
بسم الله الرحمن الرحیم

دانشکده علوم
گروه ریاضی

پایان‌نامه

برای دریافت درجه کارشناسی ارشد در رشته
ریاضی محض گرایش هندسه جبری

عنوان

هندسه جبری مدلهای فیلوژنتیک

نگارش

زینب صالح نژاد

استاد راهنما

دکتر حسن حقیقی

استاد مشاور

دکتر فرشته ملک

آذر ماه ۱۳۹۰

نام خانوادگی دانشجو: صالح نژاد نام: زینب

عنوان: هندسه جبری مدلهای فیلوژنتیک

استاد راهنما: دکتر حسن حقیقی

استاد مشاور: دکتر فرشته ملک

مقطع تحصیلی: کارشناسی ارشد رشته: ریاضی محض گرایش: هندسه جبری

دانشگاه: دانشگاه صنعتی خواجه نصیرالدین طوسی دانشکده علوم

تاریخ فارغ‌التحصیلی: آذر ماه ۱۳۹۰ تعداد صفحات: ۱۱۷

تقدیم به

پدر دلسوز

مادر مهربان

و

همسر عزیزم

به پاس همراهی و محبت های بی دریغشان

سپاس و ستایش مخصوص خداوندی است که انسان را آفرید و او را به فضیلت تعلیم و تعلّم بر دیگر مخلوقات خود برتری بخشید.

خداوندا، تو را حمد و سپاس می گویم که همواره یاری رسانم بوده ای و دریچه های علم و معرفت را فرارویم گشوده ای.

همچنین از استاد ارجمندم، جناب آقای دکتر حسن حقیقی که در طول این پژوهش، همواره راهنما و پشتیبان من بوده اند و نیز از سرکار خانم دکتر فرشته ملک که مشاوره این پژوهش را بر عهده داشته اند، تشکر و قدردانی می کنم.

در ادامه نیز، از جناب آقای دکتر سیامک یاسمی و جناب آقای دکتر هاشم پروانه مسیحا که زحمت داوری این پایان نامه را تقبل فرمودند، سپاسگزاری می نمایم.

زینب صالح نژاد
آذر ماه ۱۳۹۰

چکیده

فیلوژنتیک شاخه ای از علم زیست شناسی است که با استفاده از داده های موجود (به عنوان مثال DNA) تحول مولکولی را استنباط می کند. از میان مدلهای ریاضیاتی که برای تسهیل این استنباط مورد استفاده قرار می گیرد، مدلهای آماری کاربرد وسیع تری دارند. در این گونه از مدلها، می توان فرض کرد که تحول مولکولی به عنوان یک فرایند احتمالاتی در امتداد یک درخت ریشه دار که تنها برگهای آن قابل مشاهده اند و سایر گره هایش پنهانی هستند، اتفاق می افتد.

برای پارامتری سازی تمام مدلهای آماری فیلوژنتیک یک درخت خاص، می توان مفاهیم و تکنیک های هندسه جبری را به کار گرفت. این کاربست که به هندسه جبری فیلوژنتیک موسوم است، به ما این امکان را می دهد که ایده آل های همه ی چندجمله ای هایی که روی توزیع های توأم حالت های مختلف برگ های یک درخت صفر می شوند را تحت عنوان پایاهای فیلوژنتیک مورد توجه قرار دهیم.

به عنوان یک کاربرد، می توانیم مقادیر پایاهای فیلوژنتیک را، در صورتی که نزدیک صفر باشند، روی فراوانی های مشاهده شده امتحان کنیم، سپس کارایی مدل ساخته شده را نتیجه بگیریم.

در این پایان نامه، از روش فوق برای پاسخ دادن به سؤالات خاصی در فیلوژنتیک استفاده می شود. با اعمال شرایط گوناگون روی متغیرهای اختصاص داده شده به گره های یک درخت، واریته های جبری متفاوتی ساخته می شود که از این بین می توان به واریته های دترمینانی، واریته های ورونزه، واریته های قاطع واریته های سگره، واریته های توریک و ... اشاره کرد. با فرض این که مدل آماری ما از مدل مارکوف عمومی پیروی می کند، واریته ی نظیر این مدل را

ساخته و پایاهای فیلوژنتیک آن را در شرایط معین به دست می آوریم.

برای به دست آوردن پایاهای فیلوژنتیک این مدل، از تخت سازی های یالی و رأسی تانسورهایی که به برگها نظیر می شود، استفاده می کنیم و از این طریق حدسیه ای که توسط اشتورم فلس[1] مطرح شده را به اثبات می رسانیم. سرانجام نشان می دهیم که نتایج مطرح شده در این پایان نامه، در حالت کلی و برای درختهای غیر دودویی نیز معتبر است.

[1] Sturmfels

فهرست مطالب

مقدمه

معمای منشاء حیات موجودات زنده، همواره یکی از پیچیده ترین و مباحثه انگیزترین موضوعاتی بوده است که ذهن انسان ها را به خود مشغول داشته است، و حاصل تلاش بشر برای حل معمای حیات، علم زیست شناسی بوده است.

همانند هر شاخه ی علمی، در زیست شناسی نیز طبقه بندی موجودات زنده به منظور درک و فهم بهتری از معمای منشاء حیات، یکی از وظایف آن رشته محسوب می شود.

اولین گام های اساسی در این راستا را چارلز داروین [۲] زیست شناس معروف انگلیسی، در قرن نوزدهم برداشت و حاصل مشاهدات و مطالعات خود را در کتابی تحت عنوان "منشاء انواع " در دهه ی ۵۰ قرن نوزدهم میلادی منتشر ساخت. اساس نظریه ی داروین درباره ی منشاء انواع، بر دو اصل زیر استوار است:

اولا همه ی گونه ها به یکباره آفریده نشده اند.

ثانیا این گونه ها به شکل اولیه ی خود باقی نمانده اند و همواره و به تدریج در حال تغییرند.

درستی این نظریه در طی سال ها بعد از انتشار آن تقویت گردید و به شکلی منسجم، سبب طراحی برنامه ای برای یافتن منشاء حیات، از طریق مطالعه ی گونه های متشابه جانداران گردید. دانشمندان علم زیست شناسی، اطلاعات بسیاری درباره ی موجودات زنده و گونه های منقرض شده جمع آوری کرده اند و به کمک اطلاعاتی که از فسیل ها و سنگواره ها به دست آورده اند، تا اندازه ی زیادی در جهت یافتن پرسش اصلی درباره ی منشاء حیات پیش رفته اند. اما هنوز تا یافتن جواب قطعی راهی بسیار طولانی باید پیموده شود.

بر اساس یافته های علمی موجود، ثابت شده است که تنوع گونه ها، تفاوت جانداران و سلول ها، پی آمد مستقیم ویژگی های پروتئین و اسید نوکلئیک موجود در آنهاست. شباهت ها و تفاوت ها محصول واکنش شیمیایی است که در گونه های مختلف جانداران رخ می دهد. هر قدر طبیعت این واکنش ها و آهنگشان متفاوت تر باشد، تنوع گونه ای بیشتر خواهد بود.

---

[۲] Darwin

سنتز پروتئین به کنترل ویژگی نیاز دارد، بدین معنی است که باید نقشه ای موجود باشد تا ترتیب دقیق به هم پیوستن مولکول های اسید آمینه، با تعداد و انواع معین، را در هر ملکول تعیین کند. این کنترل ویژگی را در نهایت ژن ها، یعنی مولکول های DNA سلول اعمال می کنند.
هر یک از رشته های بازهای نیتروژن دار مولکول DNA، یا ژن ها، نماینده ی دستور العمل های شیمیایی است که به صورت رمز است. کنش بسیاری از این رشته ها، تعیین کردن به هم پیوستن مولکول های اسید آمینه در مولکول پروتئین است. به گفته ی دیگر، ژن های مختلف سلول، حاصل دستور العمل های مختلفی برای پروتئین سازی هستند. سلول فقط پروتئین هایی را می تواند بسازد که این دستور العمل های ژنتیکی تقریر می کنند. تحولات هنگامی روی می دهند که ژن ها تغییر یابند. ژن های تحول یافته، سیر طبیعی تغییر دستورات را به تدریج طی کرده و اندام های جدیدی می سازند.
به این ترتیب ملاحظه می شود، مولکول های DNA در بدن تمامی موجودات زنده وجود دارند و مطالعه آنها می تواند اطلاعات مفیدی درباره ی اجداد گونه های موجود به دست دهند.

با پیشرفت های علمی ای که از این رهگذر در حوزه ی زیست شناسی روی داده است، و همچنین پیدایش روش ها و ابزار های علمی جدید، در دو دهه ی آخر قرن بیستم، بشر موفق شده است دنباله های بزرگ اطلاعات ژنتیکی گونه های زنده را بخواند و اطلاعات در آنها را به دست آورد و شاخه ای جدید را بنا نهد که تحول مولکولی نامیده شده است.

فرض بر این است که تحول موجودات زنده، یک روند انشعاب است که به موجب آن جمعیت هایی از موجودات زنده، در طی زمان تغییر می کنند و بر اثر تغییرات به شاخه های جداگانه ای گونه زایی می کنند. به علاوه می توان این فرایند را به صورت یک درخت توصیف کرد که درجه ی هیچ یک ار رأس های آن از ۳ تجاوز نمی کنند. در این درخت ها برگ ها نشان دهنده ی گونه های موجودند که با قطعات DNA و آنچه اصطلاحا توالی یابی دنباله ای نامیده می شود و مقایسه کردن عناصر متناظر هر دنباله، در پی یافتن منشاء و بنای آنها می باشند.
رأس های محل انشعاب ممکن است نشان دهنده ی اجداد منقرض شده و ناشناخته ی این گونه ها باشند، به همین دلیل آنها را رأس های پنهان می نامند و عملا در شناختن منشاء گونه ها کمک مؤثر نمی تواند بکند.

بر این اساس شاخه ای از علم زیست شناسی پدید آمده است، که تلاش می کند با استفاده از اطلاعات موجود درباره ی گونه های معین و قابل مشاهده، درباره ی اجداد و نیاکان آنها و تاریخچه ی تحول آنها مطالعه می کند. این شاخه که به بررسی ارتباط تحولی بین گونه های مختلف جانداران می پردازد، فیلوژنتیک نام گرفته است.

فیلوژنتیک از دو کلمه ی یونانی فیل، به معنای قبیله یا نژاد و ژنتیکوس، به معنای زایشی و یا مربوط به زایش ساخته شده است.

مساله ی اصلی در فیلوژنتیک این است که اطلاعات ژنتیکی فقط برای موجودات امروزی وجود دارند و سوابق فسیلی حاوی شاخه های ریخت شناسی، مبهم و حاوی اطلاعات کمتری هستند. یک درخت فیلوژنتیک نشان دهنده ی فرضیه ای است درباره ی روندی که طبق آن رویدادهای تحولی منجر به گونه های مفروض مورد مطالعه گردیده است. ارتباط تحولی بین موجودات به وسیله ی درخت فیلوژنتیک نمایش داده می شود. از آنجا که تحول در طول دوره های زمانی طولانی که به طور مستقیم قابل مشاهده نیستند اتفاق می افتد، زیست شناسان مجبورند فیلوژنی ها را با استنباط روابط تحولی میان گونه های امروزی، بازسازی کنند. فسیل ها و سنگواره ها می توانند به بازسازی فیلوژنی ها کمک کنند، اما اطلاعات مربوط به فسیل ها اغلب بیش از حد کمیاب هستند و کمک خوبی به حل مساله ی مورد نظر نمی کنند. به همین دلیل، امروزه داده های مولکولی شامل پروتئین ها و رشته های DNA برای ساخت درخت های فیلوژنتیک مورد استفاده قرار می گیرند.

زیست شناسان برای مطالعه ی چنین درخت هایی، تنها به دانش زیست شناسی بسنده نکرده اند و از سایر دستاوردهای علمی موجود، از جمله ریاضیات، بهره ی فراوان برده اند. معروف است که ارنست مندل[۳]، از طریق کاشت هزاران گونه ی مختلف نخود فرنگی خود و گرده افشانی مصنوعی بر روی آنها و بررسی خصوصیات محصولات بدست آمده، توانست قوانینی کشف کند که چگونگی انتقال خصوصیات موجودات زنده را بدون در نظر گرفتن مدت انتقال آن، نشان می داد. به این ترتیب وی روشی آماری برای کشف قوانین وراثت را به کار برد.

ریاضیاتی که امروزه برای استنباط فیلوژنتیکی به کار گرفته می شود، مدل سازی، آمار، الگوریتم، ترکیبیات، جبر، هندسه و ... را شامل می شود. معمول ترین روش مورد استفاده برای استنباط

[۳] Ernest Mandel

فیلوژنی، شامل اصل کمترین فرضیات یا پارسیمونی، درست نمایی حداکثری و استنتاج بیزی می باشد. روشی نیز در اواسط قرن بیستم مورد استفاده قرار گرفت که امروزه مورد استفاده قرار نمی گیرد و تقریبا منسوخ شده است. این روش از طریق مقایسه دنباله های DNA گونه های موجود و شمارش تفاوت آنها، سعی داشت تا گونه های نزدیک به هم را شناسایی و به این طریق روند تحول آنها را مشخص نماید. در سال ۱۹۸۷ کاوندر [۴] و فلزنشتاین[۵] برای اولین بار متوجه شدند که می توان از مفاهیم جبری برای استنتاج درخت فیلوژنتیک گونه های موجود استفاده کرد و پایاهایی را در این رابطه معرفی کردند. این ایده سرآغاز کاربست روش های جبری و هندسی در مطالعه ی درخت های فیلوژنتیک گردید. در ادامه نیز آلمن [۶] و رودز[۷] و به علاوه اشتورم فلس و همکارانش با به کار بستن روش های جبری در بررسی مدل های فیلوژنتیک، تحولی نو در مطالعه ی درخت های فیلوژنتیک پدید آوردند.

در این پایان نامه قصد داریم این روش ها را در حد امکانات موجود، معرفی و به صورت مبسوط به تشریح آنها بپردازیم. مقاله های [۴] و [۶] هسته ی اصلی مطالب این پایان نامه را تشکیل می دهند و در کنار آن منابع مراجع مختلف دیگری نیز مورد استفاده قرار گرفته اند که به رسم امانت داری، نام آنها در قسمت مراجع ذکر گردیده است.

---

[۴] Cavender
[۵] Felsenstein
[۶] Allman
[۷] Rhodes

فصل ۱
پیش نیازها

به منظور پرهیز از مراجعه به منابع مختلف برای اطلاع از تعریف برخی مفاهیم به کار رفته در فصل های آتی، و همچنین مشخص کردن منظور خود از برخی مفاهیم، آنها را در این بخش معرفی و برخی خواص آنها را به اثبات می رسانیم.

۱.۱ مفاهیمی از ماتریس ها

فرض کنیم K یک میدان و m, n دو عدد صحیح مثبت باشند. در اینصورت خانواده ی تمام ماتریس های $m \times n$ با درایه های در K را می توان به صورتی کلی، با یک نماینده ی عمومی

$$\begin{pmatrix} x_{۱۱} & x_{۱۲} & \dots & x_{۱n} \\ x_{۲۱} & x_{۲۲} & \dots & x_{۲n} \\ \vdots & \vdots & & \vdots \\ x_{m۱} & x_{m۲} & \dots & x_{mn} \end{pmatrix}$$

که x_{ij} ها متغیرهایی هستند که در K تغییر می کنند، نمایش داد. به این ترتیب این ماتریس را یک ماتریس عمومی $m \times n$ می نامند.

گزاره ۱.۱.۱. *ایده آل تولید شده توسط کهادهای $k \times k$ یک ماتریس عمومی اول است.* [۱۵، نتیجه ۴.۱۱]

تعریف ۲.۱.۱. فرض کنیم d و α دو عدد صحیح مثبت و $d > \alpha$ باشد، در اینصورت ماتریسی به صورت

$$\begin{pmatrix} x_0 & x_1 & \dots & x_{d-\alpha} \\ x_1 & x_2 & \dots & x_{d+1-\alpha} \\ \vdots & \vdots & & \vdots \\ x_\alpha & x_{\alpha+1} & \dots & x_d \end{pmatrix}$$

را که از حذف عنصر اول سمت چپ سطر $i-$ام و افزودن متغیر $x_{d+i-\alpha}$به انتهای سمت راست آن ساخته می شود، را ماتریس کاتالکتیکنت می نامند. از این ماتریس تحت عنوان ماتریس هانکل[۱] نیز نام برده می شود.

۲.۱ مفاهیمی از گرافها

تعریف ۱.۲.۱. دوتایی $G = (V, E)$ که V مجموعه ای ناتهی و متناهی از عناصری به نام رئوس و E گردایه ای از زیرمجموعه های دوعضوی و متمایز V به نام یالهاست، را گراف ساده می نامیم.

چون موضوع اصلی این پایان نامه ارائه ی روشی برای شناخت اجداد و گونه های مشابه می باشد، و چون موجودات و گونه های زنده بر اثر مرور زمان تغییر می کنند و به شاخه های جداگانه ای گونه زایی می کنند، این فرایند تحول را می توان به صورت یک درخت نشان داد. به همین دلیل درخت ها نقشی اساسی در مدل سازی فرایند تحول دارند، لذا لازم است به صورتی صریح معرفی شوند.

تعریف ۲.۲.۱. یک گراف همبند بدون دور را یک درخت می نامند. اگر T یک درخت باشد که تمام رئوس داخلی آن از درجه ی ۲ است، آن را مسیر می نامند.

تعریف ۳.۲.۱. یک رأس درخت T که درجه ی آن یک می باشد را یک برگ درخت می نامیم. اگر L مجموعه ی تمام برگهای یک درخت باشد، $\tilde{V} = V - L$ را مجموعه ی رأس های داخلی

[۱] Hankel

درخت می نامیم. در حوزه ی زیست شناسی عناصر L را رأس های آشکار یا قابل مشاهده و رأس های $\tilde{V}$ را رأس های پنهان یا غیر قابل مشاهده می نامند.

قضیه ۴.۲.۱. *فرض کنید $G = (V, E)$ یک گراف باشد، آنگاه شرایط زیر هم ارزند:*

۱. *G یک درخت است*

۲. *برای هر دو رأس u و v در V، یک مسیر یکتا از u به v در G وجود داشته باشد*

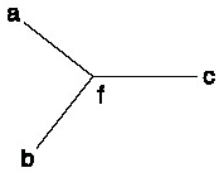
۳. *G همبند است و $|V| = |E| + ۱$.*

قرارداد ۵.۲.۱. *منظور ما از زیر جنگل یک درخت، زیر گرافی از درخت است که تمام برگ های آن، برگ های درخت اصلی هستند.*

درخت ها نشان دهنده ی فرایندهای زیستی هستند و این فرایندها ندرتا بر هم منطبق می شوند، همچنین هیچ یک از رأس های درونی درجه شان از ۳ تجاوز نمی کند. برای این درخت ها نام خاصی را انتخاب می کنیم و آن را تحت عنوان تعریف زیر بیان می کنیم.

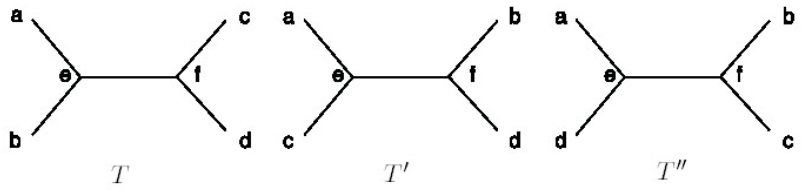
تعریف ۶.۲.۱. یک درخت را دودویی می نامیم، هر گاه هر رأس درونی آن از درجه ی ۳ باشد.

مثال ۷.۲.۱. با فرض سه ظرفیتی بودن رأس های درونی یک درخت ۳-برگی فقط می تواند به شکل زیر باشد،



شکل ۱.۱: درخت ۳-برگی سه ظرفیتی

و درختی با ۴-برگ فقط می تواند یکی از حالات زیر را بپذیرد.



شکل ۲.۱: درخت های ۴-برگی سه ظرفیتی

چون گونه زایی از یک گونه شروع می شود، و چون هدف از مطالعه ی درخت های فیلوژنتیک یافتن جد مشترک گونه های مورد نظر است، بنابراین بایستی در درخت های زیستی، رأسی را به این جد نسبت دهیم. این رأس را به صورت زیر تعریف می کنیم.

تعریف ۸.۲.۱. یک درخت ریشه دار درختی است که دقیقا یک رأس متمایز داشته باشد، این رأس متمایز را ریشه نامیده و با r نمایش می دهیم. ریشه ی r نشان دهنده ی جدیدترین جد مشترک برگ های موجود است و سایر گره های داخلی درخت، نشان دهنده ی جد مشترک برگهایی است که با آن گره از ریشه جدا می شوند. معمولا یک درخت ریشه دار را چنان رسم می کنیم که ریشه در بالای بقیه ی رأس ها قرار گیرد.

قضیه ۹.۲.۱. *اگر T یک درخت ریشه دار دودویی با n برگ باشد، آنگاه* $|E| = ۲n - ۲$.

برهان. به وضوح حکم برای یک درخت ریشه دار دودویی با ۲ برگ برقرار است. حال فرض می کنیم T یک درخت ریشه دار دودویی با $(n-۱)$ برگ باشد، که $۲(n-۱) - ۲ = ۲n - ۴$ یال دارد، ثابت می کنیم حکم برای درخت T' با n برگ نیز برقرار است.
فرض کنیم $a_۱$ یک برگ درخت T' و $v_۱$ رأس مجاور $a_۱$ باشد. همچنین فرض کنیم که $a_۲$ رأس دیگر یالی باشد که از $v_۱$ خارج می شود. چون بنابر فرض T' دودویی است، پس تنها دو یال از $v_۱$ خارج می شود، یا به عبارت دیگر تنها دو برگ در مجاورت $v_۱$ قراردارند.
با حذف دو برگ $a_۱$ و $a_۲$ از گراف T' یک درخت $n-۱$ برگی مانند T به دست می آید. اما بنابر فرض درخت T دارای $۲n-۴$ یال می باشد، و چون دو درخت T و T' تنها در دو یال با هم اختلاف دارند، پس تعداد یالهای درخت T' که n برگ دارد برابر است با: $(۲n-۴) + ۲ = ۲n - ۲$.

بنابراین به استقرا ثابت شد که تعداد یالهای یک درخت ریشه دار دودویی با n برگ برابر است با: $|E| = ۲n - ۲$. □

تعریف ۱۰.۲.۱. یک درخت فیلوژنتیک T درختی است که برگهای آن برچسب خورده باشند. اگر هر رأس داخلی T از درجه ی ۳ باشد، T را یک درخت فیلوژنتیک دودویی می نامیم، و هرگاه درخت T ریشه دار باشد، آنرا یک درخت فیلوژنتیک ریشه دار می نامیم. به همین ترتیب هرگاه T درختی ریشه دار و با رأس های داخلی از درجه ی ۳ باشد، آنرا درخت ریشه دار دودویی می نامیم.

قضیه ۱۱.۲.۱. *فرض کنید T یک درخت فیلوژنتیک با n برگ باشد، آنگاه برای هر $n \geq ۲$، درخت T دارای $۲n - ۳$ یال می باشد که $n - ۳$ یال آن درونی است.*

برهان. مشابه قضیه ی ۹.۲.۱، با استقرا روی $n \geq ۲$ حکم ثابت می شود. □

۳.۱ فرایند مارکوف

چون نوکلئوتیدها در دنباله های DNA، به هنگام تکثیر سلولی ممکن است تغییر کنند و گونه ای جدید ایجاد کنند، طبیعی است که این فرایند را در قالب فرایندهای تصادفی مطالعه نماییم. به همین دلیل لازم است مطالبی درباره ی فرایند های تصادفی در اینجا نقل نماییم. اما چون مطالعه ی خود را به فرایند های مارکوف محدود کرده ایم، تنها تعریف این فرایند را یادآوری می کنیم.

قرارداد ۱.۳.۱. *از این پس مجموعه ی $\{۱, ۲, \ldots, k\}$ را با $[k]$ نمایش می دهیم.*

تعریف ۲.۳.۱. فرض کنیم $X_۱, \ldots, X_n$ متغیرهای تصادفی با فضای نمونه ی S باشند که مقادیر خود را از یک مجموعه ی U اختیار می کنند، و فرض کنیم $A = \{۱, \ldots, t\}$باشد. برای یک زیر مجموعه ی $B \subset A$ و یک پیشامد E از S، احتمال شرطی وقوع E به شرط وقوع $\{X_i = u_i | i \in B\}$ را با $Prob(E|X_{i_۱} = u_{i_۱}, \ldots, X_{i_s} = u_{i_s}; i_j \in B)$ نمایش می دهیم.

تعریف ۳.۳.۱. فرض کنیم T یک درخت ریشه دار با مجموعه ی رئوس V باشد. یک فرایند مارکوف روی T با مجموعه حالت $[k]$، یک خانواده ی $\{X_v | v \in V\}$ از متغیرهای تصادفی است،

به طوری که وقتی u و v دو رأس T و $P : u_{i_1} = u, \ldots, u_{i_l} = v$ یک مسیر از u به v باشد، آنگاه

$$Prob(X_v = \alpha | X_{i_1} = u_1, X_{i_2} = u_2, \ldots, X_{i_l} = u_l) = Prob(X_v = \alpha | X_{i_1} = u_1 = u)$$

رابطه ی اخیر به خاصیت مارکوف موسوم است.

به عبارت دیگر، در یک فرایند مارکوف، برای هر مسیر (u, v) روی T، احتمال وقوع پیشامد α برای X_v، مستقل از وقوع پیشامدهای $X_{i_2} = u_2, \ldots, X_{i_l} = u_l$، روی مسیر u به v است و فقط به اولین پیشامد، $X_{i_1} = u_1$، بستگی دارد.

برای هر یال $e = \{u, v\}$ روی T، با فرض مقدم بودن u بر v، با مجموعه حالت $[k]$، یک ماتریس $k \times k$ موسوم به ماتریس گذر القا می کند، که آن را با M_e نمایش می دهیم. درایه ی (i, j)- ام این ماتریس، احتمال گذار حالت i روی u به حالت j روی v را نشان می دهد. روشن است که درایه های این ماتریس ها نامنفی هستند و مجموع درایه های هر سطرآن یک است. به عبارت دیگر برای هر یال e از درخت T^r داریم:

$$M_e = \begin{pmatrix} e_{11} & e_{12} & \ldots & e_{1k} \\ e_{21} & e_{22} & \ldots & e_{2k} \\ \vdots & & & \\ e_{k1} & e_{k2} & \ldots & e_{kk} \end{pmatrix}$$

که برای هر $1 \leq i, j \leq k$ بایستی: $e_{ij} \geq 0$ و $e_{i1} + e_{i2} + \ldots + e_{ik} = 1$.

بنابراین اگر برای هر یال یک درخت، یک ماتریس مارکوف تعیین کنیم، چگونگی تحول کل فرایند که در طول درخت صورت می گیرد را تعیین کرده ایم.

به علاوه یک تابع توزیع برای ریشه به صورت زیر تعریف می کنیم: $\pi_r = (\pi_1, \ldots, \pi_k)$ که π_i احتمال وقوع حالت i در ریشه است، و بدیهی است که $\pi_i \geq 0$ و $\sum_{i=1}^{k} \pi_i = 1$.

تعریف ۴.۳.۱. اگر به هر رأس درخت T یک متغیر تصادفی نظیر کرده باشیم، که این متغیرها یک فرایند مارکوف باشند، آنگاه گوییم یک مدل مارکوف عمومی روی T تعریف کرده ایم. معمولا این مدل را به صورت $(\pi(r), \{M_e\}_{e \in E})$ نشان می دهیم، و برای تمیز دادن آنها از پارامترهای درخت، آنها را پارامترهای تصادفی می نامیم.

ملاحظه ۵.۳.۱. *برای ملکول DNA، تعداد حالتها $k = ۴$ است، اما برای دنباله های پروتئین، که از ۲۰ اسید آمینه ساخته شده است، $k = ۲۰$ می باشد. برای مولکول های DNA، $k = ۲$ نیز قابل قبول است، زیرا می توان پیورین ها را با $R = \{A, G\}$ و پیرامیدین ها را با $Y = \{C, T\}$ دسته بندی کرد.*

حال فرض کنیم $a_۱, \ldots, a_n$ برگ های یک درخت T باشد. تحول در طول درخت T روی می دهد، اما ما فقط می توانیم دنباله هایی که در برگها ظاهر می شوند را مشاهده کنیم. ما علاقه مند هستیم، با مدل مارکوفی که پارامترهای آن را مشخص کردیم، توزیع توأم حالتهایی که در برگهای a_i روی می دهد را بیابیم.
واضح است که توزیع توأم P، یک جدول با آرایه n بعدی $k \times k \times \cdots \times k$ است که درایه های آن برابر

$$P(i_۱, \ldots, i_n) = Prob(a_۱ = i_۱, \ldots, a_n = i_n)$$

می باشد. $P(i_۱, \ldots, i_n)$ نشان دهنده ی احتمال وقوع حالت i_j در برگ a_j، برای $j = ۱, ۲ \ldots, n$ می باشد. به طور کلی برای سادگی در نمادگذاری، $P(i_۱, \ldots, i_n)$ را با نماد $p_{i_۱,\ldots,i_n}$ نشان می دهیم. درایه های P، فراوانی های مورد انتظار برای مشاهده ی تمامی پیشامدهای ممکن در برگ ها را به دست می دهد. جدول P را گاه یک $k \times k \times \cdots \times k$ تانسور می نامیم.

۴.۱ مفاهیمی از جبر

تعریف ۱.۴.۱. فرض کنید G یک گروه و Ω مجموعه ای غیر تهی است، و همچنین فرض کنید عضو بی اثر G را با ۱ نمایش می دهیم. نگاشت $f : \Omega \times G \longrightarrow \Omega$ را یک عمل G روی Ω می نامیم، هرگاه

$$(i) \quad f(\omega, ۱) = \omega \qquad \forall \omega \in \Omega$$

$$(ii) \quad f(f(\omega, g), h) = f(\omega, gh) \qquad \forall \omega \in \Omega, \ \forall g, h \in G$$