الگوی مارکف پنهان چیست؟

شون آر. ادی
بخش ژنتیک دانشگاه واشنگتن

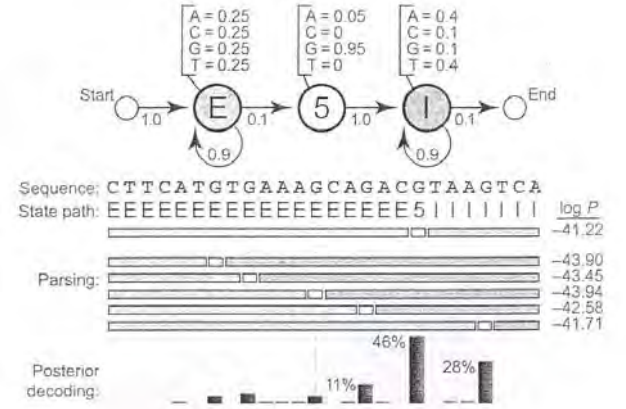
الگوهای آماری بنام الگوهای مارکف پنهان در زیست‌شناسی محاسباتی به دفعات مورد ارجاع قرار می‌گیرند. این مدل‌ها چه هستند و چرا این قدر برای مسائل متعدد مفیدند؟ تجزیه و تحلیل توالی‌های زیستی اغلب به عنوان قرار دادن برچسب درست روی هر عضو توالی تعریف می‌شود (که آن عضو را معمولاً با نام residue می‌شناسند). در شناسایی ژن می‌خواهیم برچسب روی نوکلئوتیدها را تحت عناوین اگزون‌ها، اینترون‌ها و یا توالی مابین ژنی قرار دهیم. در بحث هم ردیفی توالی‌ها به دنبال نسبت دادن اعضای یک دنباله ناشناش به اعضای همسان آن در یک مجموعه هدف از دنباله‌ها هستیم. ما همیشه می‌توانیم یک برنامهٔ خاص برای یک مسألهٔ داده شده بنویسیم اما غالباً با مشکلات مشابهی مواجه می‌شویم. یکی از مشکلات عبارت است از به کارگیری منابع ناهمگون اطلاعات. برای مثال یک ژن‌یاب باید نواحی مشترک برای برش، اریبی کودونی، ترجیه طولی اگزون اینترون و آنالیز چارچوب باز را در یک سیستم نمره‌دهی ترکیب کند. چطور این پارامترها باید تنظیم شوند؟ به اطلاعات گوناگون چگونه باید وزن داد؟ مشکل دیگر تعبیر و تفسیر نتایج به صورت احتمالی است. یافتن بهترین نمره (امتیاز) پاسخ یک هدف است ولی معنی نمره چیست و چقدر مطمئنیم که بهترین نمره درست را یافته‌ایم؟ سومین مسئله، تعمیم‌پذیری است. هم زمان با کامل ساختن ژن‌یاب ویژه آرزو می‌کنیم ای کاش اجماع اولیهٔ انتقالی ، برش‌های دیگر و علامت پلیا دنیلی را هم مدل می‌کردیم.

اغلب اوقات با افزودن حقایق بیشتر به یک برنامهٔ ویژه ظریف موجبات فروریزی آن را زیر بار وزن سنگینش پدید می‌آوریم. الگوهای مارکف پنهان(HMM) مبانی رسمی الگوهای احتمالی برچسب‌گذاری بر توالی‌های خطی را ارائه می‌دهند. آن‌ها ابزار مفهومی برای ساخت مدل‌های پیچیده با رسم تصویری ذهنی پدید می‌آورند. آن‌ها در قلب دامنه وسیعی از برنامه‌ها مثل، ژن‌یاب‌ها، جست و جوگرهای پرو فایلی، همردیف‌یاب‌های چندگانه، توالی‌ها و تعیین کننده‌های محل‌های تنظیم ژنی قرار دارند. HMMها نمادهای تجزیه و تحلیل محاسباتی توالی‌ها هستند.

یک HMM نمایشی تشخیص دهندهٔ محل بر ۵

به عنوان یک مثال ساده مسأله تشخیص محل برش ۵ در نمودار زیر را در نظر بگیرید. فرض کنید یک توالی DNA داریم که از یک اگزون شروع شده و یک محل برش ۵ دارد و به یک اینترون ختم می‌شود. مسئله تعیین محل برش ۵ است. یعنی جایی که اگزون به اینترون تغییر پیدا می‌کند ـ یا به عبارتی مسئله تعیین ۵ SS است. برای ما که می‌خواهیم حدسی هوشمندانه بزنیم توالی‌های اگزون، محل‌های برش و اینترون‌ها باید خواص آماری مختلفی داشته باشند. اجازه دهید اختلاف‌های ساده را در نظر بگیریم مثلاً در اگزون توزیع بازها به‌طور متوسط یکنواخت است (یعنی ٪۲۵ از هر کدام) در اینترون ها)A/T زیاد است (مثلاً ۴۰ درصد از هر یک از A،T و ۱۰ درصد از دو حر ف دیگر C و G) و اجماع نوکلئوتیدی در محل $5'SS$ تقریبا همیشه G است (مثلاً ٪ ۹۵ اوقات G و ٪۵ اوقات A).

با شروع از این اطلاعات، می‌توانیم یک HMM رسم کنیم (شکل ۱). این HMM سه وضعیت را به وجود می‌آورد که هر یک از این‌ها را به‌عنوان برچسبی به هر نوکلئوتید قرار می‌دهیم E(اگزون) ۵($5'SS$) و I (اینترون). هر وضعیت، دارای احتمال‌های انتشار مختص خودش است (که در بالای وضعیت‌ها نشان داده شده‌اند)، این احتمال‌ها ترکیب بازهای اگزون، اینترون‌ها و اجماع G در محل $5'SS$ را الگو می‌کنند. همچنین هر وضعیت دارای احتمال‌های انتقال است (کمان‌ها) که احتمال‌های حرکت از هر وضعیت به وضعیت دیگر را نشان می‌دهد. احتمال‌های انتقال، رتبه خطی را که طی آن انتظار داریم وضعیت‌ها رخ دهند، توجیه می‌کنند یک یا چند E ، یک یا چند ۵، یک یا چند I.



شکل ۱ یک HMM نمایشی برای تشخیص محل برش ۵.

پس چه چیز پنهان است؟

مفید است تصور کنیم که یک HMM یک توالی تولید می‌کند. وقتی وضیعتی را ملاقات می‌کنیم از توزیع احتمال انتشارها یک عضو منتشر می‌نماییم. سپس بر مبنای توزیع احتمال وضعیت‌ها، وضعیت بعدی را انتخاب می‌کنیم بنا بر این مدل، دو رشته از اطلاعات تولید می‌شود. یکی مسیر وضعیت‌ها (برچسب ها) است که ما از یک وضعیت به وضعیت دیگر منتقل می‌شویم. دیگری توالی مشاهدات(AND) است که از هر وضعیت داخل مسیر یک عضو منتشر می‌شود.

مسیر وضعیت، یک زنجیر مارکف است یعنی هر وضعیتی که قرار است برویم فقط به وضعیت فعلی‌مان بستگی دارد. از آنجایی که به ما فقط توالی مشاهدات را می دهند، این مسیر وضعیت، پنهان است ـ این‌ها برچسب‌های عضوها هستند که علاقمند به استنباط راجع به آن‌ها هستیم. مسیر وضعیت یک زنجیر مارکف

پنهان است.

احتمال $P(S,\pi|HMM,\Theta)$ که HMM با پارامترهای Θ مسیر وضعیت و توالی مشاهدات S را تولید کند برابر است با حاصل ضرب همهٔ احتمال‌های انتشارها در احتمال‌های انتقال‌های به وقوع پیوسته. برای مثال توالی ۲۶–نوکلئوتیدی به همراه مسیر وضعیت را در وسط شکل ۱ در نظر بگیرید که در آن در مجموع ۲۷ انتقال و ۲۶ انتشار وجود دارد. همهٔ ۵۳ احتمال را در هم ضرب کنید (و لگاریتم‌گیری انجام دهید، زیرا که این‌ها اعداد کوچکی هستند) شما عدد $\log P(S,\pi|HMM,\Theta) = -41.22$ را به دست خواهید آورد.

یک HMM یک الگو کامل احتمالی است ـ پارامترهای مدل و امتیاز کل توالی همگی احتمال‌ها هستند. بنابراین می‌توانیم از نظریه احتمال بیزی برای دست کاری این اعداد با استفاده از روش های توانمند استاندارد شامل بهینه‌سازی پارامترها و تفسیر معنی‌داری امتیازها، بهره جوئیم.

یافتن بهترین مسیر وضعیت

در یک مسئله تجزیه تحلیلی به ما توالی را داده‌اند و ما می‌خواهیم مسیر وضعیت پنهان را پیدا کنیم. مسیرهای زیادی پتانسیل آن را دارند که یک توالی را تولید کنند. ما به دنبال توالی هستیم که بیشترین احتمال را داراست.

برای مثال، اگر به ما HMM و توالی ۲۶ – نوکلئوتیدی شکل ۱ را بدهند، ۱۴ مسیر ممکن با احتمال غیر صفر وجود دارد، زیرا $5'SS$ باید روی یکی از ۱۴ تا A یا G داخلی بیفتد. شکل ۱ شش تا از مسیرهای با امتیاز بالا را نشان می‌دهد (مسیرهایی با G در $5'SS$). بهترین مسیر دارای لگاریتم احتمال ۲۲.۴۱– است که نتیجه می‌دهد که محتملترین وضعیت $5'SS$ در پنجمین G است.

در اکثر مسائل تعداد توالی های ممکن از وضعیت ها آنقدر زیاد است که قادر به نمایش همهٔ آن‌ها نیستیم. الگوریتم کارآی و یتربی ضمانت می‌کند که با در دست داشتن توالی مشاهدات و MMH محتملترین مسیر وضعیت را پیدا کند. الگوریتم یتربی یک الگوریتم برنامه‌ریزی پویا بسیار شبیه به آن الگوریتم‌های استانداردی است که در همردیفی توالی‌ها به کار گرفته می‌شوند.

بعد از همردیفی‌های بهترین امتیاز

شکل ۱ نمایانگر آن است که یک مسیر وضعیت مختلف دارای کمی اختلاف در امتیاز با حالتی است که $5'SS$ در پنجمین G قرار می گیرد(لگاریتم احتمال ۴۱.۷۱– در مقابل ۴۱.۲۲–). چقدر اطمینان داریم که پنجمین G، انتخاب درستی است؟ این از ویژگی‌های الگوهای احتمالی است ما می‌توانیم اطمینانمان را مستقیم محاسبه کنیم. احتمال این که یک عضو توالی مثل i توسط وضعیت K منتشر شده باشد برابر است با مجموع احتمال‌های تمام مسیرهای وضعیت که از وضعیت K برای انتشار استفاده می‌کنند (یعنی $\pi i = K$ در مسیر وضعیت) تقسیم بر مجموع همه مسیرهای وضعیت ممکن. در مثال نمایشی ما، این یک مسیر خاص در صورت کسر است که به مجموع ۱۴ مسیر وضعیت در مخرج کسر تقسیم شده است. ما برای حالتی که بهترین امتیاز پنجمین G درست باشد، احتمالی برابر ۴۶ به دست می آوریم ،حال آنکه برای حالتی که ششمین G درست باشد احتمال برابر برابر با ۲۸ را خواهیم داشت (شکل ۱، پایین). این را رمزگشایی پسین(posterior decoding) می‌نامند. برای مسائل بزرگتر، رمزگشایی پسین از دو الگوریتم به نام‌های پیشرو و پسرو استفاده می‌کنند. این‌ها اساساً شبیه الگوریتم ویتربی می‌باشند فقط به جای پیدا کردن محتمل ترین مسیر به جمع‌بندی روی تمام مسیرهای ممکن توجه می‌کنند.

ساختن الگوهای واقع‌گرایانه‌تر ساختن یک HMM یعنی مشخص نمودن چهار چیز الفبای علائم K علامت مختلف (مثل ACGT، $k = 4$) (ii) تعداد وضعیت‌های مدل M؛ (iii) احتمال‌های انتشار $ei(x)$ برای هر وضعیت i که جمع روی K علامت x برابر ۱ می‌شود یعنی $\sum ei(x) = 1$ و (vi) احتمال‌های انتقال ti(j) برای هر وضعیت i که به وضعیتی مانند j (شامل خودش) می‌رود و جمع روی M وضعیت j برابر ۱ می‌شود یعنی $\sum ti(j) = 1$. هر مدلی که این خواص را داشته باشد یک HMM است.

یعنی هرکس می تواند با رسم تصویر از مسئله ی مورد نظرش شبیه به مسئلهٔ شکل ۱ یک HMM جدید بسازد. این سادگی گرافیکی اجازه می‌دهد هرکس به تعریف بیولوژیک خاصی از یک مسئله متمرکز شود. برای مثال در الگوی نمایشی ما یعنی الگوی محل‌های برش ممکن است از قدرت ممیزی مدل راضی نباشیم ممکن است بخواهیم یک اجماع واقع‌گرایانه‌تر شش نوکلئوتیدی GTRAGT را به محل برش $5'$ اضافه کنیم می‌توانیم سطری با شش وضعیت HMM به جای وضعیت ۵ قرار دهیم تا یک اجماع شش باز را با پارامتری کردن احتمال‌های انتشار در محل‌های برش $3'$ الگو کنیم. ممکن است بخواهیم یک مدل کامل اینترونی شامل یک محل برش $3'$ را مدل کنیم و یک $3'$ به عنوان وضعیت اگزونی به مدل اضافه کنیم تا اجازه دهد یک توالی مشاهده شده به جای یک اینترون به یک اگزون ختم شود. بنابراین ممکن است علاقمند باشیم یک مدل کامل ژنی بسازیم. هر چه که اضافه می کنیم فقط کافی است به صورت ترسیمی مشخص شود.

جالب توجه HMM ها به وابستگی بین اعضا توالی مشاهدات به نحو مناسب نمی‌پردازند زیرا آن‌ها فرض می‌کنند به شرط وضعیت داده شده مشاهدات از هم مستقلند. مثالی که در آن مدل‌های HMM اغلب نامناسبند، تجزیه تحلیل ساختار دوم RNA است. جفت بازهای حفظ شده RNA همبستگی‌های جفتی با طول بلند ایجاد می‌کنند یکی از موقعیت ها می‌تواند هر یک از اعضای توالی باشد اما جفت باز شریک باید مکمل آن باشد. یک وضعیت مسیر HMM هیچ راهی به جز به یاد آوردن این مطلب که کدام وضعیت دور تولید شده است، ندارد. گاهی اوقات می توانیم قوانین HMM را بدون آن که الگوریتم‌ها را خراب کنیم، دور بزنیم. برای مثال ممکن است در بحث ژن‌یابی بخواهیم به جای انتشار سه عضو توالی، یک سه تایی وابسته از کودون‌ها را انتشار دهیم الگوریتم‌های HMM بدون هیچ مشکلی قابل تعمیم به انتشار وضعیت‌های

سه‌تایی می‌باشند. به هر حال، تا الان HMMهای پایه‌ای به کار گرفته شده‌اند. کلاس‌های الگوهای احتمالی پرتوانتری (اگرچه با کارایی کمتر) از HMMها برای تجزیه و تحلیل توالی‌ها وجود دارند.

مراجع

1. Rabiner, L.R. A Tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257-286(1989)
2. Durbin, R., Eddy, S.R., Krough, A. and Mitchinson, G.J. Biological sequence analysis: probabilistic models of proteins And nucleic acids (Cambridge University press, Cambridge UK. 1998).

این نوشته برگردانی از مقاله زیر است:

Eddy, S., R. What is a Hidden Markov Model? Nature Biotechnology Vol. 22 No. 10 October 2004.

مترجم: حمید پزشک
بخش آمار دانشگاه تهران
و هسته بیوانفورماتیک پژوهشگاه دانش‌های بنیادی

★ ★ ★

مصاحبه

مصاحبه با لاسلو لُواس

لاسلو لُواس در حال حاضر رئیس اتحادیهٔ بین‌المللی ریاضیات (IMU) است. او در سال ۱۹۴۸ در بوداپست مجارستان به دنیا آمده و تمام تحصیلات خود را در مجارستان گذرانده است. لُواس یکی از افراد سرشناس در ترکیبیات می‌باشد و به خاطر کارهای عمیقی که انجام داده، جوایز ولف و کنوت را در سال ۱۹۹۹ و جایزه بولیایی را در سال ۲۰۰۷ از آن خود کرده است. لُواس در طول تحصیل در دبیرستان ۳ بار برنده مدال طلا از المپیاد جهانی ریاضیات شده است. پسر او نیز در سال ۲۰۰۸ موفق به کسب این مدال گردید.

لُواس در بهار ۱۳۸۶ به دعوت پژوهشکده ریاضیات پژوهشگاه دانش‌های بنیادی در کنفرانس نظریه جبری گراف شرکت کرد و چند سخنرانی ارائه داد. در خلال این کنفرانس از ایشان برای یک مصاحبه اختصاصی با خبرنامه دعوت به عمل آمد که با گشاده‌رویی دعوت را پذیرفتند.

این مصاحبه توسط آقایان دکتر: **مهدی بهزاد**، **سعید اکبری**، **منوچهر ذاکر** و **رشید زارع‌نهندی** انجام شد. از پژوهشکده ریاضیات که امکانات این مصاحبه را فراهم کرد و از خانم **عاطفه پارسا** به خاطر ضبط مصاحبه و عکس‌ها و آقای **فرزین منیعی** و خانم **سیده‌صادقه حق‌شناس** که در پیاده‌سازی قسمتهایی از مصاحبه کمک کردند، سپاس‌گزاری می‌شود. پیاده‌سازی نهایی و ترجمه مصاحبه توسط **منوچهر ذاکر** و **رشید زارع‌نهندی** انجام شده است.

- **اکبری: نخست از این که دعوت ما را برای مصاحبه پذیرفتید، متشکریم. لطفاً اگر خاطراتی از دوران نوجوانی خود و نحوه آشنایی با افرادی مانند اردوش (P. Erdös)، پلیکان (J. Pelikan) و سایر ترکیبیات‌دانان زبده دارید، بفرمایید.**

اولین قدم من در ورود به ریاضیات زمانی بود که در دبیرستان بودم. آن زمان در مجارستان کلاس‌های ویژه‌ای برای دانش‌آموزان مستعد بر پا کرده بودند و من هم وارد یکی از آنها شدم. دانش‌آموزان بسیار خوبی در آن کلاس بودند و من خیلی خوش‌شانس بودم که فرصت هم‌کلاسی با آنان را داشتم. استادانی از دانشگاه به این کلاس‌ها می آمدند و درس‌های ویژه‌ای ارائه می‌کردند و موضوعات