"همانند سازی DNA"

همانند سازی ماده ژنتیک به طور دقیق صورت میگیرد، خواص سلول های دختر پس از تقسیم سلول مادر، مشابه یکدیگر و مشابه خواص سلول مادر میباشند، دقیق بودن همانندسازی ماده ژنتیک در سطح تولید مثل جاندار نیز مشهود میباشد. این که زاده های اسب، اسب هستند و نه جاندار دیگر، تکثیر دقیق ماده ژنتیک را میرساند.

ماده ژنتیک و بالطبع تغییرات ناشی از اشتباه در همانندسازی آن به ارث میرسد، یعنی تغییرات در زاده های سلولی که اشتباه در آن رخ داده است، باقی خواهند ماند. چنانچه سلول مورد نظر سلول جنسی باشد، اشتباه در نسل های بعدی آن موجود زنده حفظ خواهد شد.

مسیرهای گوناگونی برای حفظ ماده ژنتیک و جلوگیری از هر گونه تغییر کمی و کیفی در آن، تکامل یافته اند. از مشخص ترین این مسیرها تقسیم میتوز و میوز است که در جریان این تقسیم ها کروموزوم ها بدون هیچ کم و کاستی بین سلول های حاصل از تقسیم توزیع میشوند.

۱. انواع DNA پلیمرازها

سه نوع DNA پلیمراز در سلول های باکتری وجـود دارد، DNA پلیمـراز3،DNA پلیمرازO ، DNA پلیمراز Σ . DNA پلیمـراز Σ مسـؤلیت اصـلی DNA سـازی، یعنـی DNA سازی به منظور تهیه نسخه های ژنومی دختر را به عهده دارد. اگر از فعالیـت ایـن آنزیم به نحوی جلوگیری شود، مثلاً ژن آن جهش یابد یا از عامـل بازدارنـده اختصاصـی برای فعالیت آن آنزیم استفاده شود، ساخته شدن DNA دختر انجـام نمیگیـرد. جهـش در ژن پلیمرازهای دیگر یا جلوگیری از فعالیت آنها چنین تأثیری را ندارد.

DNA پلیمراز3 نقش ثانوی در همانندسازی DNA دارد اما نقش اصلی آن در ترمیم DNA است که در ترمیم، DNA سازی به مقدار کم انجام میشود. نقش آنزیم DNA پلیمرازO در سلول هنوز مشخص نشده است.

هر سه آنزیم DNA پلیمراز، DNA سازی را از سمت'۵ ملکول به سمت'۳ آن کاتالیز میکنند، به این معنی که ساخت رشته DNA از سمت '۵ به سمت '۳ پیش میرود و از این رو سمت '۵ قبل از سمت '۳ ساخته میشود.

۲. فعالیت نوکلئازی DNA پلیمرازها و فرآیند ویرایش

آنزیم DNA پلیمراز علاوه بر فعالیت DNA سازی، توانایی تجزیه DNA را نیز دارد. DNA پلیمراز 3 میتواند DNA را از سمت '۵ به'۳ ، و هر سه DNA پلیمراز پروکاریوتی قادرند DNA را از سمت '۳ به '۵ تجزیه کنند. فعالیت تجزیه DNA ، فعالیت نوکلئازی نامیده میشود و آنزیم هایی که این واکنش را کاتالیز میکنند نوکلئاز هستند. فعالیت نوکلئازی '۳ به '۵ به ویرایش معروف است و هر سه پلیمراز پروکاریوتی (موجودات غیر پیشرفته که DNA در غشای هسته قرار ندارند.) قادر به انجام آن هستند.

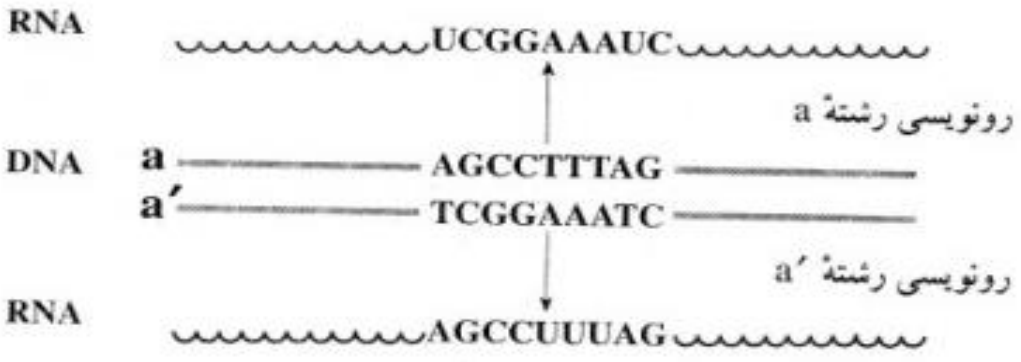
برای ساختن رشته جدید DNA ، آنزیم DNA پلیمراز نوکلئوتیدها را طوری کنار یکدیگر قرار میدهد که رابطه مکملی با نوکلئوتیدهای رشته الگو داشته باشد (رابطه مکملی یعنی مثلاً اگر توالی نوکلئوتیدها در قسمتی از یک رشته'۳-GGCATTAC-'۵ باشد، ترتیب در رشته دیگر الزاماً '۵-CCGTAATG-'۳ خواهد بود.) برای اضافه شدن هر نوکلئوتید ، انتهای'۳ نوکلئوتیدی که در رابطه مکملی (A/T یا G/C) با نوکلئوتید الگو باشد، لازم است . اضافه شدن نوکلئوتید "غلط" به رشته جدید ضمن DNA سازی نادر است اما گاهی رخ میدهد ، در چنین حالتی ، هیچ DNA پلیمراز پروکاریوتی نمیتواند نوکلئوتید بعدی را به این نوکلئوتید غلط متصل کند. در برخورد با این نوکلئوتید اشتباه ، مثلاً انتهای C که با A در الگو جفت شده باشد ، آنزیم پلیمراز غلط بودن رابطه مکملی را

تشخیص میدهد و سپس مراحلی طی میشود تا رابطه مکملی بـه صـورت درسـت انجـام گیرد ، ویرایش یکی از عوامل مؤثر در پایین نگه داشـتن میـزان اشـتباه در همانندسـازی است.

" رونویسی "

پروتئین ها اکثر فعالیت های سلولی را انجام میدهند و DNA اطلاعـات لازم بـرای ساختن پروتئین ها را رمز میکند ، اما DNA مستقیماً برای ساختن پـروتئین هـا بـه کـار نمیرود بلکه RNA ملکول میـانجی میباشـد و مسـتقیماً اطلاعـات را از DNA دریافـت میکند، ساختن RNA از روی DNA رونویسی نامیـده میشـود و یکـی از مراحـل و در حقیقت اولین مرحله تجلی اطلاعات ژنتیک میباشد.

معمولاً RNA فقط از روی یک رشته DNA رونویسی میشود به این معنـی کـه در برخی مناطق ژنوم ، یکی از دو رشته DNA برای ساخته شـدن RNA الگـو اسـت و در مناطق دیگر ، رشته دیگر DNA الگو میشود ولی در یک منطقه هر دو رشته DNA الگو نیست.



۱.مراحل رونویسی

سلول های پروکاریوتی یک نوع RNA پلیمراز (آنزیم RNA ساز) دارند که انواع RNA سلول را از روی DNA رونویسی میکند. واکنش رونویسی را که این آنزیم و در حقیقت تمام RNA های پلیمراز انجام میدهند ، خلاصه میشود در خواندن رشته DNA الگو از جهت '۳ به '۵ و ساختن RNA مکمل آن از جهت '۵ به '۳ .

رونویسی معمولاً در سه مرحله آغاز ، ادامه و پایان بررسی میشود . ساختار آنزیم RNA پلیمراز در این سه مرحله تغییرات جزئی پیدا میکند . مرحله آغاز رونویسی شامل شناسایی محل صحیح شروع RNA سازی تا تشکیل چند پیوند اولیه فسفودی استر در RNA است. هنگام آغاز RNA سازی ، دو رشته DNA در محل آغاز از یکدیگر جدا میشوند تا یکی از رشته های DNA نقش الگو را ایفا کند و ریبونوکلئوتیدهای ملکول RNA با دِاُکسی ریبونوکلئوتیدهای رشته الگو پیوند مکملی تشکیل دهند.

مرحله ادامه رونویسی شامل تکرار تشکیل پیوند فسفودی استر بین نوکلئوتید جدید و انتهای '۳ RNA است. در این مرحله ، منطقه باز شده DNA در جهت پیشرفت رونویسی حرکت میکند به نحوی که بخش جدیدی از DNA در جلو باز ، و منطقه ای در عقب مجدداً دو رشته ای میشود . ضمناً پیوندهای جدید بین ریبونوکلئوتیدها و DNA در جلو تشکیل و پیوندهای قبلی در عقب باز میشوند . افزوده شدن ریبونوکلئوتیدها تا رسیدن به محل پایان رونویسی تکرار میشود . در محل پایان رونویسی ، RNA کامل رها و آنزیم از DNA جدا میشود .

۲.راه اندازها

یک تعریف برای راه انداز این است که راه انداز قسمتی از ژن است که امکان شروع ساختن RNA مربوط به آن ژن یا رونوشت آن ژن را در جایگاه صحیح فراهم میسازد . اگر این قسمت وجود نمیداشت ممکن بود رونویسی از مکان نادرست ، مثلاً از وسط یک

ژن شروع شود و در داخل ژن مجاور به اتمام برسد . در این حالت ، محصول رونویسی کارا نخواهد بود . بخش مهم رویداد آغاز RNA سازی اتصال RNA پلیمراز به DNA است و از این رو جایگاه اتصال آنزیم میتواند تعریف دیگری برای راه انداز باشد .

۳. پایان رونویسی

در نزدیکی انتهای '۳ RNA ای که رونویسی شده است ، ساختاری مشابه سنجاق سر تشکیل میشود که در پایان دادن به رونویسی نقش دارد . این ساختار حاصل تشکیل پیوندهای هیدروژنی بین ریبونوکلئوتیدها در دو بخش ملکول RNA رونوشت است. ریبونوکلئوتیدهایی که در پیوندهای هیدروژنی شرکت دارند ساقه سنجاق را به وجود میاورند ، و توالی بین دو بخش که در رابطه مکملی شرکت ندارند به شکل حلقه تک رشته ای در میاید . معمولاً بعد از این ساختار ، ردیفی از ریبونوکلئوتیدهای U در انتهای 'RNA۳ ها وجود دارد . این ساختار سنجاق سر مانعی برای ادامه حرکت آنزیم RNA پلیمراز به حساب میاید . منطقه ای در DNA که توالی ساختار سنجاق سر و ردیف U را رمز میکند ، جایگاه پایان نامیده میشود . ساختار سنجاق سر و ردیف U هر دو برای پایان رونویسی مهم هستند ، زیرا جهش هایی که هر یک از اینها را تغییر میدهند در پایان رونویسی اختلال به وجود میاورند .

"ترجمه"

ترجمه روندی است که در آن از اطلاعات رمز شده در RNA برای ساختن پروتئینها استفاده میشود . ترجمه در این جا کلمه مناسبی است ، زیرا "حروف" RNA ، نوکلئوتیدها و "حروف" پروتئین ، اسیدهای آمینه هستند و بنابراین طی فرآیند ترجمه "زبان" تغییر میابد.

درباره ترجمه RNA باید در نظر داشت که نـوعی "ملکـول تطبیقـی" بایـد وجـود داشته باشد که همچون پلی بین اسید آمینه و RNA عمل میکند ، ایـن ملکـول تطبیقـی ، آنزیمی است که اتصال اسیدهای آمینه را به ملکولهای اسـید نوکلئیـک کـوچکی کاتـالیز میکند و از نوع RNA میباشد کـه بـه آن "RNA ناقـل" یـا tRNA (transfer RNA) میگویند. آنزیمی که اتصال اسید آمینه به RNA را کاتالیز میکند آمینواَسـیلtRNA سـنتاز است.

۱. mRNA های پروکاریوتی

RNA های پیک یا mRNA ها مولکول هایی هستند که اطلاعات DNA را بـرای ساختن پروتئین ها دریافت میکنند و در پروتئین سازی نقش الگو دارند. جهـت خوانـدن mRNA در پروتئین سازی '۵ به '۳ است یعنی انتهای RNA که اول سـاخته میشـود اول هم ترجمه میشود . تمام طول mRNA های پروکاریوتی ترجمه نمیشـود. جایگـاه آغـاز ترجمه به طور متوسط ۱۰۰ نوکلئوتید از انتهای '۵ فاصله دارد و جایگاه پایان ترجمه نیـز در جایی نرسیده به انتهای '۳ مولکول mRNA قرار دارد. توالی رمزگـردان RNA پیـک بین جایگاه آغاز و پایان ترجمه قرار دارد.

mRNA های پروکاریوتی تک ژنی یا چند ژنی هستند. mRNA های تک ژنـی از روی یک ژن رونویسی میشوند و الگوی ساختن تنها یک پروتئین هستند. mRNA هـای چند ژنی ، ویژه سلول های پروکاریوتی و مظهری از نحوه تنظـیم تجلـی ژن هـا در ایـن سلول ها هستند که بعداً توضیح داده خواهد شد. هر یک از RNA های چند ژنی از روی دو یا چند ژن مجاور رونویسی میشود و بنابراین الگـوی سـاختن همـان تعـداد پـروتئین است. فاصله بین دو ژن متوالی در این واحدها نیز رونویسی میشود و در بخش بـین ژنـی RNA های پیک ظاهر میشود.

۲. mRNA های یوکاریوتی

mRNA های یوکاریوتی (یوکاریوتها موجودات پیشرفته مثل انسان و حیوان و بیشتر گیاهان هستند که DNA آنها در غشای هسته قرار دارد.) با mRNA های پروکاریوتی تفاوت بسیار دارند در درجه اول ، mRNA های یوکاریوتی همیشه تک ژنی هستند. تفاوت های دیگر در این خلاصه میشوند که محصولات اولیه رونویسی ژن های یوکاریوتی، پیش سازهای RNA های پیک هستند و برای تبدیل آنها به mRNA بالغ پردازش های گوناگون لازم است.

۳.پروتئین سازی

پروتئین سازی را همچون همانند سازی و رونویسی میتوان به سه مرحله آغاز ، ادامه و پایان تقسیم کرد. کلیه واکنش ها در سیستمهای پروکاریوتی و یوکاریوتی مشابه میباشند.ملکول ها و ساختارهایی که در مراحل مختلف پروتئین سازی شرکت دارند عبارتند از RNA پیک ، tRNA ها، انواع عوامل ترجمه و ریبوزوم . عوامل ترجمه پروتئین هایی هستند که هر یک در مرحله ویژه ای از پروتئین سازی فعالیت دارند. ریبوزوم ها کارخانه ایی هستند که پروتئین سازی در آنها انجام میگیرد. ریبوزوم ها ساختارهای بسیار بزرگی هستند که در پروکاریوت ها از حدود ۵۰ نوع پروتئین و سه نوع ملکول rRNA (RNA ریبوزومی) و در یوکاریوت ها از حدود ۸۰ نوع پروتئین و ۴ نوع rRNA تشکیل میشوند .

پروتئین سازی در پروکاریوت ها و یوکاریوت ها در انتهای '۵ ملکول mRNA شروع نمیشود ، بلکه در جایگاهی درونی تر در ملکول ، شروع میشود . با توجه به این که رمزها در RNA پیک پشت سر هم خوانده میشوند یعنی نوکلئوتیدی بین دو رمز متوالی وجود ندارد ، رمزهای پی در پی بسته به این که اولین رمز کجا شروع میشود ، متفاوت خواهند بود . برای هر mRNA ، سه چارچوب خواندن را میتوان تصور کرد که

هر یک توالی خواصی از اسیدهای آمینه را رمز میکند . در عمل فقط یکـی از چارچوبهـا استفاده میشود و این چارچوب صحیح را جایگاه رمز آغازین تعیین میکند .

"جهش"

جهش یعنی تغییر در توالی نوکلئوتیدها درDNA ، برخی جهـش هـا بـر فعالیـت محصول ژن جهش یافته تأثیر میگذارند. به عنوان مثال ممکـن اسـت جهـش در فعالیـت آنزیمی که محصول ژن جهش یافته اسـت اخـتلال بـه وجـود آورد . جهـش از دیـدگاه تکاملی اهمیت بسزایی دارد زیرا در نتیجه آن تنوع به وجود میاید و عملکرد تکامـل بـه وجود تنوع در جوامع گونه ها وابسته است . جهش تصادفی است به این مفهوم که احتمال تغییر در تمام نوکلئوتیدهای ژنوم، کم و بیش یکسان است . جهش رویدادی شیمیایی است که بدون توجه به تأثیر بالفعل مثبت یا منفی آن بر جانـدار رخ میدهـد . بـر اثـر انتخـاب طبیعی، جهش یافته های زیانبار کمتر تولید مثل میکنند یا اصلاً تولید مثل نمیکننـد و لـذا تعداد زاده هایی که در این جهش ها باشند به نسبت در نسل های بعـد کـم خواهـد بـود . جهش یافته هایی با جهش سودمند بیشتر تولید مثل میکنند و در نسل های بعد فراوان تـر میشوند . احتمال این که جهش در ژن یا ژنوم خاصی رخ دهد ، به تعداد نوکلئوتیدهای آن ژن یا ژنوم بستگی دارد . برای موجودات زنده ای که از راه جنسی تولید مثل میکنند ، تنها جهش هایی که در سلول های جنسی رخ داده هند به نسل های بعـد میرسـند. جهـش در سلول غیر جنسی فقط در دودمان همان سلول در فرد جهش یافته تأثیر میگذارد.

۱. جهش های خود به خود و جهش های القایی

یک نوع تقسیم بندی جهش ها بر حسب خـود بـه خـود بـودن یـا القـایی بـودن آنهاست. جهش های خود به خود بر اثر عوامـل و شـرایط روزمـره زنـدگی جانـدار رخ

میدهند ، مثلاً بر اثر کیفیت عملکرد DNA پلیمراز در همانند سازی DNA یا تابش هـای فرابنفش که از خورشید به زمین میرسند .

جهش های القایی در اثر به کارگیری مواد جهش زا به وجود میایند. مواد جهـش زا باعث افزایش مقدار جهش از میزان زمینه ای میشوند . بعضی از مواد شیمیایی ، تابشـهای فرابنفش و پرتو ایکس از عوامل جهش زا هستند . جهش چه خود به خود صورت گیـرد و چه القا شود ، همواره به علت واکنش های شیمیایی و فیزیکی رخ میدهد .

۲.جهش های بزرگ و جهش های کوچک

جهش های بزرگ جهش هایی هستند که با میکروسکوب نوری میتوان تأثیر آنها را بر ماده ژنتیک تشخیص داد. این جهش ها باعث تغییر در تعداد یا شـکل کرومـوزوم هـا میشوند. سایر جهش ها، جهش های کوچک به شمار میایند. هر دو نوع جهش بر فنوتیپ (شکل ظاهری موجودات) تأثیر میگذارند اما تشخیص تغییر ایجاد شده در ماده ژنتیـک در جهش های بزرگ راحت تر است. روند ایجاد هر دو نـوع جهـش اساسـی مولکـولی دارد.

۳.جهش های ساختاری و جهش های تنظیمی

جهش های ساختاری آن دسته از جهـش هـا هسـتند کـه باعـث تغییـر در تـوالی اسیدهای آمینه در ملکول پروتئین میشوند. جهـش هـایی کـه تـوالی نوکلئوتیـدها را در RNAهایی که ترجمه نمیشوند ، تغییر میدهند ماننـد rRNA وtRNA نیـز جهـش هـای ساختاری هستند . جهش های تنظیمی توالی نوکلئوتیدها را در بخش های تنظیمی ماننـد راه اندازها و جایگاه اتصال ریبوزوم در mRNA تغییر میدهند . این جهش ها بـر میـزان تجلی تأثیر میگذارند ، به این معنی که تجلی ژن مربوط را زیاد، کم یا ناممکن میسـازند . این گونه جهش ها در برخی بیماری ها شناخته شده اند.

"جزیره های CpG " "CpG islands"

در ژن انسان زمانی که دینوکلئوتید CG دیده میشود (که اغلب به صورت CpG نشان داده میشود تا از باز C-G که بعد از دو رشته DNA قرار میگیرد متمایز شود) نوکلئوتید C (سیتوزین) از نظر شیمیایی با متیلاسیون شناخته میشود (یعنی در جزیره های CpG آنچه که متیله میشود نوکلئوتید C است.). احتمال نسبتاً زیادی برای این که متیل C به T جهش پیدا کند وجود دارد ، در نتیجه در ژن احتمال وجود جزیره های CpG از آنچه با توجه به احتمال های مستقل C وG بدست میاید کمتر میباشد. (یعنی از حد انتظار ما کمتر است.) با توجه به دلایل با اهمیت زیست شناسی ، مرحله متیلاسیون در مقطع کوتاهی از ژن متوقف میشود، مثل نواحی شروع (یا promoter ها) در بسیاری از ژنها . در این نواحی CpG دینوکلئوتیدهای بیشتری نسبت به نواحی دیگر دیده میشود، یعنی در واقع نوکلئوتیدهای C و G بیشتر دیده میشوند . به این نواحی جزیره های CpG گفته میشود ، که عموماً طولی به اندازه چند صد باز تا چند هزار باز دارند .

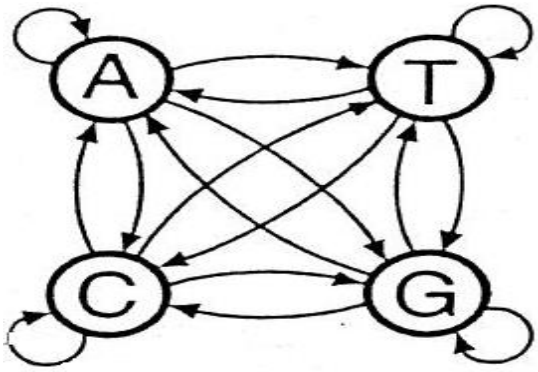
در این بحث ما به دو سئوال توجه میکنیم : اولاً : اگر طول کوتاهی از یک رشته ژن داشته باشیم چگونه پی میبریم که به جزیره های CpG تعلق دارد یا نه؟ و ثانیاً : اگر تکه بلندی از یک رشته ژن داشته باشیم ، چگونه میتوانیم اگر دارای جزیره های CpG باشد آن را بیابیم ؟ ما بحث خود را با سئوال اول شروع میکنیم.

" زنجیرهای مارکف "

میخواهیم بدانیم چه نوعی از مدلهای احتمالی را میتوان برای نواحی جزیره های CpG استفاده کرد:

میدانیم که دینوکلئوتیدها عوامل مهمی هستند و همچنین احتیاج به مدلی داریم که رشته هایی تولید کند که در آن احتمال یک وضعیت به وضعیت ماقبلش بستگی داشته باشد. ساده ترین مدل در این حالت زنجیرهای مارکف میباشند . میتوان زنجیر مارکف را به صورت موقعیت هایی نشان داد که هر کدام با فلشهایی به یکدیگر متصل شده اند که بینگر یک عدد میباشند .

مثلاً یک زنجیر مارکف برای DNA میتواند به صورت زیر باشد:



در شکل بالا برای هر چهار حرف A,C,G,T در حروف DNA یک موقعیت (وضعیت) در نظر گرفته شده است . پارامتر احتمال مربوط به فلش هایی است که در شکل کشیده شده است به این صورت که هر فلش بیانگر احتمال رخ دادن یک وضعیت بعد از وضعیت قبلی میباشد. این احتمال ها ، احتمال های انتقال نامیده میشوند که به صورت a_{st} داده میشوند :

$$a_{st} = P(x_i = t \mid x_{i-1} = s) \qquad (۱)$$

برای هر مدل احتمالی مربوط به رشته ها میتوانیم احتمال رخ دادن یک رشته را به صورت زیر نشان دهیم :

$$P(x) = P(x_L, x_{L-1}, \ldots, x_1)$$

$$= P(x_L | x_{L-1}, \ldots, x_1) P(x_{L-1} | x_{l-2}, \ldots, x_1) \ldots P(x_1)$$

در رابطه قبل از فرمول $P(X,Y)=P(X|Y)P(Y)$ به دفعات استفاده شده است. خصوصیت اصلی زنجیر مارکف این است که احتمال رخ دادن هر وضعیت x_i فقط به موقعیت قبلی آن یعنی x_{i-1} بستگی دارد و نه به کل رشته قبل از آن ، با توجه به این مطلب میتوان نوشت : $P(x_i| x_{i-1},\dots x_1) = P(x_i| x_{i-1}) = a_{xi-1xi}$

و در نتیجه تساوی قبلی به صورت زیر نوشته میشود:

$$P(x)=P(x_L|x_{L-1})P(x_{L-1}|x_{l-2})\dots P(x_2|x_1)P(x_1)$$
$$=P(x_1)A^{L}_{i=2}\, a_{xi-1xi} \qquad (۲)$$

تساوی قبل برابر با یک میباشد.

" نمایش شروع و پایان رشته ها "

با توجه به این که احتمالهای انتقال را معرفی کردیم ، میخواهیم برای احتمال شروع یک رشته یعنی$P(x_1)$ یک نماد خاص معرفی کنیم . میتوانیم یک موقعیت شروع جدید به مدل اضافه کنیم و در این حالت حرف $\exists$ را برای آن تعریف کنیم . به این صورت که $x_0=\exists$ و به عنوان مثال احتمال این که وضعیت اول یعنی x_1 برابر با s باشد به صورت زیر تعریف میشود:

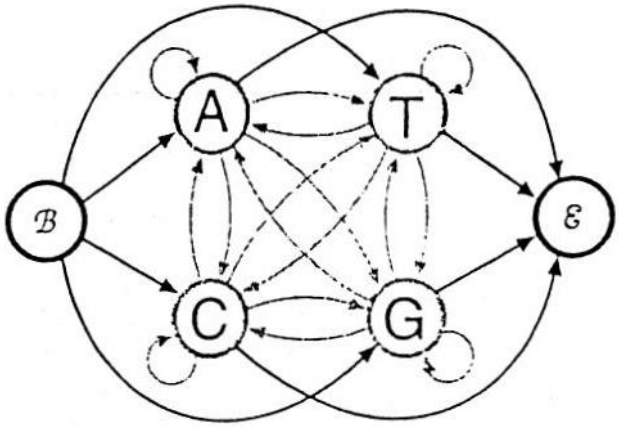
$$P(x_1= s) = a_{\exists s}$$

به طور مشابه میتوان نماد , را برای انتهای یک رشته در نظر گرفت تا مطمئن شویم که آخر رشته مدلسازی شده است. در این حالت احتمال پایان یافتن رشته به طوری که $x_L=t$ باشد برابر است با:

$$P(,|x_L= t) = a_{t,}$$

وضعیت های شروع و پایان تعریف شده را به مدل DNA اضافه میکنیم:

(شکل ۱)



وضعیت های شروع و پایان تعریف شده یعنی $\exists$ و, را موقعیت های خاموش (Silent States) مینامیم چون تنها برای نمایش موقعیت شروع و پایان میباشند و هیچ وظیفه دیگری بر عهده ندارند.

اصولاً انتهای یک رشته در زنجیر مارکف نمایش داده نمیشود، و این به این خاطر است که رشته میتواند هر کجا پایان یابد. دلیل این که وضعیت پایان (,) را به انتهای یک رشته اضافه میکنیم این است که بتوانیم توزیع تمام رشته های ممکن را با هر طولی بدست آوریم. این توزیع به صورت تصاعدی افزایش میابد .

مطلب مهم: فرض کنید مدل دارای وضعیت پایان باشد و احتمال انتقال از هر وضعیتی به وضعیت پایان برابر ϑ باشد، آنگاه مجموع احتمالهای تمام رشته ها با طول L (و به این صورت که رشته بعد از یک انتقال به وضعیت پایان، پایان یابد) برابر است با :

$$\vartheta(1-\vartheta)^{L-1}$$

مطلب مهم: مجموع احتمال های تمام رشته های ممکن با هر طولی برابر یک میباشد . این نشان میدهد که زنجیر مارکف یک توزیع احتمال مناسب با فضای وضعیتی شامل تمامی رشته های ممکن میباشد.

"استفاده از زنجیرهای مارکف برای تشخیص"

یک استفاده اولیه از تساوی (۲) این است که مقادیر آزمون نسبت احتمال (likelihood ratio test) را بدست آوریم. در این قسمت ما با توجه به مثال جزیره CpG و داده های واقعی این مطلب را توضیح میدهیم . از یک گروه از رشته های DNA انسانی به طور فرضی ۴۸ جزیره CpG بیرون کشیده ایم و از آنها تعداد دو زنجیر مارکف بدست آوردیم ، به طوری که یکی از آنها مربوط به نواحی جزیره CpG میباشد (مدل با علامت "+") و دیگری مابقی رشته میباشد (مدل با علامت "_"). در این حالت احتمالهای انتقال برای هر مدل به صورت زیر بدست میاید :

$$a^{+}_{st}=c^{+}_{st} / E_{t'}c^{+}_{st'} \qquad (۳)$$

به صورت مشابه a^{-}_{st} نیز بدست میاید . در تساوی بالا، c^{+}_{st} تعداد دفعاتی است که در ناحیه مورد بررسی (علامتگذاری شده) موقعیت t بعد از موقعیت s میاید . a^{+}_{st} ها برآورد کننده هایی با بیشترین احتمال (maximum likelihood estimators) یا ML برای احتمال های انتقال میباشند. (و a^{-}_{st} ها)

(در این مثال ۶۰۰۰۰ نوکلئوتید مورد بررسی قرار گرفته اند و روش ML در این جا مناسب میباشد ، اگر تعداد گونه های مورد بررسی کم باشد بهتر است از برآورد Bayesian استفاده کرد که در این جا توضیح داده نمیشود .) جدول نتایج با توجه به تساوی بالا عبارت است از :

+	A	C	G	T
A	۰/۱۸۰	۰/۲۷۴	۰/۴۲۶	۰/۱۲۰
C	۰/۱۷۱	۰/۳۶۸	۰/۲۷۴	۰/۱۸۸
G	۰/۱۶۱	۰/۳۳۹	۰/۳۷۵	۰/۱۲۵
T	۰/۰۷۹	۰/۳۵۵	۰/۳۸۴	۰/۱۸۲

_	A	C	G	T
A	۰/۳۰۰	۰/۲۰۵	۰/۲۸۵	۰/۲۱۰
C	۰/۳۲۲	۰/۲۹۸	۰/۰۷۸	۰/۳۰۲
G	۰/۲۴۸	۰/۲۴۶	۰/۲۹۸	۰/۲۰۸
T	۰/۱۷۷	۰/۲۳۹	۰/۲۹۲	۰/۲۹۲

که در آن ردیف اول در هر دو حالت بیانگر نسبت دفعاتی است که A بعد از هر کدام از چهار باز بیان شده آمده است، و به همین ترتیب برای ردیف های بعدی میباشد، مجموع هر ردیف برابر با یک میباشد، این احتمال ها یکسان نیستند ، مثلاً احتمال اینکه G بعد از A بیاید بسیار بیشتر از احتمال آمدن T بعد از A میباشد. توجه کنید که جدولها متقارن نیستند . در هر جدول احتمال آمدن G بعد از C کمتر از احتمال آمدن C بعد از G میباشد و همان طور که انتظار میرود این تأثیر در جدول "_" یعنی نواحی جدا از جزیره های CpG بیشتر است چون در این نواحی به دلیل نبودن جزیره های CpG ، به طور منطقی باید تعداد دفعاتی که G بعد ازC آمده است کمتر باشد .

برای این که از مدل های بالا برای تشخیص (بین نواحی با جزیره های CpG و مابقی نواحی) استفاده کنیم، log-odds ratio را محاسبه میکنیم :

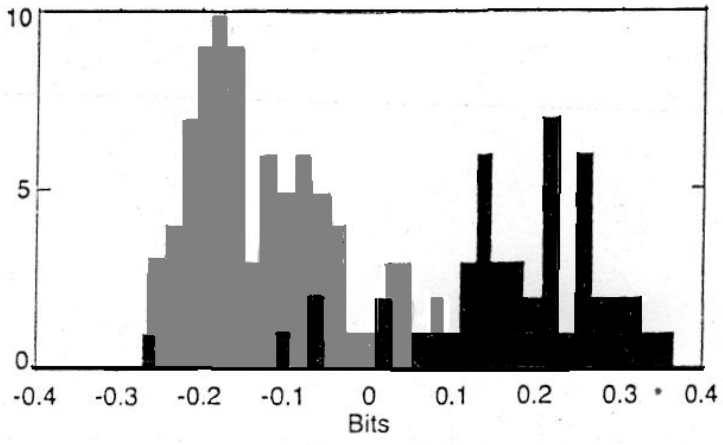
$$S(x) = \log[\, P(x|\ model +)/\ P(x|\ model -)] = E^{L}_{i=1} \log[a^{+}_{xi-1xi} / a^{-}_{xi-1xi}\ ]$$

$$= E^{L}_{i=1} \exists_{xi-1xi}$$

که در آن منظور از x رشته DNA میباشد و منظور از $\exists_{xi-1xi}$ لگاریتم نسبت احتمالِ (log likelihood ratios) احتمال های انتقال داده شده میباشد . در جدول زیر مقادیر$\exists$ آمده است :

$\exists$	A	C	G	T
A	-۰/۷۴۰	۰/۴۱۹	۰/۵۸۰	-۰/۸۰۳
C	-۰/۹۱۳	۰/۳۰۲	۱/۸۱۲	-۰/۶۸۵
G	-۰/۶۲۴	۰/۴۶۱	۰/۳۳۱	-۰/۷۳۰
T	-۱/۱۶۹	۰/۵۷۳	۰/۳۹۳	-۰/۶۷۹

(شکل ۲)



شکل بالا جدول هیستوگرام طول های نرمال شده تمام رشته ها میباشد، جزیره های CpG با هیستوگرام پررنگ ، بقیه با هیستوگرام های کم رنگ نشان داده شده اند.

جدول قبل نشان دهنده توزیع$S(x)$ است که با تقسیم شدن برطولشان نرمال شده اند مانند میانگین تعداد ذرات در یک مولکول. اگر آنها را با تقسیم بر طولشان نرمال نکنـیم ، آنگاه توزیع بسیار پراکنده خواهد بود.

در جدول تفاوت مشهودی بین نواحی جزیره CpG و بقیه نواحی دیده میشود ، ایـن تفاوت و تشخیص چندان تحت تأثیر نرمال شدن طول رشته ها واقع نمیشود بلکه با همان علامت های "+" و "_" شناخته میشود . بهتر است مورد توجه قرار گیرد که در مواردی که مدل اشتباه پارامتریزه شود و یا علامت های اشتباه بـرای داده مـورد بررسـی در نظـر گرفته شود ، ممکن است اشتباهاتی در نتایج رخ دهد .

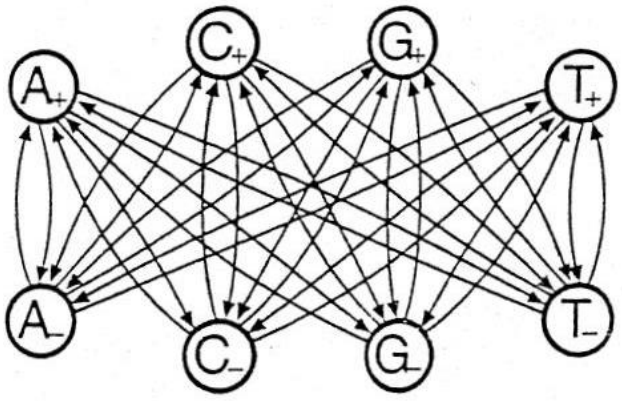
"زنجیرهای مارکف مخفی"

یکی از شاخه های مورد بررسی در زنجیرهای مارکف ، زنجیرهای مـارکف مخفـی میباشد. در قسمت اولیه بحث زنجیرهای مارکف و جزیره های CpG ، دو سـئوال مطـرح شد که در مباحث بالا به سئوال اول یعنی چگونگی پی بردن و متمـایز کـردن نـواحی کـه متعلق به جزیره های CpG هستند در برابر قسمتهایی از DNA کـه دارای ایـن خاصـیت نیستند ، پاسخ داده شد و حال میخواهیم سئوال دوم را مورد بررسی قـرار دهـیم یعنـی در یک رشته بلند ژن ، چگونه میتوان نواحی دارای جزیره های CpG را مشخص کرد؟(البته اگر این نواحی وجود داشته باشد.)

برای رسیدن به این سئوال با روش زنجیرهای مارکف به این صـورت عمـل میشـود که: برای هر نوکلئوتید در رشته مورد بررسی ۱۰۰ نوکلئوتید اطراف آن را بدست گرفته و log-odds score را در هر مورد بدست میاوریم و آن را رسم میکنـیم و در ایـن حالـت انتظار داریم که جزیره های CpG با مقادیر مثبت ظاهر شوند. با وجود این نمیتوان فـرض کرد که حوزه اطراف جزیره های CpG با این دقت مشخص میشوند و کاملاً متمایزند و در عین حال بررسی ۱۰۰ نوکلئوتید اطراف هر نوکلئوتید در رشته ، کار بسیار سختی میباشد. به همین خاطر روش بسیار بهتری مورد بررسی قرار میگیرد که عبـارت از سـاختن یـک مدل واحد برای کل رشته به طوری که هر دو زنجیر مارکف در هم یکی شوند. برای اینکه

بتوانیم هر دو قسمت جزیره های CpG و مابقی رشته را با هم نمایان کنیم ، لازم است که هر دو زنجیر مارکفی را که در مبحث قبل بیان شد در یک مدل داشته باشیم (در شکل بعدی به وضوح دیده خواهد شد) به این صورت که احتمال انتقال کمی از یک زنجیر به زنجیر دیگر وجود داشته باشد ، در این حالت ما دو وضعیت داریم که هر دو شبیه به نمونه نوکلئوتید میباشد و این مشکل را باید به گونه ای حل کرد. برای حل این مسأله ، موقعیتها را علامتگذاری میکنیم ، به این صورت که A_+ ,C_+ ,G_+,T_+ به ترتیب نشان دهنده (یا ساطع کننده) A,C,G,T در نواحی جزیره های CpG میباشند و به همین صورت A_- ,C_- ,G_- ,T_- نشان دهنده A,C,G,T در نواحی فاقد جزیره های CpG میباشند.

(شـــــــکل ۳)



این شکل نشان دهنده یک زنجیر مارکف مخفی برای جزیره های CpG میباشد . علاوه بر نشان دادن انتقالها ، تمام انتقالها بین هر گروه به صورت کامل نشان داده شده است .(همان طور که قبلاً در زنجیرهای ساده تر مارکف دیده بودیم.)

در مدلی که توضیح داده شد، احتمالهای انتقال به گونه ای در نظر گرفته شده است که در هر گروه احتمالهای انتقال بین وضعیت ها بسیار نزدیک به احتمالهای انتقال در همان بخش از مدل اصلی باشد ، ولی در این مدل احتمال کمی برای انتقال از یک بخش به بخش دیگر (یعنی از بخشی با جزیره های CpG که بخش مثبت میباشد به بخش بدون جزیره های CpG که بخش منفی میباشد) نیز وجود دارد . به طور کلی احتمال بیشتری برای انتقال از بخش "+" به بخش "_" وجود دارد تا برای انتقال از بخش "_" به بخش

"+" ، به همین خاطر مدل تمایل بیشتری برای گذراندن زمان در بخش "_" (یعنی بخش بدون جزیره های CpG) دارد تا در بخش "+" (یعنی بخش جزیره های CpG).

علامتگذاری وضعیت ها قدم بسیار مهمی در مسأله مطرح شده میباشد. تفاوت اساسی بین یک زنجیر مارکف و یک زنخیر مارکف مخفی این است که در مدل مارکف مخفی یک شباهت یک به یک بین وضعیت ها وجود ندارد ، در این حالت وقتی که x_i تولید شده، فقط با در نظر گرفتن x_i نمیتوان پی برد که مدل در چه وضعیتی قرار داشته است . (مثالی که کمی بعد توضیح داده خواهد شد مطلب را روشن تر میسازد.) .در مثال قبلِ ما هنگامی که وضعیت C مشاهده میشود نمیتوان بیان کرد که این C توسط C_+ تولید شده یا C_-. منظور از این مطلب این است که مشاهده C به تنهایی نمیتواند به ما نشان دهد که این C از بخش جزیره های CpG آمده است یا از بخش دیگر.

"توضیحی درباره مدل مارکف مخفی یا HMM "

در این قسمت میخواهیم علامت ها و نشانه های مدل مارکف مخفی را فرمول بندی کنیم تا بتوانیم احتمالهای مربوط به وضعیت ها و سمبول ها را بدست آوریم. در حال حاضر لازم است بتوانیم بین رشته وضعیت ها و رشته سمبول ها تمایز قائل شویم. در این جا نام رشته وضعیت ها را مسیر B (path) میگذاریم. یک مسیر به تنهایی دارای خصوصیت زنجیر مارکف ساده میباشد، در نتیجه احتمال مشاهده یک وضعیت فقط به وضعیت ماقبلش بستگی دارد. i اُمین وضعیت در مسیر، B_i نامیده میشود و زنجیر با پارامترهای زیر مشخص میشود:

$$a_{kl}=P(B_i=l \mid B_{i-1}=k) \qquad (۴)$$

برای این که شروع روند را مدل سازی کنیم، وضعیت اولیه را معرفی میکنیم ، همانطور که قبلاً در مورد شروع زنجیر در زنجیرهای مارکف نیز این کار را کردیم. (همانند

شکل ۱) احتمال انتقال a_{0k} از این وضعیت اولیه به وضعیت k مثل احتمال شروع از وضعیت k میباشد . در عین حال میتوانیم مانند زنجیرهای مارکف برای این حالت نیز وضعیت پایان در نظر بگیریم ، به این صورت که همیشه زنجیر با یک انتقال به وضعیت پایانی به اتمام برسد . برای راحت بودن، هر دو وضعیت شروع و پایان را با ۰ (عدد صفر) نشان میدهیم. (در این جا مشکلی پیش نمیاید، چون فقط میتوان از موقعیت اولیه خارج شد و تنها امکان پذیر است که به وضعیت پایانی وارد شد، در نتیجه متغیرها بیشتر از یک بار مورد استفاده قرار نمیگیرند.)

به این خاطر که ما سمبول های b را با وضعیت های k جفت کردیم ، باید برای مدل پارامترهای $e_k(b)$ را نیز تعریف کنیم. برای مدل CpG ما، هر وضعیت به یک سمبول وابسته میباشد، ولی این حالت همیشگی و لازم نیست، در حالت کلی یک وضعیت میتواند یک سمبول از بین توزیع تمام سمبولهای ممکن تولید کند. بنابراین تعریف میکنیم:

$$e_k(b)=P(x_i=b \mid B_i=k) \qquad (۵)$$

تساوی بالا بیانگر احتمال این است که سمبولb دیده شود به شرط این که در وضعیت k باشیم ، که به این احتمالها ($e_k(b)$ ها) احتمالهای پخش یا ارسال (emission probabilities) گفته میشود.

برای مدل جزیره های CpG ما، احتمالهای پخش همواره ۰ یا ۱ میباشد. در این جا برای توضیح بیشتر در مورد احتمالهای پخش مثالی در مورد کازینو (قمارخانه) آورده میشود .

مثال: قمارخانه (قسمت اول)

در قمارخانه معمولاً افراد از تاس سالم استفاده میکنند، ولی گاهی اوقات به جای آن از تاس ناسالم (یا باردار) استفاده میکنند. تاس ناسالم دارای احتمال ۰/۵ برای عدد ۶ و احتمال ۰/۱ برای اعداد ۱ تا ۵ میباشد و این یعنی ما دارای فضای وضعیتی به

صورت{تاس ناسالم ، تاس سالم}=S هستیم و توزیع احتمالهای پخش برای تـاس سـالم برابر است با: (اگر در نظر بگیریم: تاس سالم =A)

$$e_a(1) = e_a(2)= e_a(3)= e_a(4)= e_a(5)= e_a(6)=1/6$$

و توزیع احتمالهای پخش برای تاس ناسالم برابر است با : (اگـر در نظـر بگیـریم : تـاس ناسالم=B)

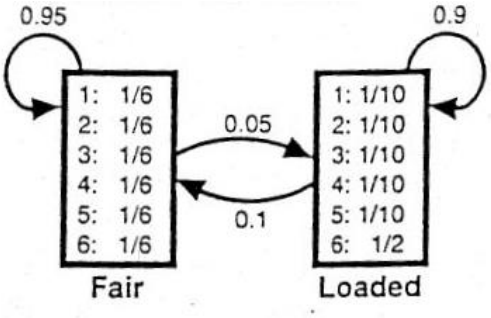
$$e_b(1)= e_b(2)= e_b(3)= e_b(4)= e_b(5)=1/10 \qquad , e_b(6)=1/2$$

به عنوان مثال احتمالهای بالا به صورت زیر بدست آمده اند:

$$e_a(1)=Pr(x_i=1 \mid B_{i=}A)=1/6$$

یعنی احتمال آمدن سمبول ۱ به شرط انداختن تاس (وضعیت) A برابر با ۰/۱۶۶۶ میباشد.

حال فرض کنید احتمال این که کازینو، قبل از هر پرتاب تـاس ناسـالم را بـه سـالم تغییر دهد، برابر با ۰/۰۵ و احتمال برعکس آن یعنی تغییر تاس از ناسالم به سالم برابر بـا ۰/۱ باشد، در نتیجه انتقال بین تاسها دارای رونـد مـارکف میباشـد. همـانطور کـه دربـالا توضیح داده شد در هر وضعیت از روند مارکف مشـاهدات هـر پرتـاب دارای احتمالهـای متفاوتی میباشند، یعنی بسته به این که کدام تاس پرتاب شده است احتمـال مشـاهده عـدد پرتاب شده متفاوت است. همیشه باید در نظر داشته باشیم که ما به عنوان بیننـده اطلاعـی از این که کدام تاس پرتاب شده است نداریم، در نتیجه کل روند بالا به عنوان مدل مارکف مخفی شناخته میشود ، که میتوانیم آن را به صورت زیر نشان دهیم:



به طوری که احتمالهای پخش( )e در مستطیل های وضعیت نشان داده شده اند.

چه چیزی در مدل قبلی مخفی میباشد؟ همانطور که گفتیم اگر شما فقط بتوانید رشته ای از پرتاب ها (رشته ای از مشاهدات) را ببینید نمیتوانید بگویید در کدام پرتاب ها از تاس سالم استفاده شده است و در کدامیک از تاس ناسالم. چون این مسأله توسط کازینو از شما مخفی نگاه داشته میشود، یعنی زنجیر وضعیت (در این جا زنجیر تاس ها) مخفی است. در زنجیر مارکف شما همیشه دقیقاً میدانید که یک مشاهده به چه وضعیتی تعلق دارد، اما به طور روشن کازینو به شما اطلاع نمیدهد که از تاس ناسالم استفاده میکند و احتمال هر وجه تاس چقدر خواهد بود . حال برای این وضعیت پیچیده که بعداً باز هم به آن خواهیم پرداخت، میتوانیم احتمال های مدل مارکف مخفی آن را بدست آوریم. (وقتی که به این مسأله که از دو تاس متفاوت در پرتاب ها استفاده شده شک داریم.)

دلیل این که از نام احتمالهای پخش(emission) برای مدلهای مارکف مخفی استفاده میکنیم این است که معمولاً بهتر است این مدلها را مدلهای تولیدکننده در نظر بگیریم که یا زنجیر تولید میکنند و یا دفع میکنند. به این مفهوم که با پرتاب هریک از تاس های سالم یا ناسالم زنجیری تولید میشود که از پدید آمدن زنجیری که توسط تاس دیگر میتوانست به وجود آید جلوگیری میشود. در مفهوم کلی یک زنجیر توسط یک مدل مارکف مخفی میتواند به صورت زیر تولید شود: در ابتدا وضعیت B_1 با توجه به احتمال های انتقال a_{0i} انتخاب میشود، در این وضعیت یک مشاهده با توجه به توزیع e_{B1} برای همان وضعیت B_1 رخ میدهد، در ادامه یک وضعیت جدید B_2 با توجه به احتمال های انتقال a_{B1i} انتخاب میشود و همین طورتا آخر. به این ترتیب یک زنجیر از مشاهدات تصادفی و ساختگی تولید شده است. به همین خاطر گاهی اوقات میگوییم $P(x)$ احتمال این است که x توسط مدل تولید شده باشد.

حال میتوانیم به راحتی احتمال توأم رشته مشاهده شده x و رشته وضعیتهای B را به صورت زیر بدست آوریم:

$$P(x,B)=a_{0B1}A^{L}_{i=1}e_{Bi}(x_i)a_{Bi\ Bi+1} \qquad (۶)$$

در این جا باید فرض کنیم $B_{L+1}=0$ ، به عنوان مثال، احتمال این که زنجیر CGCG توسط زنجیر وضعیت (C_+,G_-,C_-,G_+) تولید شود برابر است با :

$$a_{0.C+}\times 1 \times a_{C+.G-}\times 1\times a_{G-.C-}\times 1\times a_{C+.G+}\times 1\times a_{C-.0}$$

(گفتیم در جزیره های CpG احتمالهای پخش همواره ۰ یا ۱ میباشند که در این جا چون تمام $e_k(b)$ ها مشاهده شده اند پس تمام احتمالهای پخش برابر با یک میباشند.)

تساوی (۶) نوع مشابه تساوی (۴) میباشد، البته در مورد مدل های مارکف مخفی. با این حال در عمل چندان مورد استفاده نمیباشد چون ما در واقع مسیر(یا همان وضعیت) B را نمیدانیم. در مبحث بعدی توضیح میدهیم که چگونه مسیر را برآورد کنیم، چه از طریق پیدا کردن محتمل ترین مسیر و چه از طریق استفاده از توزیع وضعیت قبلی. بعد از آن نشان میدهیم که چگونه میتوان پارامترها را برای مدل مارکف مخفی برآورد کرد.

"محتمل ترین وضعیت(مسیر): الگوریتم ویتربی"

The Viterbi Algorithm

با وجود این که دیگر این امکان وجود ندارد که با نگاه کردن به یک سمبول متناسب بیان کنیم که سیستم در چه وضعیتی قرار دارد، معمولاً رشته هایی از وضعیت های بنیادی و مهم مورد توجه ما میباشند. به فهمیدن اینکه رشته مشاهدات با در نظر گرفتن وضعیتهای اساسی چه مفهومی میدهد، رمزگشایی (decoding) گفته میشود. روش های چندی برای رمزگشایی وجود دارد. منظور از گفته های ذکر شده این است که ما عملاً در قضیه زنجیرهای مارکف مخفی با مدلهای پیچیده تری نسبت به مثال ذکر شده در مورد کازینو روبه رو هستیم که فقط با نگاه کردن به یک رشته از مشاهدات نمیتوانیم متوجه شویم که این رشته به کدام یک از وضعیت ها تعلق دارد. (مثلاً در مثال قبل باید بفهمیم اعداد مشاهده شده به کدام تاس مربوط میشوند.) در نتیجه با روش های پیچیده تری باید به این

مسأله پی ببریم و به این که مشاهدات ما به کدام وضعیت مربوط میشوند رمزگشایی گفته میشود.

در این جا ما معمول ترین روش رمزگشایی را که الگوریتم ویتربی نام دارد توضیح خواهیم داد.در حالت کلی ممکن است رشته های وضعیت زیادی وجود داشته باشد که بتوان آنها را به یک رشته خاص از سمبول ها مرتبط کرد. به عنوان مثال، در مورد جزیره های CpG زنجیرهای وضعیت(C_+, G_+, C_+, G_+) ،(C_-, G_-, C_-, G_-) و (C_+, G_-, C_+, G_-) همگی میتوانند زنجیر سمبول CGCG را تولید کنند. با این حال این عمل برای هر کدام از زنجیرهای وضعیت ذکر شده با احتمالات متفاوتی صورت میگیرد. سومین زنجیر حاصل ضرب احتمال های کوچک جابه جایی بین اجزاء میباشد، در نتیجه احتمال این که این زنجیر، CGCG را بسازد از دو تای اول بسیار کوچک تر است. دومین زنجیر به وضوح از اولین زنجیر احتمال کمتری برای رخدادن دارد به این خاطر که این زنجیر شامل دو انتقال از C به G میباشد که این وضعیت به وضوح در حالت "_" احتمال کمتری نسبت به حالت "+" دارد زیرا بر طبق آنچه که در ابتدای بحثِ زنجیرهای مارکف در مورد جزیره های CpG گفته شد احتمال وجود C و G در کنار هم در جزیره های CpG بیشتر است چون در قسمتهای دیگر رشته احتمالِ تبدیلِ C به باز T در اثر متیله شدن بیشتر است. در نتیجه در سه انتخاب بالا احتمال این که رشته CGCG از وضعیت های "+" آمده باشد بیشتر است.

مسیر پیش بینی شده در زنجیرهای مارکف مخفی به ما نشان خواهند داد که کدام قسمت از رشته به عنوان جزیره های CpG معرفی خواهند شد، به این خاطر که همانطور که در بالا ذکر شد، هر وضعیت تعیین کننده جزیره های CpG و یا عدم حضور آنها در یک ناحیه میباشد. اگر قرار باشد ما حتماً یک مسیر برای پیش بینی خود انتخاب کنیم، باید مسیری را انتخاب کنیم که دارای بالاترین درجه احتمال باشد:

$$B^*=\arg_B \max P(x,B) \qquad (۷)$$

مسیر با بیشترین احتمال B^* میتواند به روش تکرار به دست آیـد. فـرض کنیـد احتمـال $<_k(i)$ مربوط به مسیر با بیشترین احتمال که در وضعیت k پایان میابد و دارای مشـاهده i میباشد، برای تمام وضعیت های k تعریف شده باشد، به ایـن ترتیـب ایـن احتمالهـا بـرای مشاهده x_{i+1} میتوانند به صورت زیر محاسبه شود:

$$<_l(i+1)= e_l(x_{i+1})max_k(<_k(i)a_{kl}) \qquad (۸)$$

همه رشته ها باید در وضعیت ۰ (وضعیت شروع) شروع شوند، در نتیجه شرط اولیـه ایـن است که: $<_0(0)=1$ باشد.

(شکل ۴)

v		C	G	C	G
$\mathcal{B}$	1	0	0	0	0
A_+	0	0	0	0	0
C_+	0	**0.13**	0	**0.012**	0
G_+	0	0	**0.034**	0	**0.0032**
T_+	0	0	0	0	0
A_-	0	0	0	0	0
C_-	0	0.13	0	0.0026	0
G_-	0	0	0.010	0	0.00021
T_-	0	0	0	0	0

نتیجه محاسبات احتمالهای< در مورد جزیره های CpG نشان داده شده در شکل ۳ و رشـته CGCG . مسیر با بیشترین احتمال به صورت پررنگ نشان داده شده است.

مسیر اصلی از روش الگوریتم ویتربی میتواند به صورت عقبگرد به دست آید یعنـی برای بدست آوردن احتمالات مربوط به یـک مشـاهده از مشـاهدات ماقبـل آن اسـتفاده میکنیم.

Algorithm: Viterbi **الگوریتم ویتربی:**

Initialisation(i=0): $<_0(0)=1, <_k(0)=o \text{ for } k>0$ (شروع)

Recursion(i=1,…,L): $<_l(i)=e_l(x_i)\max_k(<_k(i-1)a_{kl})$ (تکرار)

$Ptr_i(l)=\text{argmax}_k(<_k(i-1)a_{kl})$

Termination : $P(x, B^*)=\max_k(<_k(L)a_{k0})$ (پایان)

$B^*_L=\text{argmax}_k(<_k(L)a_{k0})$

Trace back(i=L,…,1): $B^*_{i-1}=ptr_i(B^*_i)$ (بازگشت)

توجه به این نکته لازم است که وضعیت پایانی که دلیلی است برای وجود a_{k0} در مرحله پایان (termination step) در نظر گرفته شده است. اگر پایان مرحله مدل سازی نشود آنگاه دیگر این a وجود نخواهد داشت.

روش های قابل اجرایی برای الگوریتم ویتربی و الگوریتمی که در آینده در مورد آن توضیح میدهیم وجود دارد. مشکل اساسی در این مورد این است که ضرب کردن تعداد زیادی از احتمالات باعث به وجود آمدن اعداد بسیار کوچکی میشود که عموماً باعث ایجاد خطا در هر محاسبه کامپیوتری میشود، به همین خاطر الگوریتم ویتربی همیشه باید در فضای لگاریتم محاسبه شود مثلاً محاسبه $\log(<_l(i))$ ، که باعث میشود جواب ها و محاسبات دارای جوابهای بسیار منطقی تری باشند. جدول (۴) نشان دهنده جدول کاملی از مقادیر > برای رشته CGCG و مدل جزیره CpG میباشد. وقتی که ما از الگوریتم مشابهی برای رشته بلندتری استفاده میکنیم، مسیر B^* بهینه بدست آمده، بین موقعیت های "+" و "_" عوض میشود، در نتیجه محدوده دقیق تری از جزیره CpG پیش بینی شده در اختیار ما قرار میدهد.

مثال: قمارخانه (قسمت دوم)

حال برای زنجیری از پرتاب تاس ها با مدل نشان داده شده در ابتدای مثال قمارخانه (قسمت اول) میتوانیم محتمل ترین مسیر را بیابیم. مجموعه ای از ۳۰۰ پرتاب

تصادفی از مدل طرح شده در مثال قبل تولید شده است، هر پرتاب ممکن است توسط تاس سالم(F) یا تاس ناسالم (L) تولید شده باشد، در شکل بعد نتیجه این پرتاب ها مشاهده میشود. در این جا الگوریتم ویتربی برای این استفاده میشود که پیش بینی کند کدام یک از تاس ها برای هر کدام از پرتاب ها در نظر گرفته شده است. در حالت کلی مشاهده میکنید که این پیش بینی تا مقادیر بسیار زیادی درست میباشد و تعداد بسیار زیادی از پرتاب ها را درست پوشش داده است.

```
Rolls   3151162464466442453113216311641521336251445436316566266566666
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL

Rolls   6511664531326512456366646316366631623264552362666666625151631
Die     LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls   2225554416665665635643243641315134651463534111264146262533 56
Die     FFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls   3661636664662325344136616611632525624622552652522664353533 36
Die     LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   2331216253644144323351632436336655624666626326666123552452 42
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
```

(شکل ۵)

اعداد نشان دهنده نتیجه ۳۰۰ بار پرتاب تاس ها همانطور که در مثال توضیح داده شد، میباشند. خط پایین آن نشان دهنده این است که در اصل کدام تاس پرتاب مورد نظر را انجام داده است. آخرین خط نشان دهنده پیش بینی انجام شده توسط الگوریتم ویتربی میباشد.

مطلب مهم: $B^{*}= \arg_B \max P(B|x)=\arg_B \max P(x,B)$

"الگوریتم جلوروندہ " The forward algorithm

برای زنجیرهای مارکف، ما احتمال مربوط به یک رشته یعنی P(x) را توسط تساوی (۲) بدست آوردیم. نتیجه های بدست آمده به عنوان مثال برای تشخیص بین جزیره های CpG و سایر نواحی DNA استفاده میشدند. ما میخواهیم بتوانیم این احتمال را برای زنجیرهای مارکف مخفی نیز بدست آوریم. به این خاطر که مسیرهای متفاوتی از وضعیت میتواند منجر به یک رشته واحد x شود ما باید احتمالهای همه مسیرهای ممکن را با هم جمع کنیم تا بتوانیم احتمال کامل رشته x را بدست آوریم:

$$P(x)=E_B P(x,B) \qquad (۹)$$

تعداد مسیرهای ممکن B با افزایش طول زنجیر به صورت تصاعدی افزایش میابد، به همین خاطر بررسی تساوی بالا به وسیله شمردن همه مسیرهای ممکن منطقی نمیباشد. یک روش این است که با استفاده از تساوی (۶) که توسط محتمل ترین مسیر B^* تخمین زده شده است برآوردی از P(x) بدست آوریم. این حالت به طور قطع نشان دهنده این است که تنها مسیر با احتمال قابل توجه B^* میباشد، فرضیه ای نسبتاً تکان دهنده که به هر حال در بسیاری از موارد بسیار خوب عمل میکند. در واقع در این جا تخمین زدن لازم است، به این دلیل که احتمال کامل توسط یک روش تعیین شده شبیه به الگوریتم ویتربی، که قدمهای ماکزیمم کردن توسط جمع کردن جایگزین شده قابل محاسبه است. به این روش محاسبه، الگوریتم جلو رونده میگویند.

کمیت مشابه $<_k(i)$ در محاسبه الگوریتم ویتربی، در این حالت یعنی در الگوریتم جلو رونده به صورت زیر محاسبه میشود:

$$f_k(i)=P(x_1,\ldots,x_i, B_i=k) \qquad (۱۰)$$

تساوی قبل احتمال مشاهده زنجیر است تا آنجایی که شامل x_i شود، به طوری که $B_i=k$ و یا میتوان گفت محتمل ترین مسیر کـه در وضـعیت k پایـان میابـد و شـامل مشـاهده i میباشد. تساوی بازگشتی به صورت زیر میباشد.

$$f_l(i+1)=e_l(x_{i+1})E_k f_k(i)a_{kl} \qquad (۱۱)$$

الگوریتم کامل به صورت زیر تعریف میشود:

الگوریتم جلوروندہ: Forward algorithm

(شروع) Initialisation (i=0) : $f_0(0)=1$, $f_k(0)=0$ for $k>0$

(بازگشت) Recurtion(i=1,…L) : $f_l(i)= e_l(x_i)E_k f_k(i-1)a_{kl}$

(پایان) Termination: $P(x)=E_k f_k(L)a_{k0}$

الگوریتم جلوروندہ (و همچنین الگوریتم معکوس که در مبحث بعد مطالعه میشود) نیز مثل الگوریتم ویتربی میتواند موقع اجرا کردن در کامپیوتر دچار خطـا شـود. در این حالت نیز مشکل با استفاده از کار کردن در محیط لگاریتم حل میشود، البتـه بـه زیبایی استفاده از لگاریتم در الگوریتم ویتربی نخواهد بود.

"الگوریتم معکوس و احتمالهای وضعیت های پسین"

الگوریتم ویتربی مسیر با بیشترین احتمال را در مدل پیدا میکرد، امـا همـانطور کـه مشاهده شد، این الگوریتم ممکن است همیشه برای نتیجـه گیـری هـای بیشـتر در مـورد زنجیر مناسب ترین روش نباشد. به عنوان مثال ممکن اسـت مـا بخـواهیم محتمـل تـرین وضعیت را برای مشاهده x_i بدانیم. به طور کلی ممکن است بخـواهیم در زنجیـر مشـاهده شده احتمال این که مشاهده x_i از وضعیت k بدست آمده باشد یعنی $P(B_i=k|x)$ را بدست آوریم.این احتمال، احتمال پسین وضعیت k در زمان i میباشد وقتی که زنجیر تولیـد شـده

را میشناسیم. این توضیح را اضافه میکنیم که در حالت های قبلی ما با توجه به مشاهدات، محتمل ترین مسیر را بدست میاوردیم اما در این جا مسیر را داریم و حال میخواهیم بدانیم هر x_i که مشاهده میکنیم به کدام وضعیت از این زنجیر بستگی دارد.

روش ما برای رسیدن به احتمال پسین تا حدودی غیر مستقیم است. در ابتدا ما احتمال تولید کل زنجیر مشاهده شده را به صورتی که سمبول i توسط وضعیت k تولید شود محاسبه میکنیم:

$$P(x,B_i=k)=P(x_1\ldots x_i, B_i=k)P(x_{i+1}\ldots x_L|x_1\ldots x_i, B_i=k)$$
$$=P(x_1\ldots x_i, B_i=k)P(x_{i+1}\ldots x_L| B_i=k) \qquad (۱۲)$$

دلیل نوشتن خط دوم تساوی این است که هر چیزی بعد از وضعیت k فقط به وضعیت k بستگی دارد. جمله اول در قسمت دوم این تساوی مشابه $f_k(i)$ در تساوی (۱۰) میباشد که توسط الگوریتم جلورونده محاسبه شده است جمله دوم $b_k(i)$ نامیده میشود:

$$b_k(i)= P(x_{i+1}\ldots x_L| B_i=k) \qquad (۱۳)$$

$b_k(i)$ مشابه متغیر الگوریتم جلورونده است، ولی در کنار آن شامل یک روند بازگشتی معکوس است که از پایان زنجیر شروع میشود:

الگوریتم معکوس: Backward algorithm

(شروع) Initialisation (i=L): $b_k(L)=a_{k0}$ برای همه k

(بازگشت) Recursion(i=L-1,…,1): $b_k(i)=E_l a_{kl} e_l(x_{i+1}) b_l(i+1)$

(پایان) Termination : $P(x)=E_l a_{0l} e_l(x_1) b_l(1)$

مرحله پایانی بسیار به ندرت مورد استفاده واقع میشود زیرا معمولاً $P(x)$ توسط الگوریتم جلورونده محاسبه میشود، و در این جا فقط برای کامل شدن الگوریتم نشان داده شده است.

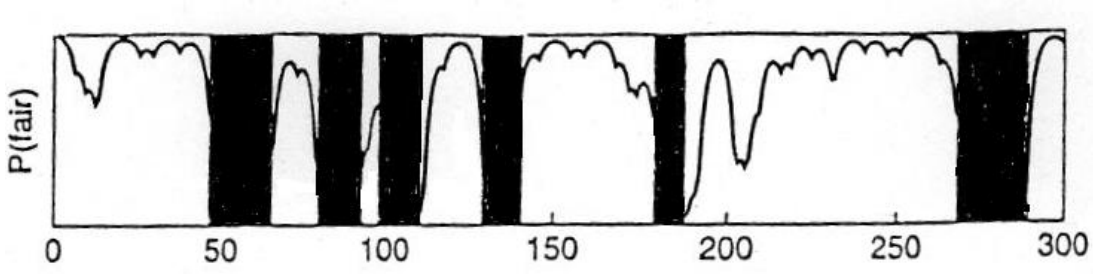
تساوی (۱۲) حالا میتواند به صورت $P(x,B_i=k)=f_k(i)b_k(i)$ نوشته شود وبا توجه به این تساوی میتوانیم احتمال های پسین مورد نیاز را بدست آوریم:

$$P(B_i=k|x)=[f_k(i)b_k(i)]/p(x) \qquad (۱۴)$$

به طوری که $P(x)$ نتیجه محاسبات الگوریتم جلورونده (یا معکوس) میباشد.

مثال: قمارخانه (قسمت سوم)

در شکل (۶) احتمال پسین برای تاس سالم که در زنجیری از پرتاب ها در شکل (۵) نشان داده شده است را میبینیم. توجه کنید که احتمال پسین نشان نمیدهد که کدام تاس واقعاً در برخی مکانها استفاده شده است. البته این امر غیر منتظره نیست زیرا به طور تصادفی ممکن است زنجیر گمراه کننده ای از پرتاب ها رخ دهد. یعنی ممکن است زنجیری از پرتاب ها مشاهده شود که به ظاهر مربوط به تاس سالم باشد ولی در واقع این طور نباشد و این امر باعث اشتباه در احتمالهای پسین میشود.



احتمال پسینِ بودن در وضعیت مشابه تاس سالم در مثال قمارخانه. محور x نشان دهنده تعداد پرتاب هاست. نواحی سایه زده شده نشاندهنده زمانهایی است که پرتاب توسط تاس ناسالم صورت گرفته است.

رمزگشایی پسین Posterior decoding

استفاده اصلی از $P(B_i=k|x)$ برای دو فرم دیگر از رمزگشایی علاوه بر رمزگشایی ویتربی که در مبحث قبلی توضیح دادیم میباشد. این ها به خصوص برای زمانی که مسیرهای متفاوت زیادی داریم که دارای احتمالهای مشابه احتمال محتمل ترین مسیر میباشند قابل استفاده هستند، به این خاطر که در این حالت دیگر چندان منطقی نمیباشد که فقط مسیر با بیشترین احتمال را مورد بررسی قرار دهیم.

روش اول این است که یک زنجیر وضعیت $B^{\wedge}_i$ که بتواند به جای B^{*}_i استفاده شود مشخص کنیم:

$$B^{\wedge}_i = \arg_k \max P(B_i=k|x) \qquad (۱۵)$$

همانطور که در توضیح $B^{\wedge}_i$ دیده شد ، این زنجیر وضعیت ممکن است زمانی که ما علاقه مند به تعیین وضعیت در یک موقعیت مشخص i هستیم، نسبت به زمانی که کل مسیر را بررسی میکنیم مناسب تر باشد. در واقع زنجیر وضعیت که با $B^{\wedge}_i$ مشخص میشود ممکن است کاملاً شبیه به مسیر درون مدل کامل نباشد، حتی زمانی که برخی از انتقالها مجاز نیستند ممکن است یک مسیر معقول نیز به نظر نیاید (چون ممکن است شامل این انتقالها شود) که این اتفاق بسیار زیاد رخ میدهد.

دومین و احتمالاً مهم ترین روش کدگذاری زمانی رخ میدهد که خود زنجیر وضعیت نیست که مورد توجه و اهمیت است بلکه خصوصیتی که از آن بدست میاید دارای اهمیت است . فرض کنید تابع $g(k)$ را داریم که با توجه به وضعیت ها تعریف میشود. آنگاه مقدار طبیعی که مورد توجه میباشد عبارت زیر است:

$$G(i|x) = E_k P(B_i=k|x)\, g(k) \qquad (۱۶)$$

یک موقعیت مهم در این عبارت زمانی است که $g(k)$ مقدار عددی یک را برای زیر مجموعه ای از وضعیت ها و مقدار عددی صفر را برای مابقی زیرمجموعه ها اختیار

میکند. در این حالت $G(i|x)$ احتمال پسین این است که سمبول i از یک وضعیت درون یک گروه مشخص بیرون بیاید. به عنوان مثال ، در مورد مدل جزیره های CpG ما، آنچه که واقعاً مورد توجه ماست این است که یک باز مربوط به قسمت جزیره CpG است یا خیر، برای این هدف ما میخواهیم $g(k)=1$ را برای $k \mu \{A_+,C_+,G_+,T_+\}$ و $g(k)=0$ را برای $k \mu \{A_-,C_-,G_-,T_-\}$ تعریف کنیم. در نتیجه $G(i|x)$ دقیقاً احتمال پسین است با توجه به مدلی که در آن باز i در جزیره CpG قرار داشته باشد، در حالتی که ما علامتگذاری وضعیت هایی را داریم که توسط قسمتی از آنها تعریف میشوند (همانطور که در واقع در مدل جزیره های CpG ، وضعیت ها را با " + " و " _ " علامتگذاری میکنیم.) یعنی مثلاً قسمتی از آنها با " + " و قسمتی دیگر با " _ " نمایش داده میشوند. میتوانیم از معادله (۱۶) برای پیدا کردن محتمل ترین علامت (مثلاً " + " یا " _ " در مدل جزیره های CpG) در هر موقعیت از زنجیر استفاده کنیم. این روش البته روش چندان مناسبی برای علامتگذاری زنجیرها نمیباشد.

مثال: پیش بینی جزیره های CpG

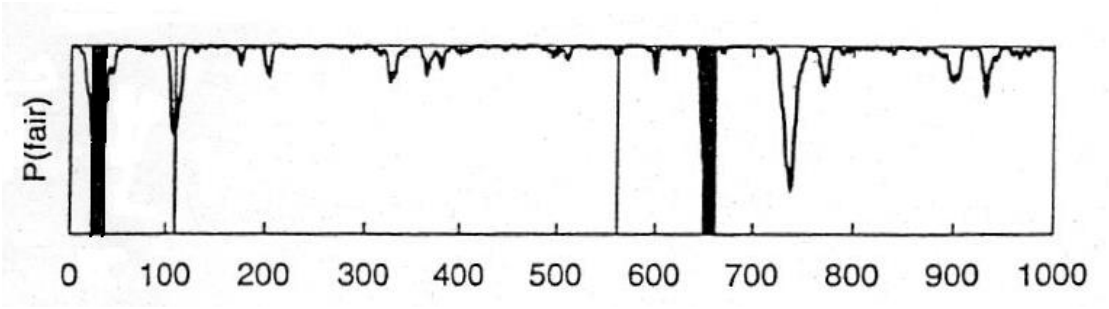
حال جزیره های CpG میتوانند از روی مدل ما پیش بینی شوند. توسط الگوریتم ویتربی ما میتوانیم محتمل ترین مسیر را در مدلمان پیدا کنیم. وقتی که این مسیر وارد وضعیت های " + " میشود ، یک جزیره CpG پیش بینی شده است. برای گروهی از ۴۱ زنجیر ، که هر کدام شامل یک جزیره CpG فرضی میباشد، همه جزیره ها به جز ۲ تا (منفی هایی اشتباه پیش بینی شده بودند یعنی در واقع مثبت بوده اند) پیدا شده اند و ۱۲۱ جزیره جدید نیز پیش بینی شده اند (با در نظر گرفتن مثبت هایی که اشتباه پیش بینی شده اند). جزیره های CpG واقعی تقریباً طولانی هستند (تقریباً حدود ۱۰۰۰ باز را شامل میشوند) . در حالی که جزیره های پیش بینی شده کوتاه میباشند، و جزیره های CpG معمولاً به صورت تعدادی باز کوتاه پیش بینی میشوند. با استفاده از دو قدم ساده:

(۱) پیش بینی های زنجیره ای شامل کمتر از ۵۰۰ بـاز جـدا از هـم (۲)دور انـداختن پیش بینی های کوتاه تر از ۵۰۰ باز، تعداد پیش بینی های مثبت اشتباه به ۶۷ تـا کـاهش میابد.

هنگام استفاده از رمزگشایی پسین، دو جزیره CpG مشابه مشـاهده نشـدند و ۲۳۶ مثبت اشتباه، پیش بینی شده اند. استفاده از دو مرحله ذکر شده در بالا این عـدد را بـه ۸۳ کاهش میدهد. برای این مسأله ، چندان تفاوتی بین دو روش ذکر شده وجود ندارد، به جز اینکه روش کد گذاری پسین تعداد بیشتری جزیره های CpG بسیار کوچکی را پیش بینی میکند. این امکان وجود دارد که برخی مثبت های اشتباه، جزیره های CpG واقعی باشند. دو منفی اشتباه احتمالاً به صورت اشتباه برچسب گذاری شده اند، ولی با ایـن حـال ایـن امکان وجود دارد که یک مدل پیشرفته تر برای بدست آوردن تمام خصوصیات این علائم لازم باشد.

مثال: قمارخانه (قسمت چهارم)

فرض میکنیم مدل قبلی برای قمارخانه تغییر پیدا کرده و احتمال گذر از تاس سـالم به ناسالم احتمالی معادل ۰/۰۱ میباشد. به وضوح احتمال باقی ماندن تـاس سـالم در هـر پرتاب معادل ۰/۹۹ میباشد، ولی بقیه احتمال ها تغییر نکرده باقی میمانند. برای این مـدل ۱۰۰۰ پرتاب تصادفی تولید شده است، برای این پرتاب ها محتمل ترین مسیری که توسط الگوریتم ویتربی بدست آمد هیچگاه وضعیت تاس ناسالم را مشاهده نمیکند. در شکل (۷) احتمال پسین این که در این پرتاب ها از تاس سالم استفاده شـده باشـد نشـان داده شـده است. البته روش کدگذاری پسین نه به صـورت کامـل ولـی بسـیار نزدیـک بـه واقعیـت، وضعیت ها را پیش بینی میکند.



(شکل ۷)

احتمال پسین برای تاس سالم ولی دارای احتمال ۰/۰۱ برای گدر به تاس ناسالم.