دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف

ژنومیک محاسباتی

مطهری و فروغمند
پاییز ۱۴۰۰

بازسازی درخت تبارزایی (۲)

جلسه پنجم

نگارنده: مجتبی زمانی

۱ مروری بر مباحث گذشته

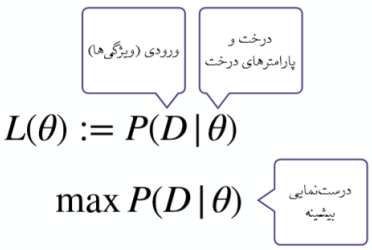
در درخت تبارزایی یک سری موجود داریم و می‌خواهیم از روی آن درخت بسازیم به طوری که موجودات شبیه به هم نزدیکتر باشند. با دو نوع ورودی می‌توانیم این درخت را بسازیم: ۱.به ازای هرموجود برداری از ویژگی‌ها داشته باشیم و موجوداتی که ویژگی‌هایشان شبیه به هم است نزدیک هم باشند ۲.ماتریس شباهت (اختلاف) مانند edit distance توالی‌ها.

در ویژگی_مبنا چند معیار برای انتخاب بهترین درخت وجود دارد: بیشینه صرفه‌جویی، مدل احتمالاتی و ...

در مسئله صرفه‌جویانه‌ترین درخت، یک ماتریس داریم که سطرهای آن گونه‌ها و ستون‌های آن ویژگی‌ها هستند و هدف ساخت یک درخت است که گونه‌ها روی برگ‌ها هستند، یک سری بردار ویژگی به راس‌های میانی نسبت دهیم به طوری که مجموع هزینه یال‌ها (تغییرات) کمینه شود. این حالت برای وقتی که درخت را داده باشند الگوریتم خوبی دارد اما در صورتی که درخت را نداده باشند NP-Hard است.

اگر پارامترهای یک مدل احتمالاتی (درخت و پارامترهای درخت) را داشته باشیم و بتوانیم احتمال مشاهده را حساب کنیم، به آن تابع likelihood گویند. درست‌نمایی بیشینه یعنی داده را داریم و می‌خواهیم پارامترها را به گونه‌ای پیدا کنیم که $P(D|\theta)$ بیشینه شود.

مسئله صرفه‌جویانه‌ترین درخت برای بعضی مدل‌ها خیلی شبیه به درست‌نمایی بیشینه است اما حل هر دو سخت است. این دو الگوریتم را روی داده‌های واقعی خیلی می‌سنجند، الگوریتم صرفه‌جویانه‌ترین درخت معمولا جواب‌های بهتری می‌دهد.



۲ مدل تکامل DNA

در مدل felsenstein cavender تکامل DNA را بررسی نکردیم و فقط احتمال تغییر یال را بررسی کردیم یعنی زمان را نادیده گرفتیم. همچنین احتمال تغییر حروف مختلف را درنظر نگرفتیم.

۱.۲ مدل تکامل مارکوف

P(t) داریم که یک بردار است که نشان می‌دهد در زمان t احتمال مشاهده یک اسید نوکلئیک چقدر است؟ فرض می‌کنیم که اسید نوکلئیک‌ها در مکان‌های مختلف، به طور مستقل تغییر می‌کنند. ابتدا با یک احتمالی هر کدام از نوکلئیک اسیدها را داریم و همینطور که زمان می‌گذرد با یک روندی تغییر می‌کنند. پس می‌توانیم یک بردار p(t) تعریف کنیم که در زمان t احتمال هرکدام را نشان دهد. فرض می‌کنیم که p از رابطه زیر تبعیت می‌کند:

$$p'(t) = p(t)Q$$

مشتق p یعنی تغییرات احتمال مشاهده هرکدام از نوکلئیک اسیدها به یک ماتریس وابسته است که نرخ اضافه‌شدن به هرکدام به شرط اینکه یک مقداری را داشته باشیم چقدر است. Q نرخ تغییرات است. جمع سطر ماتریس Q برابر ۰ است. با حل معادله بالا، جواب زیر بدست می‌آید:

$$p(t) = p(0)e^{tQ}$$

e به توان یک ماتریس یعنی بسط تیلور e را بنویسیم.
اگر این مدل شرط‌های خوبی داشته باشد یک توزیع پایای یکتا وجود دارد:

$$\pi e^{tQ} = \pi$$

توزیع پایای یکتا یعنی زمان که بگذرد احتمال‌ها تغییر نمی‌کند.
معمولا مدل‌هایی که داریم بازگشت پذیرند یعنی اگر در زمان t توزیع x را روی نوکلئیک اسید داشته باشیم و زمان بگذرد و توزیع y بدست آید، از y هم شروع کنیم زمان بگذرد x بدست آید. با این شرایط اگر درخت بسازیم درخت بدون ریشه و بدون جهت بدست می‌آید.

$$xe^{tQ} = y \Leftrightarrow ye^{tQ} = x$$

۲.۲ مدل JC۶۹

اگر بدانیم اسیدآمینه‌ها را با چه احتمالی داریم بعد از زمان t احتمال اسیدآمینه‌ها را با ماتریس P می‌توان بدست آورد. این مدل یک پارامتر μ دارد که نرخ جهش در یک واحد زمانی است و باید از داده‌ها پیدا شود. بنابراین با داشتن t می‌توان احتمال مشاهده هر اسیدآمینه را محاسبه کرد.

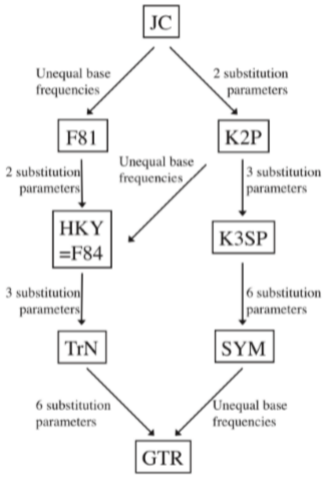
$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

۳.۲ سلسله مراتب مدل تکاملی DNA

GTR کلی‌ترین مدل مارکوف است که پارامترهای زیادی دارد. وقتی تعداد پارامتر زیاد شود overfitting اتفاق می‌افتد. اگر تعداد داده زیاد باشد حق داریم از مدل‌های بهتر با پارامترهای بیشتر استفاده کنیم.

اگر مدل تکاملی داشته باشیم می‌توانیم احتمال درخت و ویژگی‌های راس‌های میانی را حساب کنیم. اگر یک درخت و طول یال‌ها را داشته باشیم می‌توانیم احتمال درخت و طول یال‌ها را حساب کنیم.
ورودی: ماتریس ویژگی‌ها، خروجی: درخت، الگوریتم:؟



MCMC ۴.۲

فرض کنیم که یک سری چیز داریم و به هرکدام یک ارزش نسبت دادیم می‌خواهیم از آن‌ها نمونه بگیریم به طوری که احتمال نمونه‌گیری متناسب با f باشد. پیدا کردن جمع همه f ها سخت است. یک نقطه ابتدایی x_0 داریم. هر دفعه یک همسایه از توزیع $g(x'|x_t)$ انتخاب می‌کنیم. با این احتمال تغییر را می‌پذیریم:

$$A(x', x_t) = \min\left(1, \frac{P(x')}{P(x_t)} \frac{g(x_t \mid x')}{g(x' \mid x_t)}\right)$$

اگر الگوریتم خیلی ادامه پیدا کند برای t بزرگ، احتمال $P(x_t) = x_t$ یعنی متناسب با f است. مسئله دیگر این است که ماکسیمم likelihood را مانند همین روش بدست آوریم.

۳ بازسازی فاصله مبنای درخت تبارزایی

در این مسئله ماتریس فاصل دو به دو داریم و می‌خواهیم از روی آن درخت بسازیم. ماتریس فاصله باید خاصیت‌های زیر را داشته باشد:
متقارن باشد، فاصله هر گره با خودش صفر است و رابطه نامساوی مثلثی برقرار است.

$$\textbf{Symmetric: } M_{ij} = M_{ji} \text{ and } M_{ii} = 0; \text{ and}$$

$$\textbf{Triangle Inequality: } M_{ij} + M_{jk} \geq M_{ik}.$$

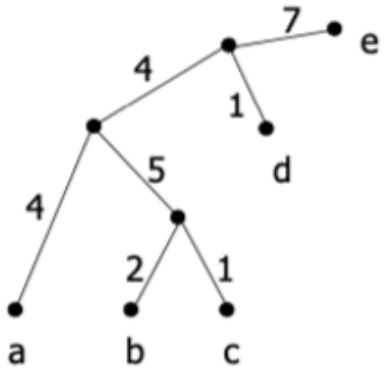
خروجی یک درخت فیلوژنی است که بدون ریشه است، درجه همه راس‌های میانی ۳ است و یال‌های درخت طول دارند.

۱.۳ حالت خاص: ماتریس فاصله جمعی

ماتریس M ، ماتریس فاصله جمعی است اگر و فقط اگر:

- یک درخت با وزن یال‌های مثبت وجود داشته باشد.
- فاصله دو برگ = جمع فاصله یال‌های بین آن‌ها

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



۲.۳ حالت خاص: ماتریس ابر متریک

- M جمعی باشد.
- درخت ریشه دار T با وزن‌های مثبت
- فاصله همه برگ‌ها با ریشه برابر باشد.

M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0