روش‌های محاسباتی برای پیش‌بینی ساختار دوم RNA

محمد گنج‌تابش
دانشگاه تهران، پردیس علوم، دانشکده ریاضی، آمار و علوم کامپیوتر
نویسنده مسئول: *mgtabesh@ut.ac.ir*

چکیده

عملکرد مولکول‌های RNA اغلب به ساختار فضایی آن‌ها بستگی دارد. ساختار فضایی یک مولکول RNA را می‌توان با روش‌های آزمایشگاهی مانند NMR[1] و یا کریستالوگرافی اشعه‌ی ایکس به طور دقیق مشخص کرد، ولی این کار مستلزم صرف زمان و هزینه‌ی بالایی است. به همین دلیل، استفاده از روش‌های محاسباتی برای پیش‌بینی ساختار فضایی یک مولکول RNA بسیار مورد توجه قرار گرفته است. چون مسئله پیش‌بینی ساختارهای فضایی RNA بسیار پیچیده بوده و از نظر محاسباتی پرهزینه است، لذا اکثر تحقیقات انجام شده در این زمینه بر روی مسائلی تمرکز دارند که به ساختار دوم RNA[2] مرتبط‌اند. این نوع ساختار را می‌توان به صورت مجموعه‌ای از موقعیت‌های جفت‌شده در یک دنباله‌ی RNA توصیف کرد. مسئله پیش‌بینی ساختار دوم RNA در حدود ۳۰ سال پیش ارائه شده و تا کنون تحقیقات زیادی در این مورد انجام شده است. کمینه‌سازی سطح انرژی به عنوان یکی از رویکردهای مهم برای حل این مسئله پیشنهاد شده و بر اساس آن الگوریتم‌های نوسینف[3] و زوکر[4] ارائه شده‌اند. در این مقاله، پس از ارائه تعاریف اولیه مربوط به ساختارهای RNA ، این دو الگوریتم با جزئیات کامل ارائه و تحلیل می‌شوند.

کلمات کلیدی: ساختار RNA ، انرژی آزاد کمینه، توالی RNA

مقدمه

یکی از شاخه‌های مهم علوم زیستی، علم ژنتیک است که از زمان ارائه تجربیات مندل بر روی گیاه نخودفرنگی (سال ۱۸۶۵ میلادی) پایه‌های علمی آن ریخته شد و قریب ۱۵۰ سال قدمت دارد. تحقیقات ژنتیکی در طی این مدت، منجر به کشف وجود ژن، کروموزوم، DNA، RNA، ارتباط وراثت با هر یک از ژن‌ها، نقش این مواد در بدن موجودات زنده، ارتباط آن‌ها با فعالیت پروتئین‌ها در بدن و بسیاری مسائل دیگر شده است. با کشف این پدیده‌ها، مجددا مسائل جدیدی در حوزه زیست‌شناسی مطرح شدند. مسائلی از این دست که شجره‌نامه هر یک از موجودات زنده چیست، این موجودات از لحاظ وراثتی و ژنتیکی چه ارتباطی با هم دارند، هر یک از صفات ظاهر شده در یک موجود ناشی از وجود کدام ژن در بدن است، وراثت کدام صفات ناشی از عمل ژن‌های متعدد است، یک ژن چگونه ساختار فضایی پروتئین را تعیین می‌کند، ساختار فضایی پروتئین‌های مختلف به چه شکلی است و بسیاری مسائل دیگر. زیست‌شناسان و متخصصین ژنتیک در طی این سال‌ها توانسته‌اند حجم عظیمی از اطلاعات موجودات زنده را استخراج و ثبت کنند، اما تفسیر این اطلاعات در همه جا مشخص نیست. این که این اطلاعات چه معنایی دارند و در بدن موجود زنده چه نقشی بازی می‌کنند از جمله مسائل مهم علم ژنتیک‌اند. وظایف قسمت‌های مختلف ماده ژنتیکی به ساختاری که آن قسمت در داخل سلول به خود می‌گیرد بستگی دارد و به همین دلیل، تشخیص ساختار این قسمت‌ها به یکی از مسائل مهم پیش روی زیست‌شناسان تبدیل شده است. زیست‌شناسان در چند دهه اخیر با تک‌تک این مسائل

[1] Nuclear Magnetic Resonance
[2] RNA Secondary Structure
[3] Nussinov
[4] Zuker

دست و پنجه نرم کرده‌اند و در مورد برخی از این مسائل، به دلیل حجم بالای داده‌های مسئله و زمان‌بر بودن انجام کارهای آزمایشگاهی، به علوم کامپیوتر، آمار و ریاضیات روی آورده‌اند. این امر در دهه‌ی اخیر که ژنگان[1] بسیاری از موجودات زنده به طور کامل مشخص شده و یا در حال مشخص شدن است، بیش از پیش شدت گرفته است. تبدیل داده‌ها به مدل‌های کامپیوتری، طراحی الگوریتم‌ها برای بررسی داده‌ها و در نهایت استخراج نتایج با معنی و تفسیر آن‌ها از جمله خدمات رایجی هستند که علوم کامپیوتر برای علوم زیستی فراهم می‌آورد. علم حاصل از ادغام این دو شاخه‌ی علم، یعنی طراحی الگوریتم‌ها، توسعه‌ی برنامه‌ها و بانک‌های اطلاعاتی در جهت رشد و پیشرفت تحقیقات زیست‌شناسی را بیوانفورماتیک یا زیست‌شناسی محاسباتی می‌نامند.

در بین مسائل جالب مطرح‌شده در حوزه بیوانفورماتیک، پیش‌بینی ساختار RNA اخیرا مورد توجه قرار گرفته است. مولکول RNA از نظر زیست‌شناسی، مولکول مهمی به شمار می‌رود. در واقع اطلاعات موجود در DNA توسط این مولکول به پروتئین تبدیل می‌شود. یک مولکول RNA از نوکلئوتیدهای آدنین (A)، سیتوزین (C)، گوانین (G) و یوراسیل (U) ساخته می‌شود. این نوکلئوتیدها پس از ایجاد پیوند با یکدیگر، در کنار هم به صورت یک رشته درمی‌آیند که دو سر آن آزاد است. سپس بین برخی از این نوکلئوتیدها و تحت شرایط خاصی پیوندهای هیدروژنی برقرار شده و مولکول RNA یک ساختار فضایی به خود می‌گیرد. شکل فضایی این مولکول، نحوه عملکرد آن را مشخص می‌کند. مسئله‌ی پیش‌بینی ساختار RNA در حدود ۳۰ سال پیش مطرح شده و تا کنون تحقیقات زیادی بر روی آن انجام شده است. برای یک مولکول RNA سه نوع ساختار در نظر گرفته می‌شود. ساختار اول، یک دنباله از حروف (نوکلئوتیدها) است که واحدهای سازنده آن مولکول را مشخص می‌کند. ساختار دوم، یک شکل دوبعدی است که نحوه ایجاد پیوند بین واحدهای سازنده را مشخص می‌کند و ساختار سوم همان شکل فضایی مولکول است. شکل فضایی یک مولکول RNA را می‌توان با روش‌های آزمایشگاهی مانند NMR و یا کریستالوگرافی اشعه ایکس به طور دقیق مشخص کرد، ولی این کار مستلزم صرف زمان و هزینه‌ی بالایی است. گاهی ممکن است مشخص کردن شکل فضایی یک مولکول RNA با این روش‌ها، تا یک سال به درازا بکشد. به همین دلیل، استفاده از روش‌های محاسباتی برای پیش‌بینی شکل فضایی یک مولکول RNA مورد توجه قرار گرفته است. از جمله رویکردهای مهمی که برای حل این مسئله ارائه شده‌اند، رویکردهای مقایسه‌ای، رویکردهای کمینه‌سازی سطح انرژی و گرامرهای مستقل از متن تصادفی است. رویکرد مقایسه‌ای [۱] که در حال حاضر مطمئن‌ترین روش برای حل این مسئله است، اولین رویکردی بود که ارائه شد. رویکردهای کمینه‌سازی سطح انرژی از مفاهیم ترمودینامیکی مولکول‌ها استفاده می‌کنند. الگوریتم برنامه‌نویسی پویای نوسینف [٤، ۳، ۲] و الگوریتم برنامه‌نویسی پویای زوکر [۷، ٦، ٥] از مهم‌ترین رویکردهای این دسته هستند. الگوریتم نوسینف سعی در بیشینه کردن تعداد پیوندهای هیدروژنی در بین نوکلئوتیدهای RNA را دارد ولی این کار لزوما سطح انرژی یک مولکول RNA را به حداقل مقدار ممکن نمی‌رساند. در طرف دیگر، الگوریتم زوکر با استفاده از مفاهیم ترمودینامیک و پارامترهای انرژی مرتبط با مولفه‌های ساختاری مختلف که به صورت آزمایشگاهی به دست آمده‌اند [۹، ۸]، برای یک دنباله‌ی داده شده، ساختاری را مشخص می‌کند که دارای کمترین میزان انرژی باشد.

تعاریف اولیه و نمادگذاری

اگرچه DNA به صورت دو رشته‌ای حالت پایدار به خود می‌گیرد، RNA به صورت تک رشته‌ای در درون سلول ظاهر می‌شود. حالت نسبتا پایدار RNA به لطف پیوندهای هیدروژنی بین بازهای مکمل آن برقرار می‌شود. همانند DNA ، در RNA نیز بین بازهای مکمل A با U و G با C پیوند هیدروژنی برقرار می‌شود. علاوه بر این دو نوع پیوند، بین G با U نیز ممکن است پیوند هیدروژنی ضعیفی به صورت غیرمتعارف تشکیل شود.

تعریف ۱. مجموعه جفت بازهای کانونی[2] در توالی‌های *RNA* به صورت زیر تعریف می‌شود:

$$\Sigma^{BP} = \{A-U, U-A, C-G, G-C, G-U, U-G\}$$

ترتیب چیدمان بازها در توالی نوکلئوتیدهای RNA ، ساختارهای متفاوتی از RNA را در درون سلول به وجود

[1] Genome
[2] Canonical Base Paris
[1] Transitivity

می‌آورد. باتوجه به اینکه فعالیت RNA تابعی از ساختار آن است، لذا با استفاده از خاصیت تعدی[۱]، تنوع در فعالیت‌های RNA متاثر از ترتیب چیدمان توالی نوکلئوتیدهای RNA است. برای توالی‌های RNA سه نوع ساختار قابل تعریف است که به ترتیب ساختار اول، ساختار دوم و ساختار سوم نام‌گذاری می‌شوند. هرکدام از این ساختارها بخشی از اطلاعات شکل فضایی RNA را در بر می‌گیرد.

در ساختار اول RNA فقط ترتیب مؤلفه‌های سازنده‌ی RNA اهمیت دارد و معمولا این ساختار برای تشخیص الگوهایی ویژه در توالی RNA مورد استفاده قرار می‌گیرد. الگویابی همترازی یکی از این روش‌ها است که هدف آن یافتن الگوهای تکراری در توالی‌های RNA مشابه است. اگر یک مجموعه الفبای $\Sigma = \{A, C, G, U\}$ وجود داشته باشد توالی RNA به صورت زیر نمایش داده می‌شود:

$$R = r_۱ r_۲ r_۳ ... r_n \in \Sigma^n.$$

بنا به قرارداد، توالی RNA از کربن پنجم در سمت چپ توالی شروع شده و به کربن سوم در سمت راست توالی ختم می‌شود.

پیوندهای کانونی مستحکم‌ترین پیوندهای هیدروژنی هستند که بین نوکلئوتیدهای یک توالی RNA تشکیل می‌شوند. استحکام بالای این پیوندها باعث شده که به طور معمول فقط این نوع از پیوندها در یک توالی RNA مشاهده شوند. برای نمایش پیوند بین نوکلئوتید i ام و نوکلئوتید j ام یک توالی RNA از نمایش $i.j$ استفاده می‌شود.

تعریف ۲. اگر R نشان‌دهنده یک توالی *RNA* به طول n باشد، ساختار دوم آن، که با S نشان داده می‌شود، یک مجموعه از جفت‌های $i.j$ است که $i, j \in \{۱, ..., n\}$ و $i < j$ باشد. برای هر دو جفت $i_۱.j_۱$ و $i_۲.j_۲$ در S یکی از سه حالت زیر اتفاق می‌افتد:

۱) $i_۱ = i_۲ \iff j_۱ = j_۲$
۲) $i_۱ < i_۲ < j_۲ < j_۱$ (تو در تو)
۳) $i_۱ < j_۱ < i_۲ < j_۲$ (جدا از هم)

ساختار سوم RNA به شکل فضایی آن گفته می‌شود. با داشتن مختصات سه بعدی نوکلئوتیدها در ساختار سوم، شکل فضایی و نحوه اتصالات بین نوکلئوتیدها به طور کامل مشخص می‌شود. اطلاعات ساختار سوم به مراتب جامع‌تر از اطلاعات ساختار دوم است. ساختار سوم RNA به روش کریستالوگرافی اشعه ایکس به دست می‌آید. تصاویر کریستالوگرافی در اختیار محققین قرار گرفته و آنها با استفاده از تجربه و مهارتی که دارند، مختصات سه بعدی اجزای تشکیل دهنده توالی RNA را در یک قالب استاندارد ذخیره می‌کنند. با توجه به اینکه به دست آوردن این تصاویر پرهزینه، مشکل و زمان‌بر است، تعداد ساختارهای سومی که با این روش به دست آمده‌اند محدود هستند. بنابراین یکی از حوزه‌های تحقیقاتی که در این زمینه به وجود آمد، پیش‌بینی ساختار سوم از طریق انجام محاسبات بود. الگوریتم‌هایی که برای پیش‌بینی ساختارهای سوم طراحی شدند کارآمد نبوده و طراحی این الگوریتم‌ها نیز بسیار دشوار است. به همین علت و با توجه به سادگی ساختارهای دوم، این ساختارها در طراحی الگوریتم‌های بیوانفورماتیکی حضور پررنگ‌تری دارند.

با توجه به اهمیت ساختار دوم RNA ، مؤلفه‌هایی که معمولا در اغلب ساختارهای دوم وجود دارند نام‌گذاری شده‌اند. این مؤلفه‌ها شامل استم[۲]، هیرپین[۳]، حلقه داخلی[۴]، برآمدگی[۵]، چندحلقه[۶] و رشته‌های آویزان[۷] می‌باشند (شکل ۱) که در ادامه به تعریف آنها می‌پردازیم.

تعریف ۳. به توالی پیوندهای $i.j, (i+۱).(j-۱), ..., (i+\ell).(j-\ell)$ در ساختار دوم یک **استم** به طول ℓ گفته می‌شود.

به عبارت دیگر، به چند جفت نوکلئوتید که با هم به صورت متوالی پیوند ایجاد کرده‌اند استم می‌گویند. معمولا استم‌های با طول بیشتر، از پایداری بیشتری نیز برخوردار هستند.

تعریف ٤. اگر داشته باشیم $i.j$ و هیچ یک از نوکلئوتیدهای $i+۱$ تا $j-۱$ در هیچ پیوندی شرکت نکنند، به این ناحیه **هیرپین** گفته می‌شود.

[۲] stem
[۳] hairpin
[۴] internal loop
[۵] bulge
[۶] multi-loop
[۷] dangle

تعریف ۵. اگر یک ساختار دوم با پیوندهای $i_1.j_1$ و $i_2.j_2$ را به شرط $i_1 + 1 < i_2 < j_2 < j_1 - 1$ داشته باشیم و اگر همزمان نوکلئوتیدهای بین i_1 و i_2 و همین طور نوکلئوتیدهای بین j_2 و j_1 در هیچ پیوندی شرکت نکنند، به این دو ناحیه که در پیوند شرکت نکرده‌اند **حلقه داخلی** گفته می‌شود.

به زبان ساده‌تر، وقتی دو استم متوالی شکل می‌گیرد، به قسمتهایی از توالی RNA که بین این دو استم متوالی قرار گرفته‌اند و پیوندی بین آنها برقرار نشده است حلقه داخلی می‌گویند.

تعریف ۶. پیوند $i.j$، پیوند $p.q$ را **احاطه**[1] کرده است هرگاه داشته باشیم $i < p < q < j$.

تعریف ۷. اگر پیوند $i.j$ دو یا چند پیوند دیگر مانند $p.q, r.s, \cdots$ را احاطه کند، ولی این پیوندها یکدیگر را احاطه نکنند، می‌گوییم که یک **چندحلقه** شکل گرفته است.

به عبارت ساده‌تر، به ناحیه‌ای از توالی RNA که حداقل به سه استم ختم شود چندحلقه گفته می‌شود.

تعریف ۸. نوکلئوتیدهایی که با هیچ پیوندی احاطه نشده باشند **رشته های آویزان** نامیده می‌شوند.

در انجام فرآیندهای شیمیایی، تغییرات انرژی آزاد گیبس که به اختصار آن را با ΔG نشان می‌دهند، مورد بررسی قرار می‌گیرد. تغییرات انرژی آزاد شده در یک فرآیند شیمیایی، مانند تشکیل ساختار RNA، می‌تواند در تعیین جهت فرآیند کمک کند. چنانچه ΔG برابر صفر باشد، واکنش در هر دو جهت به صورت یکسان پیش می‌رود. اگر $\Delta G > 0$ آنگاه واکنش در جهت غیرمطلوب و اگر $\Delta G < 0$ آنگاه واکنش در جهت مطلوب حرکت می‌کند [۱۰]. تغییرات انرژی آزاد را معمولا به صورت فرمول زیر که تابعی از آنتالپی یا ΔH (میزان انرژی یک سیستم ترمودینامیکی برای تبادل با محیط)، دما یا T (معیاری برای تخمین میزان انرژی جنبشی سیستم) و آنتروپی یا ΔS (معیاری برای سنجش آشفتگی نسبی در سیستم) است، نشان می‌دهند:

$$\Delta G = \Delta H - T \times \Delta S$$

در فرآیندی مانند تشکیل ساختار RNA، محاسبه‌ی انرژی آزاد گیبس با استفاده از مفاهیم ذکر شده بسیار دشوار است و به جای آن می‌توان از روش‌های بیوفیزیکی جهت اندازه‌گیری تغییرات انرژی آزاد استفاده کرد. همچنین از آنجا که در فرآیند تشکیل ساختار RNA خصوصیات بیوفیزیکی تاثیرگذار هستند، به نظر می‌رسد روش‌هایی که این خصوصیات را در محاسبات خود لحاظ می‌کنند از دقت بالایی برخوردار باشند. برای محاسبه‌ی میزان انرژی آزاد یک RNA که یک ساختار خاص را به خود گرفته است، می‌توان از مدل نزدیک‌ترین همسایه استفاده کرد. این روش در برنامه‌های RNAeval از بسته‌ی نرم‌افزاری وینا[2] [۱۱] و RNAstructure از گروه تورنر[3] [۹] نیز به کار رفته است. در مدل نزدیک‌ترین همسایه، هر مولفه از ساختار دوم، میزان انرژی خود را در انرژی کلی ساختار دخالت می‌دهد. به عبارت دیگر، انرژی هر ساختار در این مدل برابر است با مجموع انرژی مولفه‌های مختلف آن که به صورت زیر نوشته می‌شود:

$$\Delta G = \Delta G_{Stems} + \Delta G_{Hairpins} + \Delta G_{Bulges} + \Delta G_{Internals} + \Delta G_{Multiloops} + \Delta G_{External} \quad (۱)$$

پیش‌بینی ساختار دوم RNA

همان‌طور که گفته شد، توالی RNA از منظر زیست‌شناسی بسیار مهم بوده و در فرآیندهای سلولی نقش مهمی را ایفا می‌کند. با توجه به اهمیت موضوع و لزوم استفاده از روش‌های محاسباتی، مسئله‌ی پیش‌بینی ساختار دوم RNA در حدود ۳۰ سال پیش توسط لوینتال[4] مطرح شد و تا کنون روش‌های زیادی برای حل آن ارائه شده‌اند. لازم به ذکر است تمامی روش‌های محاسباتی ارائه شده معمولا تقریبی از ساختار فضایی RNA را پیدا می‌کنند.

رویکرد مقایسه‌ای

رویکرد مقایسه‌ای، که در حال حاضر مطمئن‌ترین روش برای حل مسئله پیش‌بینی ساختار دوم RNA است [۱]، اولین رویکردی بود که برای حل این مسئله ارائه شد. در این رویکرد، تعدادی رشته RNA هم‌خانواده که ساختار دوم

[1] Surround
[2] Vienna RNA Package
[3] Turner
[4] Levinthal

بعضی از آن‌ها معلوم است، هم‌تراز شده و از نتیجه‌ی این هم‌ترازی برای تعیین ساختار دوم رشته‌هایی که ساختار دوم آن‌ها مشخص نیست استفاده می‌شود. دلیل استفاده از چنین رویکردی این است که ساختار فضایی مولکول‌های RNA کارکرد آنها را تعیین می‌کند و معمولا ساختار در طول تکامل حفظ شده است. به عنوان مثال، احتمال دارد نوکلئوتیدهای یک توالی RNA در طول نسل‌ها بر اثر جهش و یا هر دلیل دیگری تغییر کنند، ولی اگر جهش در نوکلئوتیدها ساختار فضایی آن را تغییر ندهد مشکلی در روند کلی کارکرد این RNA ایجاد نمی‌شود. محققان با همین استدلال و با کمک گرفتن از ساختار دوم چند رشته‌ی RNA هم‌خانواده که ساختار دوم آنها معلوم است و با تعیین میزان تشابه میان این رشته‌ها و رشته‌هایی که ساختار دوم آن‌ها معلوم نیست، سعی می‌کنند ساختار دوم رشته‌های هم‌خانواده را حدس بزنند. از جمله نقاط قوت این رویکرد، دقت بالای آن در پیش‌بینی ساختار دوم است و از نقاط ضعف این رویکرد می‌توان به محدود بودن آن اشاره کرد؛ به این معنی که برای پیش‌بینی ساختار دوم یک رشته RNA باید حداقل یک رشته، که از همان خانواده بوده و ساختار دوم آن معلوم است، وجود داشته باشد.

رویکرد کمینه‌سازی سطح انرژی

در رویکرد کمینه‌سازی سطح انرژی از مفاهیم ترمودینامیکی مولکول‌ها استفاده می‌شود. همان‌طور که می‌دانیم هر مولکول در حالت عادی پیوندهای درون‌مولکولی خود را به گونه‌ای ایجاد می‌کند که سطح انرژی خود را به پایین‌ترین سطح ممکن برساند. بنابراین می‌توان با شناسایی پیوندهای هیدروژنی بین نوکلئوتیدهای رشته RNA و انتخاب هوشمندانه برخی از آن‌ها، سطح انرژی آن را کمینه کرد. در هنگام کار با مفاهیم ترمودینامیکی به منظور حل مسئله‌ی پیش‌بینی ساختار دوم RNA استفاده از دو تابع زیر بسیار مفید خواهد بود:

- $E(R,S)$: انرژی آزاد دنباله‌ی R هنگامی که ساختار دوم آن S باشد (با توجه به مدل نزدیک‌ترین همسایه برای محاسبه انرژی).
- $\phi(R)$: برای دنباله R، این تابع ساختار دومی مانند S که دارای کمترین انرژی آزاد است را محاسبه می‌کند. به عبارت دیگر:

$$\phi(R) = Argmin_S E(R,S).$$

با توجه به دو تابع بالا، مسئله پیش‌بینی ساختار دوم یک RNA مانند R از منظر مفاهیم ترمودینامیکی معادل است با به دست آوردن مقدار تابع $\phi(R)$. از جمله نقاط قوت این رویکرد این است که فقط با داشتن توالی RNA می‌توان ساختار دوم آن را پیش‌بینی کرد. از نقاط ضعف این رویکرد این است که تا کنون هیچ مدل ترمودینامیکی کاملی که بتواند میزان انرژی آزاد هر ساختار دومی را محاسبه کند ارائه نشده است؛ یعنی قوانین ترمودینامیکی ارائه شده در حال حاضر نمی‌توانند میزان انرژی آزاد یک ساختار دوم را به صورت دقیق محاسبه کنند.

الگوریتم نوسینف

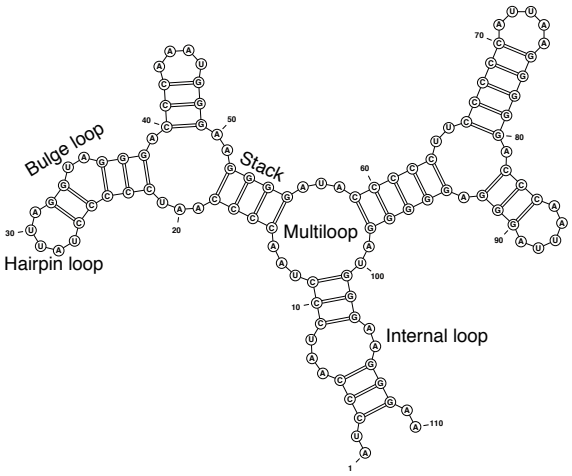
این الگوریتم، ساده‌ترین الگوریتم موجود در زمینه‌ی پیش‌بینی ساختار دوم RNA است که ساختار دوم (بدون شبه‌گره‌ها) را بدون در نظر گرفتن قوانین ترمودینامیک حل می‌کند؛ ولی علی‌رغم این شرایط، این الگوریتم پایه و اساس سایر الگوریتم‌های ارائه شده برای حل این مسئله به شمار می‌رود. در این الگوریتم ابتدا با استفاده از روابط بازگشتی، یک ماتریسی محاسبه شده و سپس از روی این ماتریس، ساختار دوم به دست می‌آید.

پر کردن ماتریس. همان‌طور که گفته شد، این روش یک دنباله اولیه RNA، یعنی دنباله‌ای از نوکلئوتیدها، را به عنوان ورودی می‌گیرد و سعی می‌کند با استفاده از روابط بازگشتی، بیشینه تعداد جفت‌های ممکن برای آن دنباله را پیدا کند؛ در واقع ساختاری را به دست می‌آورد که اگر دنباله‌ی ورودی، آن ساختار را به خود بگیرد، بیشترین تعداد جفت‌های ممکن را تشکیل داده است. رابطه‌ای که نوسینف از آن استفاده می‌کند در زیر آورده شده است:

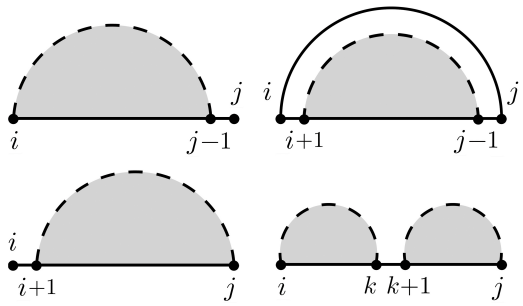
$$S(i,j) = max \begin{cases} S(i+1,j) \\ S(i,j-1) \\ S(i+1,j-1)+1 \\ max_{i<k<j}\ S(i,k)+S(k+1,j) \end{cases} \quad (۲)$$

در رابطه‌ی بالا S یک ماتریس $n \times n$ است که n، طول دنباله RNA می‌باشد. در ابتدا درایه[1]های روی قطر اصلی و زیر آن، صفر در نظر گرفته می‌شوند و فقط باید درایه‌های بالای قطر اصلی را با استفاده از رابطه‌ی بالا پیدا کرد. در رابطه‌ی بالا $S(i,j)$ نشان‌دهنده درایه سطر iام و ستون jام ماتریس است که بیشترین تعداد جفت‌های ممکن را برای زیردنباله‌ای از دنباله‌ی RNA که به نوکلئوتیدهای iام و jام

[1] Entry



شکل ۱: مؤلفه‌های مختلف ساختار دوم RNA .



شکل ۲: حالت‌های بررسی شده در رابطه‌ی بازگشتی نوسینف.

محدود شده است، در خود ذخیره می‌کند. در این رابطه چهار حالت در نظر گرفته شده‌اند. حالت اول، حالتی است که در آن، نوکلئوتید iام، آزاد است و حالت دوم، حالتی را بیان می‌کند که در آن، نوکلئوتید jام، به صورت آزاد در نظر گرفته شده است. سومین حالت، وضعیتی را بیان می‌کند که در آن، نوکلئوتیدهای iام و jام با یکدیگر پیوند تشکیل داده‌اند. اگر این دو نوکلئوتید نتوانند با یکدیگر جفت شوند (مکمل یکدیگر نباشند)، این حالت در محاسبه $S(i,j)$ در نظر گرفته نمی‌شود. در حالت آخر، نوکلئوتیدهای iام و jام با یکدیگر پیوندی تشکیل نمی‌دهند ولی ممکن است با بازهایی که در بین آن‌ها وجود دارند جفت شوند. این چهار حالت ممکن بررسی شده و مقدار بیشینه آنها به عنوان مقدار نهایی $S(i,j)$ در نظر گرفته می‌شود. شکل ۲ حالت‌های این رابطه بازگشتی را به صورت شماتیک نشان می‌دهد. در این شکل از نمایش فاین‌من[1] خطی برای به تصویر کشیدن رابطه بازگشتی الگوریتم نوسینف استفاده شده است. خطوط افقی، بیانگر دنباله RNA و دایره‌های کوچک توپر، بیانگر نوکلئوتیدها هستند. کمان‌های پیوسته، نشان‌دهنده‌ی وجود پیوند بین دو نوکلئوتید دو طرفشان هستند. کمان‌های خط‌چین، بیانگر این هستند که رابطه‌ی بین نوکلئوتیدهای دو طرف آن‌ها نامشخص است و وضعیت آن‌ها باید با استفاده از رابطه‌ی بازگشتی مشخص شود. نواحی خاکستری، ناحیه‌هایی هستند که وضعیت آن‌ها باید مشخص شود و نواحی سفید، ناحیه‌هایی هستند که وضعیتشان مشخص شده و هیچ پیوندی در داخل آن‌ها رخ نداده است. با توجه به رابطه بازگشتی ارائه شده، ترتیب محاسبه درایه‌های ماتریس از قطر بالای قطر اصلی به سمت بالاترین قطر و در هر قطر از پایین‌ترین درایه به سمت بالاترین درایه می‌باشد.

برگشت روی ماتریس. پس از پر کردن ماتریس، الگوریتم نوسینف، شروع به برگشتن روی ماتریس و پیدا کردن مسیری برای ساخت ساختار دوم می‌کند. برای این منظور، اولین درایه‌ای که کار با آن شروع می‌شود درایه $S(1,n)$ می‌باشد، یعنی آخرین درایه محاسبه شده در ماتریس و در واقع درایه‌ای که مقدار آن برابر است با بیشترین تعداد ممکن پیوند در رشته RNA داده شده. پس از این درایه به درایه‌ای حرکت می‌کند که با استفاده از آن به درایه‌ی فعلی رسیده است، یعنی درایه‌ای که بیشترین مقدار را در رابطه‌ی نوسینف برای درایه فعلی ایجاد کرده است. به عنوان مثال اگر در محاسبه $S(1,n)$، بین چهار حالت مذکور، $S(2,n)$ مقدار بیشینه بوده باشد، الگوریتم به درایه $S(2,n)$ بازگشت می‌کند و کار را از آنجا ادامه می‌دهد. این کار تا زمانی انجام می‌شود که به قطر اصلی برسیم و سپس با توجه به مسیری که پیمایش شده است، ساختار دوم RNA داده شده به دست می‌آید. بدیهی است که ممکن است برای یک دنباله RNA داده شده، چندین ساختار دوم به دست آیند.

پیچیدگی الگوریتم نوسینف از مرتبه $O(n^3)$ است. این الگوریتم، حلقه‌های هیرپین با اندازه کمتر از سه را نیز در

[1] Feynman

نظر می‌گیرد. برای این که الگوریتم نوسینف، قادر به تولید هیرپین‌های با طول کمتر از سه نباشد، کافی است درایه‌های سه قطر بالای قطر اصلی نیز با عدد صفر مقداردهی شوند و گام پر کردن ماتریس از چهارمین قطر بالای قطر اصلی آغاز شود. مشخص است که در این صورت، باید گام برگشت روی ماتریس نیز تا رسیدن به سومین قطر بالای قطر اصلی ادامه یابد. چون الگوریتم نوسینف، قوانین ترمودینامیک و بحث‌های مربوط به انرژی را در نظر نمی‌گیرد، امروزه خیلی قابل اعتماد نیست ولی به عنوان یک نقطه‌ی شروع خوب از آن نام برده می‌شود. دقت این الگوریتم در پیش‌بینی ساختار دوم RNA حدود ۷۰ درصد است.

الگوریتم زوکر

این الگوریتم نیز همانند الگوریتم نوسینف از برنامه‌نویسی پویا استفاده می‌کند، با این تفاوت که الگوریتم زوکر، به دنبال ساختاری است که دنباله داده شده بر روی این ساختار، دارای کمترین مقدار انرژی آزاد باشد. برای راحتی کار، در اینجا الگوریتم زوکر با این فرض توضیح داده می‌شود که حلقه‌های چندشاخه دارای انرژی صفر هستند، همان‌طور که در [۷] آورده شده است. در ادامه، $S(i,j)$ عبارت است از زیردنباله‌ای از دنباله‌ی S که به نوکلئوتیدهای i و j محدود شده است.

الگوریتم زوکر از دو ماتریس $n \times n$ به نام‌های V و W استفاده می‌کند. برای هر i و j که $۱ \leq i < j \leq n$، $W(i,j)$ برابر است با انرژی آزاد کمینه زیردنباله $S(i,j)$ و $V(i,j)$ برابر است با انرژی آزاد کمینه زیردنباله $S(i,j)$، با این فرض که نوکلئوتیدهای i و j با یکدیگر جفت شده باشند. اگر نوکلئوتیدهای i و j نتوانند با یکدیگر جفت شوند (مکمل یکدیگر نباشند)، آنگاه $V(i,j) = \infty$ در نظر گرفته می‌شود. درایه‌های ماتریس‌های V و W به صورت بازگشتی محاسبه شده و ترتیب محاسبه درایه‌های این دو ماتریس همانند ترتیب موجود در الگوریتم نوسینف است. در ابتدا، برای هر i و j که $|j-i| = ٤$، قرار می‌دهیم $W(i,j) = ۰$، چرا که زیردنباله‌های به طول پنج، هیچ ساختار پایداری را تشکیل نمی‌دهند. از طرفی، مقدار V برای همین i و j برابر است با انرژی آزاد هیرپینی که با پیوند فرضی بین این دو نوکلئوتید، بسته می‌شود. اگر $|j-i| > ٤$، آنگاه مقادیر $V(i,j)$ و $W(i,j)$ به صورت بازگشتی و با توجه به مقادیر درایه‌های محاسبه شده‌ی قبلی به دست می‌آیند. فرض کنید $FH(i,j)$ بیانگر هیرپینی باشد که با پیوند بین نوکلئوتیدهای i و j محدود شده است. همچنین فرض کنید $FL(i,j,i',j')$ مولفه‌ای باشد که بین دو پیوند (i,j) و (i',j')، که $i < i' < j' < j$، به وجود آمده است. این مولفه می‌تواند استک ($i' = i + ۱$ و $j' = j - ۱$)، برآمدگی ($i' = i + ۱$ یا $j' = j - ۱$) و یا حلقه داخلی ($i' > i + ۱$ و $j' < j - ۱$) باشد. با توجه به این تعاریف، مقدار $V(i,j)$ به صورت زیر محاسبه می‌شود:

$$V(i,j) = min \begin{cases} E(FH(i,j)) \\ min_{i<i'<j'<j} E(FL(i,j,i',j')) + V(i',j') \\ min_{i+۱<k<j-۲} W(i+۱,k) + W(k+۱,j-۱) \end{cases} \quad (۳)$$

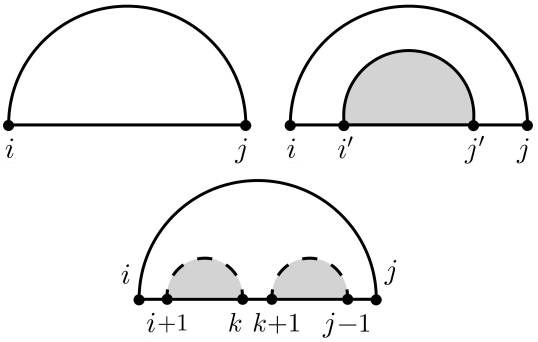
در رابطه‌ی بالا، حالت اول بیانگر این است که نوکلئوتیدهای i و j به عنوان پیوند انتهایی یک هیرپین در نظر گرفته شده‌اند. حالت دوم، تمامی بالج‌ها، حلقه‌های داخلی و استک‌هایی را در نظر می‌گیرد که پیوند اول آن‌ها (i,j) است. حالت سوم، حالتی است که در آن پیوند (i,j) پیوند انتهایی یک حلقه‌ی چندشاخه باشد. حالت‌های مختلف این رابطه بازگشتی در شکل ۳ به صورت شماتیک نشان داده شده است.

برای محاسبه‌ی مقدار $W(i,j)$ نیز باید سه حالت را در نظر گرفت. یا یکی از نوکلئوتیدهای i و j در هیچ پیوندی شرکت نمی‌کند، یا این دو نوکلئوتید با یکدیگر جفت می‌شوند و یا با یکدیگر جفت نمی‌شوند ولی در پیوند دیگری شرکت خواهند داشت. حالت اول را می‌توان با مقادیر $W(i+۱,j)$ و $W(i,j-۱)$ معادل کرد. توجه داشته باشید که با توجه به روابط بازگشتی، حالتی که در آن هیچ کدام از این دو نوکلئوتید در پیوندی شرکت نمی‌کنند نیز در حالت اول مورد بررسی قرار می‌گیرد. برای حالت دوم، $W(i,j) = V(i,j)$. برای حالت سوم باید یک حلقه چندشاخه را در نظر گرفت. پس رابطه‌ی بازگشتی مورد نظر به صورت زیر خواهد بود:

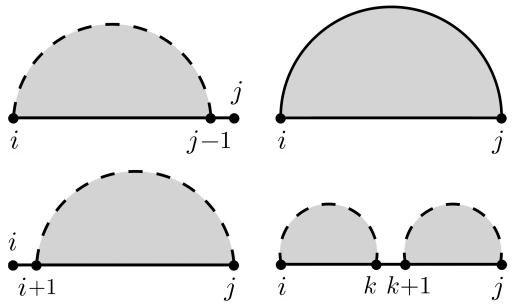
$$W(i,j) = min \begin{cases} W(i+۱,j) \\ W(i,j-۱) \\ V(i,j) \\ min_{i<k<j-۱} W(i,k) + W(k+۱,j) \end{cases} \quad (٤)$$

حالت‌های مختلف این رابطه بازگشتی نیز در شکل ٤ به صورت شماتیک نشان داده شده است.

پس از محاسبه‌ی تمامی درایه‌های این دو ماتریس، مقدار $W(۱,n)$ برابر با انرژی آزاد کمینه دنباله RNA داده شده به طول n است. برگشت روی این ماتریس‌ها برای به دست آوردن ساختار مورد نظر، همانند گام دوم الگوریتم نوسینف انجام می‌شود. یعنی از $W(۱,n)$ شروع کرده و به درایه‌ای از ماتریس W یا ماتریس V، می‌رود که این درایه را کمینه کرده است و سپس عمل برگشت از آنجا ادامه می‌یابد.



شکل ۳: حالت‌های مختلف برای محاسبه $V(i,j)$.



شکل ٤: حالت‌های مختلف برای محاسبه $W(i,j)$.

پیچیدگی این الگوریتم از مرتبه $O(n^4)$ است (حالت دوم رابطه بازگشتی W شامل چهار متغیر آزاد می‌باشد). معمولا برای کاهش پیچیدگی این الگوریتم فرض می‌شود که حداکثر طول بالج‌ها و حلقه‌های داخلی برابر با عدد ثابت (مثلا ۳۰ نوکلئوتید) بوده و لذا یکی از متغیرهای آزاد در حالت دوم رابطه بازگشتی حذف شده و مرتبه زمانی این الگوریتم به $O(n^3)$ کاهش می‌یابد. با این وجود، الگوریتم زوکر به دلیل انجام محاسبات زیاد، نسبت به الگوریتم نوسینف کندتر است، اما به دلیل استفاده از قوانین ترمودینامیک، دقت آن از دقت الگوریتم نوسینف خیلی بیشتر است. الگوریتم زوکر، حلقه‌های هیرپین با اندازه‌ی کمتر از سه را هیچ‌گاه به وجود نمی‌آورد، چرا که این مولفه‌ها، با توجه به مقادیری که به صورت تجربی برای انرژی مولفه‌های ساختاری به دست آمده‌اند، پایدار نیستند. این در حالی است که در طبیعت، ساختارهایی وجود دارند که دارای هیرپین‌های با اندازه‌ی کمتر از سه نیز هستند. البته در آینده ممکن است پارامترهای انرژی به گونه‌ای به‌روز شوند که امکان تولید این ساختارها نیز توسط الگوریتم زوکر وجود داشته باشد. از جمله پیاده‌سازی‌های پیشرفته‌ی این الگوریتم می‌توان به rnafold از بسته‌ی نرم‌افزاری mfold [۱۲] و RNAfold از بسته‌ی نرم‌افزاری وینا [۱۳] اشاره کرد.

جمع‌بندی

در این مقاله، ساختارهای مختلف توالی‌های RNA و اهمیت این ساختارها در عملکرد توالی‌های RNA ارائه شد. با توجه به اهمیت این گونه ساختارها و مشکلات موجود در روش‌های آزمایشگاهی برای تعیین این ساختارها، به نظر می‌رسد روش‌های محاسباتی می‌توانند جایگزین مناسبی برای پیش‌بینی آنها باشند. در این خصوص دو الگوریتم پایه‌ای معروف به همراه روابط بازگشتی آنها برای پیش‌بینی ساختارهای دوم RNA به طور کامل شرح داده شد. همچنین دو پیاده‌سازی مختلف از الگوریتم زوکر نیز معرفی شد تا علاقمندان بتوانند با استفاده از آنها ساختارهای دوم را پیش‌بینی کنند. طی چند سال اخیر و با افزایش توان محاسباتی کامپیوترها، الگوریتم‌هایی برای پیش‌بینی ساختارهای سوم RNA با مؤلفه‌های ساختاری محدود ارائه شده است. پیش‌بینی ساختار سوم RNA متعلق به رده مسائل NP–سخت بوده و به راحتی قابل حل نخواهد بود. اما نظر به اهمیت شکل فضایی در کارکرد RNA و اهمیت فزاینده انواع RNA های رمزگذار و غیر رمزگذار در فرایندهای زیستی، این زمینه مطالعاتی تا سال‌ها به عنوان زمینه پژوهشی چالش برانگیز مورد توجه خواهد بود.

References

[1] I. L. Hofacker, M. Fekete, and F. P. Stadler. Secondary structure prediction for aligned rna sequences. *Journal of Molecular Biology*, 319(5):1059–1066, 2002.

[2] R. Nussinov, E. Comay, and O. Comay. An accelerated algorithm for calculating the secondary structure of single stranded rnas. *Nucleic Acids Research*, 12:53–66, 1984.

[3] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proceedings of the National Academy of Sciences of the USA*, 77(11):6309–6313, 1980.

[4] R. Nussinov and I. Tinoco. Sequential folding of a messenger rna molecule. *Journal of Molecular Biology*, 151:519–533, 1981.

[5] A. B. Jacobson, L. Good, J. Simonetti, and M. Zuker. Some simple computational methods to improve the folding of large rnas. *Nucleic Acids Research*, 12:45–52, 1984.

[6] M. Zuker and D. Sankoff. Rna secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46:591–621, 1984.

[7] M. Zuker and P. Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[8] D. H. Turner and D. H. Mathews. Nndb: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Website, 2004. `http://rna.chem.rochester.edu`.

[9] D. H. Turner and D. H. Mathews. Nndb: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38:D280–D282, 2009.

[10] H. H. Tsang. *SARNA-Predict: A permutation Based Simulated Annealing Algorithm For RNA Secondary Structure Prediction*. PhD thesis, 2007.

[11] A. Gruber, R. Lorenz, S. Bernhart, R. Neubock, and I. Hofacker. The vienna rna websuite. *Nucleic Acids Research*, 38, 2008.

[12] M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for rna secondary structure prediction: A practical guide. Website, 1999. `http://mfold.rna.albany.edu/`.

[13] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Vienna rna package: Rna secondary structure prediction and comparison. Website, 1994. `http://www.tbi.univie.ac.at/~ivo/RNA/`.

Computational Methods for RNA Secondary Structure Prediction

Mohammad Ganjtabesh

School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran

mgtabesh@ut.ac.ir

Abstract. The function of an RNA is mostly related to it tertiary structure. These structures could be precisely determined by NMR or X-ray crystallography techniques, but it is very expensive and time consuming. Therefore, using the computational methods for predicting the RNA tertiary structure become attractive. Since the prediction of RNA tertiary structure is very complex and computationally inefficient, most of the researches are focus on the problems related to RNA secondary structures. This kind of structures could be described by a set of paired location in an RNA sequence. This problem has been introduced almost 30 years ago and many researches have been down to solve it. Minimizing the free energy is an important approach to attach it and based on this approach Nossinov and Zuker algorithms are devised. In this paper, basic concepts related to the RNA structures are introduced and the mentioned two algorithms are presented and analyzed in details.

Keywords: RNA structure, Minimum free energy, RNA sequence