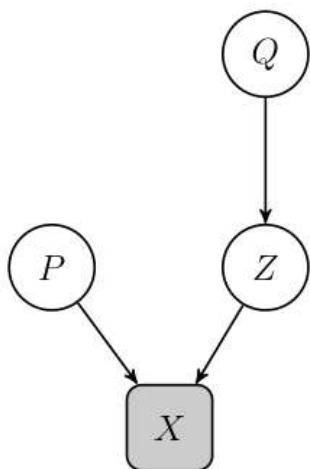


$$\begin{aligned} f(X, Y, P, Z, Q, M) = & f(X|P, Z) f(Z|Q) \\ & f(Y|X, Q, M) f(M|Q) \\ & \pi(P) \pi(Q) \end{aligned}$$



مدل پیشنهادی (ساختار جمعیت)

- زیر جمعیت‌ها در تعادل هاردی-وینبرگ هستند
- بسامد دگره‌های هر زیر جمعیت مستقل از بقیه است



$$f(X, Z, Q, P) = f(X|Z, P)f(Z|Q)\pi(P)\pi(Q)$$



مدل پیشنهادی

(ساختار جمعیت / بیز گوناگونی)

- تقریب میدان میانگین: با توزیع‌های \tilde{f} توزیع پسین f را تقریب می‌زنیم
- قضیه بیز گوناگونی $\ln \tilde{f}(\theta_i) \propto \mathbb{E}_{\tilde{f}(\theta_{-i})} [\ln f(\theta, X)]$
- کران پایین شواهد $\ln f(X) \geq \mathbb{E}_{\tilde{f}(\theta)} \left[\ln \frac{f(\theta | X)}{\tilde{f}(\theta)} \right]$

- مراحل استفاده از VB با LLBO
 - لگاریتم توزیع توأم پارامترها را بدست می‌آوریم
 - امید ریاضی را نسبت به جمع پخش می‌کنیم
 - مقادیر امیدهای ریاضی و گشتاورها را جایگزاری می‌کنیم
 - بهینه‌سازی پارامترهای توزیع‌های گوناگونی
 - به‌روزرسانی پارامترهای توزیع‌های گوناگونی در یک چرخه



مدل پیشنهادی (تشخیص بیماری)

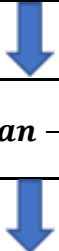
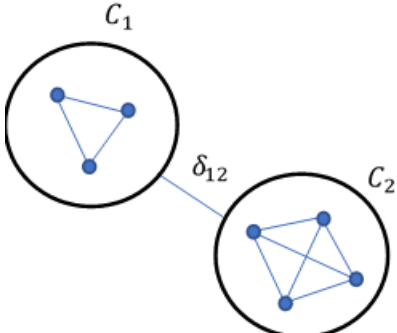
- همبستگی بین Y و X_1 وجود دارد
- ژن نمود افراد سالم به Y بستگی ندارند
- توزیع دگرها در افراد سالم و بیمار

$$f(\mathbf{X}, \mathbf{Y}, \mathbf{G}, \mathbf{I}) = f(\mathbf{X}|\mathbf{Y}, \mathbf{G}, \mathbf{I}) f(\mathbf{G}|\mathbf{I}) \pi(\mathbf{I}) \pi(\mathbf{Y})$$

$$\propto \frac{f_A(X_1|Y, G)}{f_0(X_1)} f(G|I) \pi(I)$$



مدل پیشنهادی (مقدار اولیه‌ی مدل بیماری)

1	$I_k \quad \overbrace{\begin{bmatrix} 0 & 1 & \dots & 1 & 0 \end{bmatrix}}^L$	$\pi(I_k)$
2	 <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 0 auto;">Pitman - Yor</div>	<div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> $new: P(x_i \in C_{W_{i+1}}) = \frac{n_k - \beta}{i + \alpha}$ </div> <div style="border: 1px solid black; padding: 5px;"> $old: P(x_i \in C_i) = \frac{\alpha + W_i \beta}{i + \alpha}$ </div>
3		$\pi(\delta_{ij})$



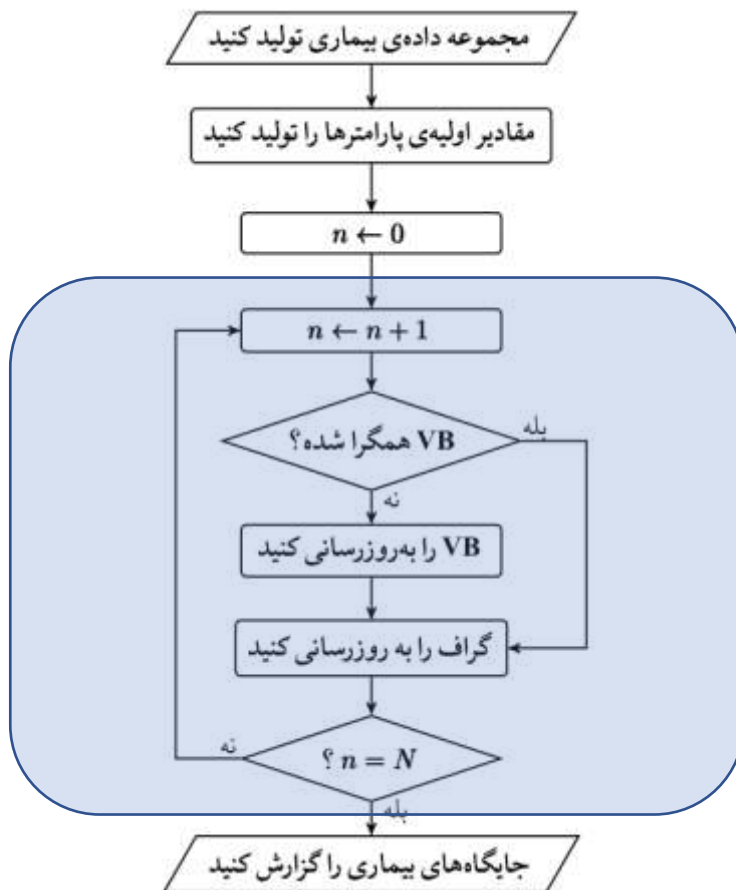
مدل پیشنهادی

(به روزرسانی گراف)

1	<p style="text-align: center;">x_i</p>	$\pi(I_k)$
2		$\gamma = \frac{f_A(x_i + X_1, G, Y, I)}{f_0(x_i + X_1)}$
3		$\frac{\gamma}{1 + \gamma}$



الگوریتم پیشنهادی





مدل پیشنهادی (تحلیل)

- تقریب میدان میانگین برای ساختار جمعیت

- تمام پارامترها را از هم مستقل در نظر گرفتیم

- پیچیدگی زمانی $O(N^2 L K)$

- تقریب گراف برای $f_0(x_i + X_1)$

- تعداد گراف‌هایی که باید برای f_0 محاسبه کرد: W^{W-2}

- اگر از گراف فعلی استفاده کنیم: $W(W - 1)$

- در واقع سعی می‌کنیم گراف فعلی را در هر چرخه بهبود بدهیم

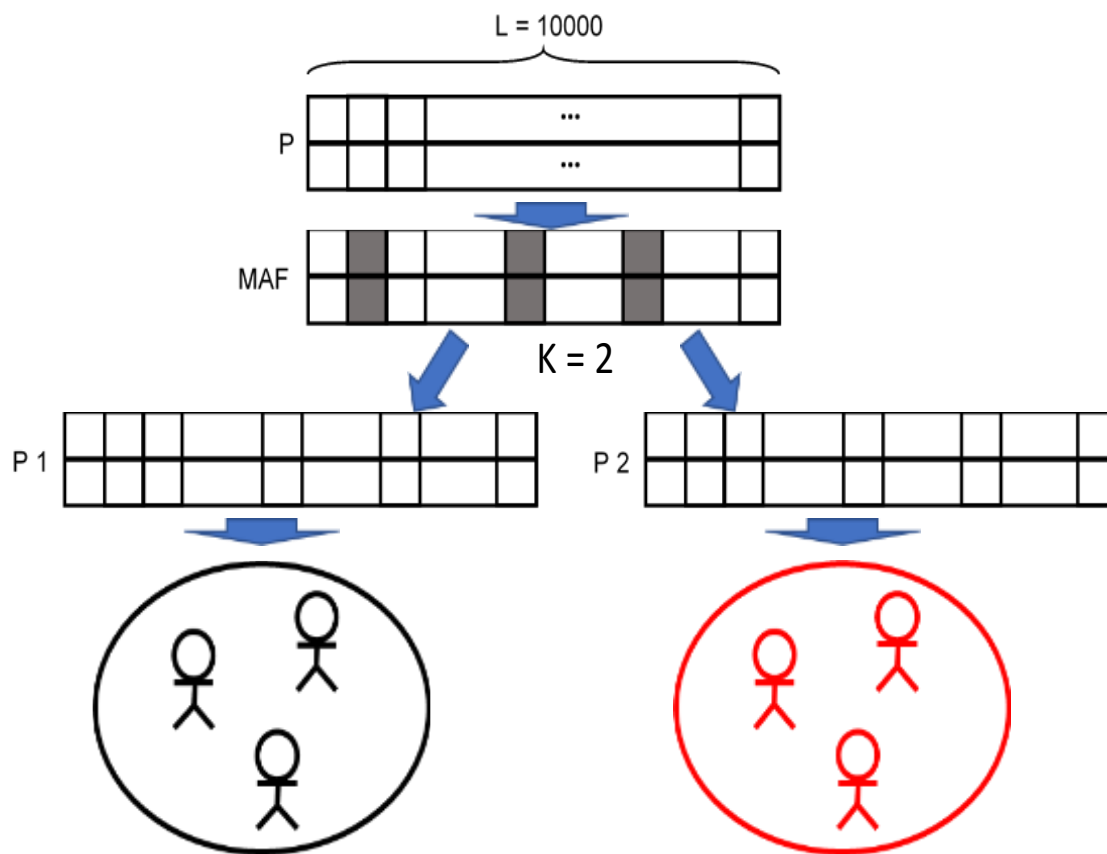


تولید داده‌های ساختگی

- شبیه‌سازی
 - جمعیت غیر مخلوطی
 - جمعیت مخلوطی
 - مدل بیماری



تولید داده‌های جمعیت





تولید داده‌ی بیماری

(مشخص کردن مدل بیماری)

$$\begin{aligned} \text{logit}(Y) = & \frac{1}{2} X_1 + \frac{1}{2} X_2 + \frac{1}{2} X_3 + \\ & + X_4 X_5 + X_6 X_7 + \\ & + 1.5 X_8 X_9 X_{10} \\ & + \text{const} \end{aligned}$$



تولید داده‌ی بیماری

(تعیین جایگاه‌ها و SNPها)

جمعیت ۲

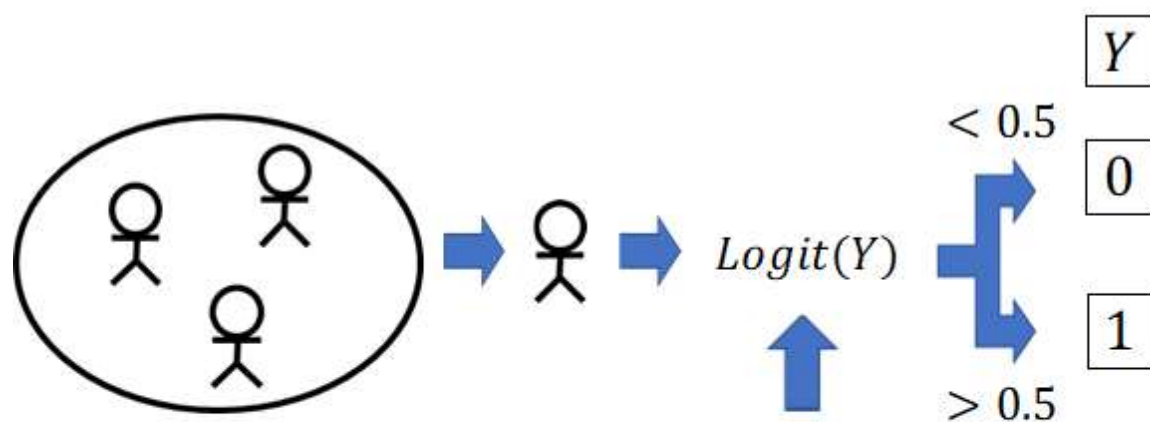
جمعیت ۱

SNP			جایگاه		
۱			۷۸		
۱			۴۲۱۶		
۱			۷۳۸۹		
۱	۰		۵۴۳۵	۹۴۹۱	
۱	۲		۲۹۶۹	۸۶۷۹	
۱	۲	۱	۵۷۸	۳۷۸۶	۹۹۷۰

SNP			جایگاه		
۰			۲۲۰		
۱			۲۳۹۲		
۲			۵۵۰۵		
۰	۲		۶۱۸۳	۸۱۳۹	
۲	۰		۵۳۹	۳۲۴۲	
۰	۰	۱	۱۷۲۴	۲۶۲۰	۶۰۰۵



تولید داده‌ی بیماری (افزودن برچسب)



SNP		جایگاه	
۰		۲۲۰	
۱		۲۳۹۲	
۳		۵۵۰۵	
۰	۲	۶۱۸۳	۸۱۳۹
۲	۰	۵۳۹	۳۲۴۲
۱	۱	۱۷۲۲	۲۲۲۰



- بررسی دقت خوشه‌بندی

- بیز گوناگونی
- زنجیر مارکوف مونت کارلو

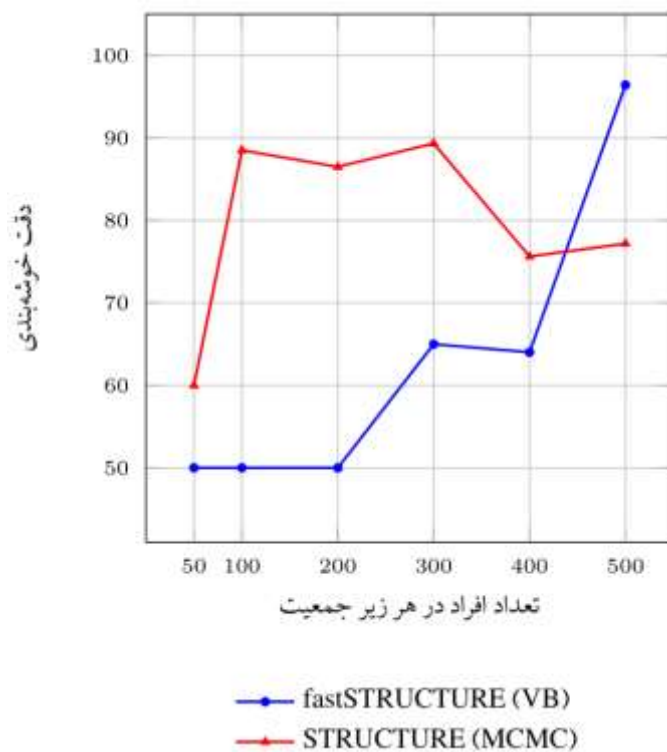
- تشخیص بیماری

- مجموعه داده‌های دارای تمام جایگاه‌های بیماری
- مجموعه داده با برهم‌کنش سه‌گانه

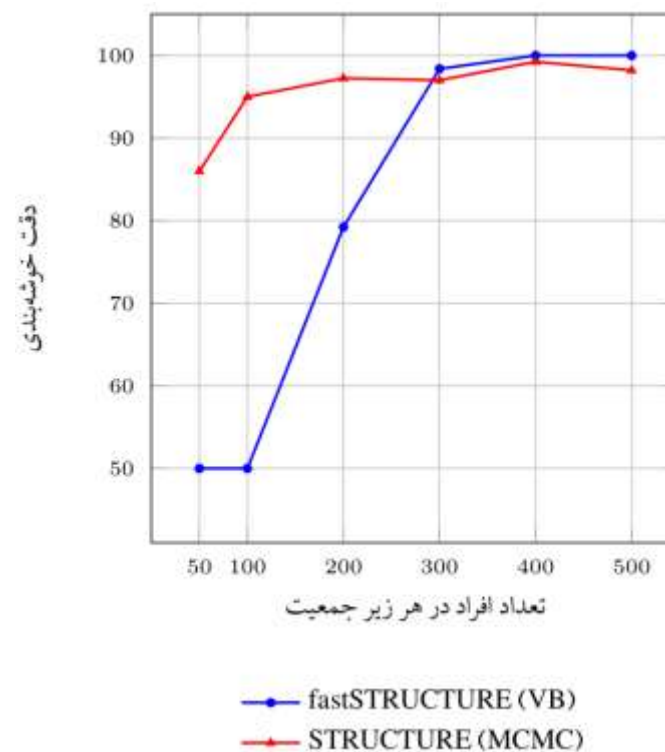


دقت خوشه‌بندی K=2

MAF=2%



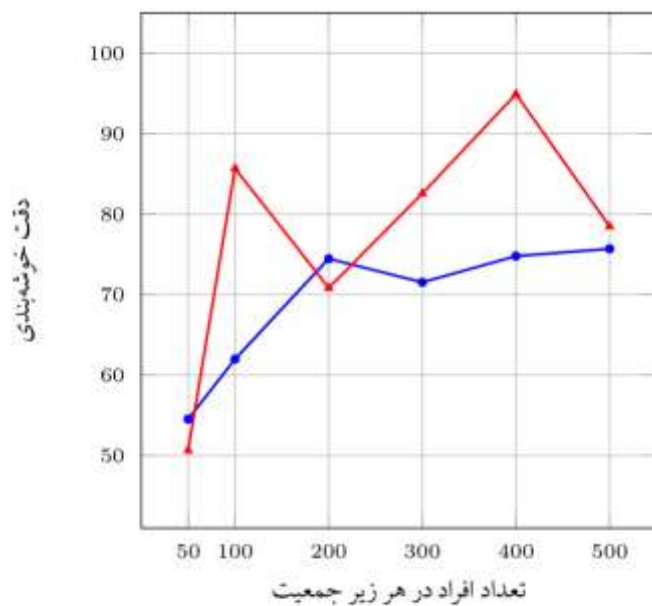
MAF=1%





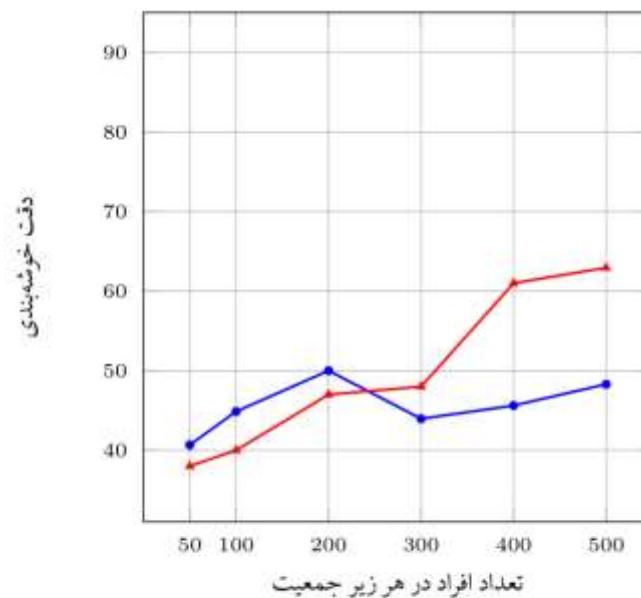
دقت خوشه‌بندی K=3

MAF=2%



—●— fastSTRUCTURE (VB)
—●— STRUCTURE (MCMC)

MAF=1%



—●— fastSTRUCTURE (VB)
—●— STRUCTURE (MCMC)



مقایسه روش‌های خوشه‌بندی

زنجیر مارکوف مونت کارلو

- نمونه‌برداری از توزیع پسین
- زمان زیاد برای نمونه‌برداری
- عملکرد بهتر در حجم کم نمونه
- نمونه‌برداری بیشتر برای نمونه‌های بزرگ برای بهبود دقت

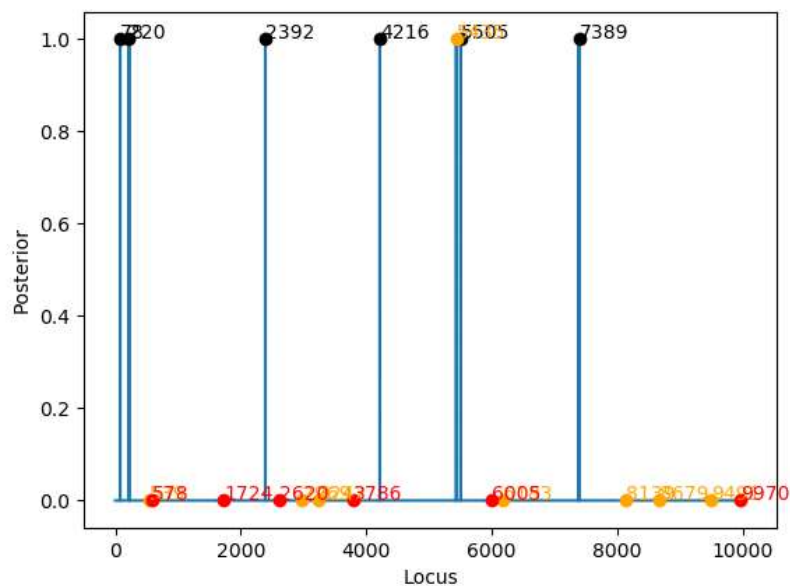
بیز گوناگونی

- بهینه‌سازی
- سرعت بیشتر در همگرایی
- نیاز به حجم نمونه‌ای به نسبت بیشتر
- کوچک کردن مقدار آستانه برای نمونه‌های بزرگ برای بهبود دقت

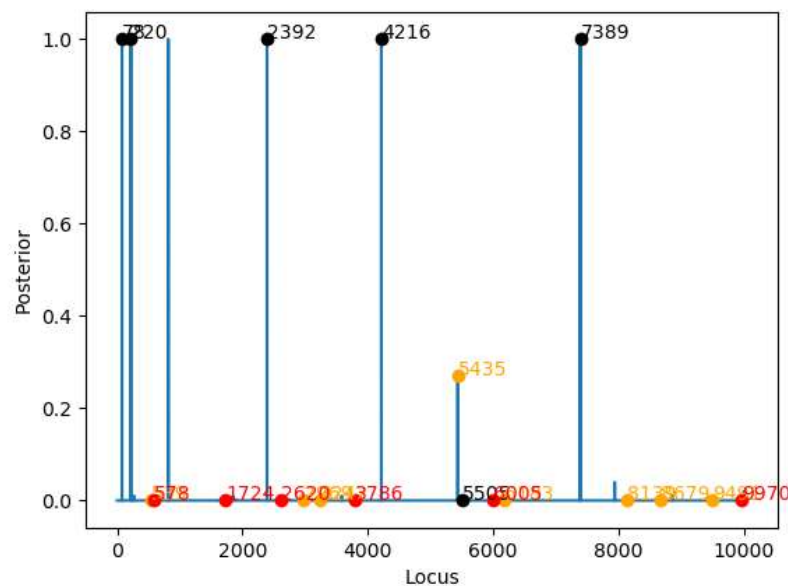


تشخیص بیماری مجموعه داده‌ی ۱

الگوریتم پیشنهادی



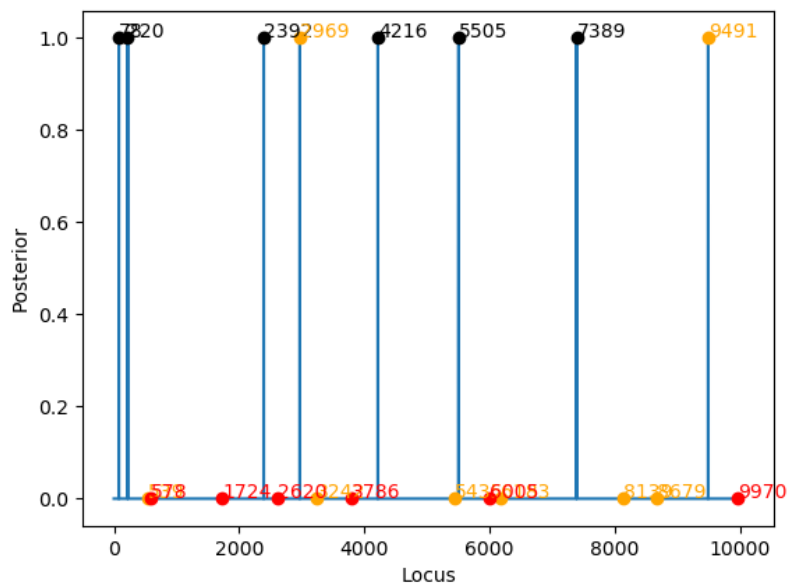
BEAM3



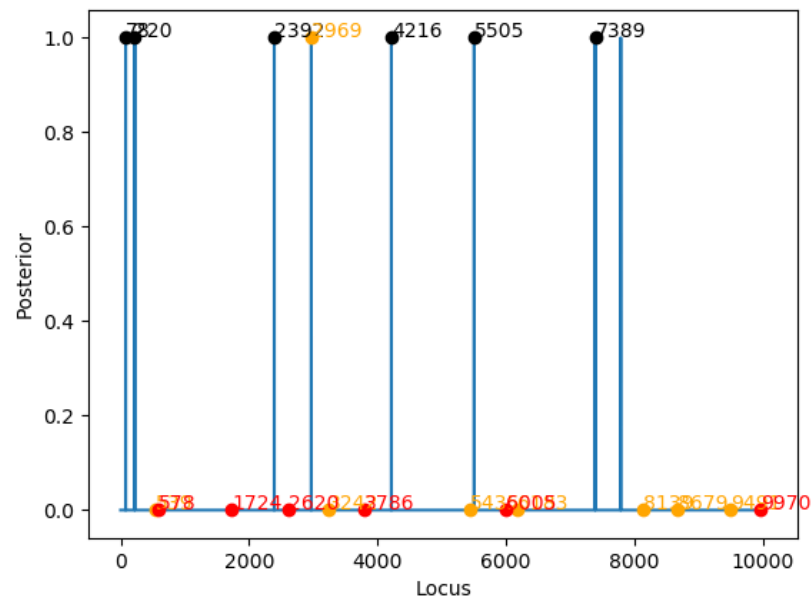


تشخیص بیماری مجموعه داده‌ی ۲

الگوریتم پیشنهادی



BEAM3





مقایسه تشخیص بیماری مجموعه داده‌های ۱ و ۲

الگوریتم پیشنهادی

- درست-کاذب‌های ندارد
- تمام تک جایگاهی‌ها شناسایی می‌شوند
- جایگاه‌های دوگانه بیشتری شناسایی می‌شود

BEAM

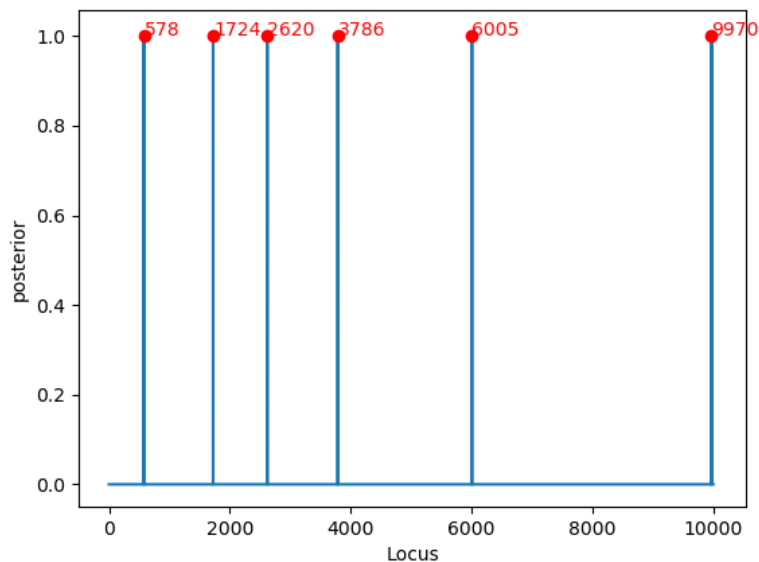
- گاهی درست-کاذب دارد
- گاهی تک-جایگاهی‌ها شناسایی نمی‌شوند
- یکی از جایگاه‌های دوگانه شناسایی می‌شود



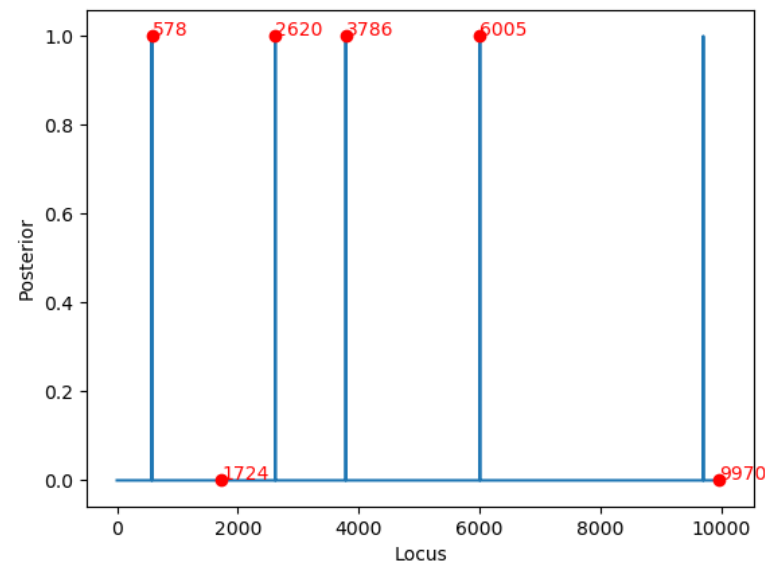
تشخیص بیماری

مجموعه داده‌ی ۳ (تأثیر خوشه‌بندی)

الگوریتم پیشنهادی



BEAM3





مقایسه تشخیص بیماری مجموعه داده‌ی ۳

الگوریتم پیشنهادی

- تمام جایگاه‌ها شناسایی می‌شوند
- درست-کاذب ندارد

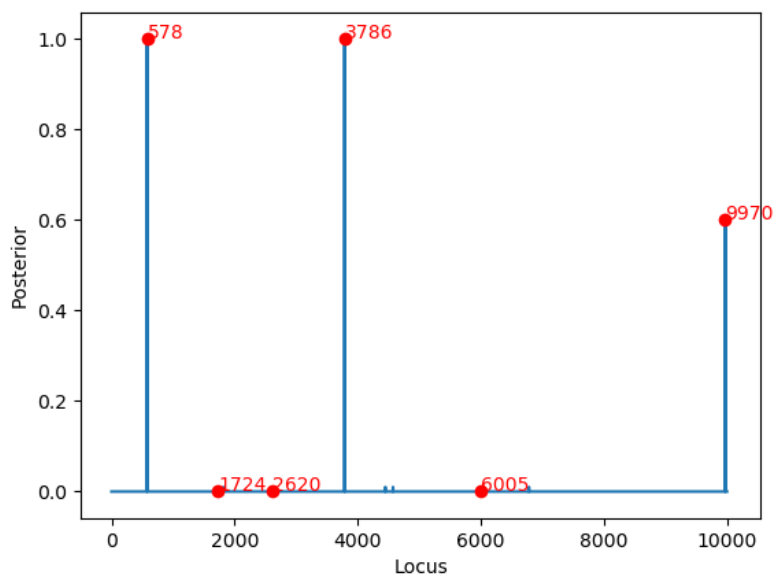
BEAM

- ۲ جایگاه‌ها شناسایی نمی‌شود
- درست-کاذب گزارش می‌شود

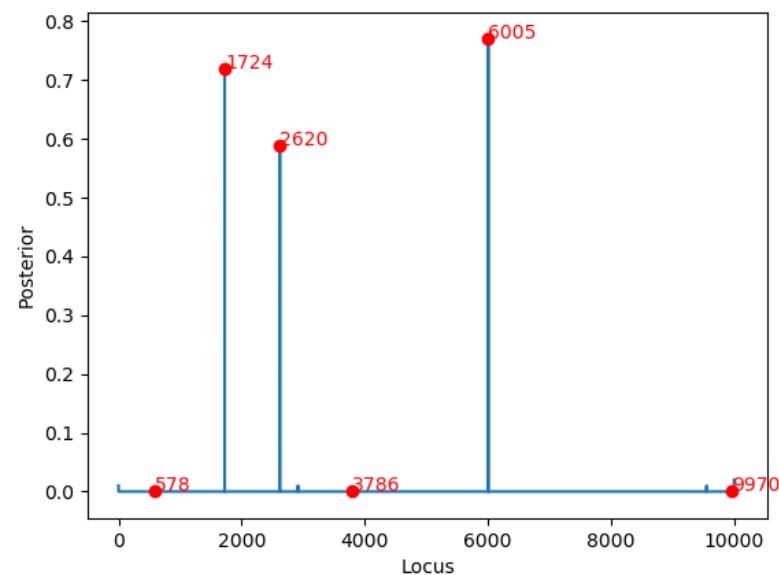


تشخیص بیماری مجموعه داده‌ی ۳ (تأثیر خوشه‌بندی) زیر جمعیت‌های جداگانه در BEAM3

زیر جمعیت ۲



زیر جمعیت ۱





تشخیص بیماری با زیر جمعیت‌های جداگانه

- افزودن مرحله پیش‌پردازش
- بهبود دقت تشخیص
- تمام جایگاه‌های خاص زیر جمعیت شناسایی می‌شوند
- درست-کاذب‌ها رفع می‌شود



- در نظر گرفتن ساختار جمعیت در تشخیص بیماری
- جایگزین کردن خوشه‌بندی MCMC با بیز گوناگونی
- جایگزین کردن مدل جستجوی فراگیر با روش یافتن روایستایی



- تولید داده‌های ساختگی با روش‌های پیچیده‌تر

- تولید داده‌های جمعیت
- تولید داده‌های بیماری

- توسعه‌ی الگوریتم برای داده‌های غیر مورد-کنترل

- داده‌های QTL

- داده‌های واقعی

- داده‌های موش (GSE2814) ۱۰۰٪ انجام شد ولی نتایج ضعیف بود
- داده‌های انسان (GSE68086) ۹۰٪ انجام شده

تشکر و تقدیر



- دکتر سید محمود طاهری
- دکتر سید مرتضی امینی
- دکتر فیروزه ریواز
- مهندس مهرداد تمیجی

