پیشگویی دقیق ساختار دوم RNA مبتنی بر الگوریتم ژنتیک

سهیلا منتصری[۱]، نصرالله مقدم-چرکری[۲]* و فاطمه زارع میرک آباد[۳]

[۱] **تهران، دانشگاه تربیت مدرس، دانشکده علوم ریاضی، گروه علوم کامپیوتر**

[۲] **تهران، دانشگاه تربیت مدرس، دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر**

[۳] **تهران، دانشگاه صنعتی امیرکبیر، دانشکده ریاضی و علوم کامپیوتر، گروه علوم کامپیوتر**

تاریخ دریافت: ۹۰/۹/۱ تاریخ پذیرش: ۹۱/۵/۴

چکیده

مولکول RNA نقش مهم و اساسی در فرآیندهای زیستی ایفاء می‌کند. در بیشتر مواقع، عملکرد RNAها توسط ساختار آنها مشخص می‌شود. با توجه به پیچیدگی و هزینه بر بودن روشهای آزمایشگاهی برای پیشگویی ساختار RNAها، از روشهای محاسباتی استفاده می‌گردد. الگوریتمهای متنوعی جهت پیشگویی ساختار دوم مولکول RNA وجود دارد. در این مقاله، یک الگوریتم ژنتیک به نام RNAG جهت پیشگویی ساختار دوم مولکول RNA براساس حداقل انرژی آزاد ارائه می‌شود. در این الگوریتم، هر فرد از جمعیت شامل تعدادی ساقه می‌باشد. افراد براساس مقدار برازندگی حداقل انرژی آزاد شده از ساقه‌ها و حلقه‌ها به‌ترتیب صعودی رتبه‌بندی شده و در ادامه به‌ترتیب عملگرهای تقاطع و جهش روی آنها برای ایجاد نسل بعد اجرا می گردد. فرآیند تولید نسل تا زمان تولید یک فرد با حداقل انرژی آزاد مناسب ادامه می‌یابد. در پایان این فرد به‌عنوان ساختار دوم بهینه در نظر گرفته می‌شود. الگوریتم پیشنهادی روی تعدادی از RNAها در *باکتریها* اجرا می‌گردد. نتایج حاصل از این تحقیق نشان می‌دهد که الگوریتم RNAG در مقایسه با سایر روشهای مشابه دارای دقت بسیار بالا است.

واژه های کلیدی: حداقل انرژی آزاد، ساقه، مقدار برازندگی.

* نویسنده مسئول، تلفن: ۸۲۸۸۳۳۰۱، پست الکترونیکی: charkari@modares.ac.ir

مقدمه

مولکولهای RNA در تمام موجودات زنده دارای نقش حیاتی هستند. شناخت ساختار RNA در درک فعالیت آن اهمیت فراوانی دارد (۸). ساختار مولکولهای RNA در بیان ژن، پیرایش RNAهای پیک (Messenger RNA)، ساخت پروتئین و عملکردهای زیستی دیگر مؤثر است (۱،۲و۱۱). به‌عنوان مثال، خاتمه رونویسی (Transcription) بعضی از ژنها در *باکتری*، براساس ساختار سنجاق‌سری انتهایی RNA پیک انجام می‌شود (۱۴). از نکات مهم در پیشگویی ساختار دوم RNA می‌توان به دنباله‌هایی اشاره کرد که هنوز ساختار آنها از راه آزمایش مشخص نشده و در نتیجه هیچ نظیری در پایگاه داده برای آنها نمی‌توان یافت. از اهداف مهم پیشگویی ساختارها حل این مشکل است (۱۶). موضوع دیگری که کاربرد مهمی جهت طراحی ساختارهای RNA دارد، برهمکنش دو مولکول RNA است (۱۳). پیشگویی ساختارهای RNA مقدمه‌ای در تعیین ساختار برهمکنش دو RNA می‌باشد.

تلاشهایی جهت پیشگویی ساختار دوم مولکولهای RNA با بیشینه کردن تعداد جفت‌بازها (Base pairs)، با استفاده از برنامه‌نویسی پویا انجام شد که در آن بهترین ساختار برای هر زیردنباله محاسبه می‌گردد (۱۰). پس از آن الگوریتم مشابهی ارائه شد که در آن از مقادیر انرژی آزاد جفت‌بازها برای محاسبه ساختاری با کمترین انرژی آزاد (Minimum

free energy) استفاده می‌شود (۹ و ۱۸). در یک روش دیگر برای پیشگویی ساختار دوم، توابع تسهیم (Partition function) مولکولهای RNA براساس برنامه نویسی پویا محاسبه می‌گردند (۷). ابزار MFold (۱۹) توسط پارامترهای موجود برای محاسبه ساختار دوم RNA (۱۵)، به پیشگویی ساختار دوم می‌پردازد. در تعدادی از رویکردها، انرژی آزاد با استفاده از مدل ترمودینامیکی نزدیکترین همسایه تعیین می‌شود. در این مدلها، انرژی آزاد ساختار به‌عنوان مجموع انرژیهای آزاد شده از هر ساقه (Stem) و حلقه با استفاده از داده‌های ترمودینامیکی محاسبه می‌گردد (۶ و ۱۷). روشی براساس گرامرهای مستقل ازمتن ارائه شد که در آن از الگوریتمهای آماری برای ایجاد ساختار دوم استفاده می‌شود (۱۲). ابزار RNAFold (۳) با استفاده از پارامترهای انرژی ایجاد شده (۵) ساختار دوم RNA را پیشگویی می‌کند.

در این مقاله، یک الگوریتم ژنتیک به نام RNAG جهت پیشگویی دقیق ساختار دوم مولکول RNA ارائه می‌شود. در این الگوریتم، یک ماتریس نقطه‌ای ایجاد می‌گردد که نشان‌دهنده تمام جفت‌بازهای ممکن در RNA است. هر زیرقطر در ماتریس نقطه ای را می‌توان به‌عنوان یک ساقه در نظر گرفت. سپس جمعیتی از ساقه‌هایی که به طور تصادفی انتخاب می‌شوند، ایجاد شده و مقدار برازندگی (Fitness value) حداقل انرژی آزاد شده از ساقه‌ها و حلقه‌ها برای هر فرد موجود در جمعیت محاسبه می‌گردد. برای ایجاد نسل جدید، عملگرهای تقاطع (Crossover) و جهش (Mutation) به‌ترتیب روی تعدادی از افراد نسل جاری انجام می‌شود. فرآیند تولید نسل ادامه می‌یابد تا زمانی که انرژی آزاد فردی به حد مطلوب برسد. در نهایت، این فرد با حداقل انرژی آزاد جهت تشکیل ساختار دوم RNA انتخاب می‌شود. الگوریتم پیشنهادی روی تعدادی از داده‌ها شامل *CopA*، *CopT*، *R1inv*، *R2inv*، *Tar*، *Tar**، *DIS*، $IncRNA_{54}$ و *RepZ* در *باکتریها* به کار رفته است. نتایج حاصل از این تحقیق نشان می‌دهد که الگوریتم RNAG در مقایسه با سایر روشهای مشابه دقت بالایی دارد.

مواد و روشها

پایگاه داده: RNAهای مورد بررسی در این مقاله شامل *CopA*، *CopT*، *R1inv*، *R2inv*، *Tar*، *Tar**، *DIS*، $IncRNA_{54}$ و *RepZ* هستند (۴).

تعاریف پایه: دنباله RNA از چهار نوع نوکلئوتید تشکیل می‌شود که شامل آدنین (A)، گوانین (G)، سیتوزین (C) و یوراسیل (U) است. هر RNA دارای دو انتهای مجزا است که به‌عنوان ′۳ و ′۵ شناخته می‌شوند. یک دنباله RNA به نام R ، $|R| = n$ در جهت ′۵ به ′۳ به صورت زیر تعریف می‌گردد:

$$\boldsymbol{R = r_1 r_2 \ldots r_n : \forall i\ (1 \le i \le n) r_i \in \{A, C, G, U\}.}$$

معکوس $\boldsymbol{R}$ با $\boldsymbol{r_n\ r_{n-1} \ldots\ r_1}$ در جهت ′۳ به ′۵ مشخص می شود. بنابراین $\boldsymbol{r_{ij} = r_i r_{i+1} \ldots r_j}$ زیردنباله‌ای از $\boldsymbol{R}$ است که از موقعیت $\boldsymbol{i}$ شروع شده و به موقعیت $\boldsymbol{j}$ ختم می‌گردد. دنباله RNA با تشکیل پیوند هیدروژنی بین بازهای آن تشکیل ساختار می‌دهد. بیشتر پیوندها بین بازهای مکمل واتسون–کریک روی می‌دهند که در آنها $\boldsymbol{G}$ با $\boldsymbol{C}$ و $\boldsymbol{A}$ با $\boldsymbol{U}$ جفت می‌شوند و برعکس. این پیوندها می‌توانند ساختار دوم RNA را تشکیل دهند.

ساختار دوم RNA از ساقه‌ها و نواحی منفرد (Single regions) تشکیل می‌شود. هر ساقه مجموعه‌ای از جفت-بازهای مجاور مانند $(\boldsymbol{r_i, r_j})$ و $(\boldsymbol{r_{i'}, r_{j'}})$ است به طوری که $\boldsymbol{i, i', j, j'}$ توسط یکی از شرایط زیر ارضاء می گردند:

$$\boldsymbol{i < i' < j' < j}$$
$$\boldsymbol{i' < i < j < j'}$$

به فرض اینکه $\boldsymbol{r_{ij}}$ و $\boldsymbol{r_{kl}}$ دو زیردنباله از RNA باشند که تشکیل ساقه می‌دهند. بنابراین زیردنباله $\boldsymbol{r_{ij}}$ به معکوس $\boldsymbol{r_{kl}}$ متصل می‌شود. به‌عبارت دیگر، بین هریک از بازهای $\boldsymbol{r_{ij}}$ و معکوس $\boldsymbol{r_{kl}}$ به‌ترتیب پیوند هیدروژنی برقرار می‌گردد.

برای نشان دادن ساقه در ساختار دوم RNA، هر باز در r_{ij} با $'('$ و هر باز در r_{kl} با $')'$ مشخص می‌شود. نواحی منفرد به‌عنوان حلقه یا تک‌رشته‌ای شده (Single-stranded) در نظر گرفته می‌شوند. لازم به ذکر است که دو انتهای هر ناحیه حلقه به ساقه‌ها متصل می گردند در حالی که تنها یک انتهای هر تک‌رشته‌ای شده به یک ساقه پیوند می خورد. هر باز در نواحی منفرد با $'.'$ نشان داده می‌شود. بنابراین $S = s_1 \dots s_n$ ساختار دوم RNA را نشان می‌دهد که در آن برای هر باز i ، این فرمول $1 \le i \le n$ ، $s_i \in \{'(','),'.'\}$ در نظر گرفته می شود.

الگوریتم ژنتیک روشی است که در حل مسائل بهینه‌سازی مورد استفاده قرار می‌گیرد و براساس فرآیندهای ژنتیکی موجودات زنده است. الگوریتم ژنتیک با جمعیتی از افراد بیان می‌شود که هر فرد نشان‌دهنده راه‌حلی برای مسئله است. مقدار برازندگی به هر فرد با توجه به میزان مناسب بودن آن به‌عنوان یک راه‌حل، اختصاص داده می‌شود. افراد براساس مقدار برازندگی جفت‌گیری کرده و نسل جدید تشکیل می‌گردد. فرآیند تولید نسل تا زمانی ادامه می‌یابد که راه‌حل بهینه برای مسئله یافت شود.

روش پیشنهادی: در مسئله پیشگویی ساختار دوم RNA، یک RNA به‌عنوان ورودی در نظر گرفته می‌شود و هدف یافتن ساختار دوم RNA است که دارای حداقل انرژی آزاد باشد. تعریف دقیق‌تر مسئله به شرح زیر است:

ورودی: یک RNA با دنباله $R = r_1 r_2 \dots r_n$ در جهت ′۵ به ′۳.

خروجی: ساختار دوم R که با دنباله‌ای از کاراکترهای $'('$، $')'$ و $'.'$ نشان داده می‌شود.

در این مقاله، روش پیشنهادی برای حل مسئله ساختار دوم RNA براساس یک الگوریتم ژنتیک به نام RNAG است که شامل مراحل ایجاد جمعیت اولیه، عملگرهای تقاطع و جهش و شرط خاتمه الگوریتم می‌باشد که در ادامه به توضیح آنها پرداخته می شود.

ایجاد جمعیت اولیه: چگونگی تولید جمعیت اولیه به‌ترتیب زیر انجام می‌شود:

۳) ماتریس نقطه‌ای $M^R_{n\times n}$ برای دنباله R براساس بازهای مکمل واتسون–کریک ایجاد می‌گردد که مقدار آن در موقعیت (i,j) به صورت زیر تعریف می‌شود:

$$M^R[i,j] = \begin{cases} 1 \ if(r_i, r_j) \in \{(A,U),(U,A),\ (C,G),(G,C)\}, \\ 0 \qquad\qquad else. \end{cases}$$

به طوری که r_i و r_j به‌ترتیب نشان‌دهنده باز i ام و j ام در دنباله R برای هر i و j، $1 \le i,j \le n$ هستند. (توجه کنید که در ایجاد این ماتریس جفت‌باز $\mathrm{G-U}$ در نظر گرفته نمی‌شود چون دقت پیشگویی را کاهش می‌دهد.)

۴) در ماتریس نقطه‌ای $M^R_{n\times n}$، تمام مقادیر مورب متوالی ۱ که روی قطر اصلی یا موازی آن قرار داشته باشند به‌عنوان یک زیرقطر در نظر گرفته می‌شوند. مجموعه‌ای از زیرقطرهای R به صورت زیر تعریف می‌گردد:

$$D^R = \{ < i, j, k, l > | 1 \le i \le k \le n \ \& \ 1 \le j \le l \le n \}$$

به طوری که (i,j) و (k,l) به‌ترتیب موقعیت شروع و پایان یک زیرقطر را مشخص می‌کنند. فرض کنید $d^R \in D^R$ و $d^R =< i,j,k,l >$ است. این زیرقطر نشان دهنده این است که زیردنباله r_{ik} در R به زیردنباله r_{jl} در معکوس R متصل می‌شود. اگر i' و j' با شرایط $i < i' < k$ ، $j < j' < l$ و $i' + j' = n + 1$ موجود باشند، آنگاه d^R بایستی از مجموعه D^R حذف و دو زیرقطر $< i,j,i',j' >$ و $< i'+1, j'+1, k, l >$ به مجموعه مورد نظر اضافه شوند. زیرا در این حالت تعدادی از بازهای تشکیل دهنده زیرقطر d^R دو مرتبه جهت تشکیل پیوند محاسبه می‌گردند.

۵) جمعیت اولیه C براساس D^R به این صورت ساخته می‌شود که برای هر i، $1 \le i \le |C|$، مراحل زیر انجام می‌پذیرد:

الف) $C[i] \subseteq D^R$، $C[i]$ به طور تصادفی از مجموعه D^R ایجاد می‌گردد. به بیان دقیق‌تر، برای هر فرد (که ابتدا تهی است) در جمعیت، ابتدا یک زیرقطر تصادفی از D^R انتخاب شده و در آن قرار می‌گیرد. زیرقطرهای بعدی نیز به طور تصادفی انتخاب شده و در صورتی در فرد قرار می‌گیرند که با هیچ یک از زیرقطرهای موجود در آن همپوشانی نداشته باشند. اگر همپوشانی وجود داشته باشد، قسمتهای همپوشان از زیرقطر جدا شده و زیرقطر حاصل به مجموعه زیرقطرهای قبلی اضافه می‌شود. فرض کنید $d_1, d_2 \in D^R$ که $d_1 =< i_1, j_1, k_1, l_1 >$ و $d_2 =< i_2, j_2, k_2, l_2 >$. همپوشانی دو زیرقطر d_1 و d_2 به صورت زیر تعریف می‌گردد:

$$\textbf{Overlap}(d_1, d_2) = \begin{cases} 1 & if\ \exists p: i_1 \le p \le k_1\ \&\ i_2 \le p \le k_2 \\ 1 & if\ \exists p: j_1 \le p \le l_1\ \&\ j_2 \le p \le l_2 \\ 0 & else. \end{cases}$$

ب) مقدار برازندگی فرد $C[i]$ به صورت زیر محاسبه می شود :

$$\textbf{Fitness}\ (C[i]) = \sum_{d' \in C[i]} MFE(d') + \sum_{l \in loop} MFE(l)$$

به طوری که d' نشان‌دهنده یک زیرقطر در مجموعه $C[i]$ است و l حلقه‌ای در مجموعه حلقه‌های هیرپین، بالج، داخلی و چندحلقه‌ای در فرد را نشان می‌دهد. حداقل انرژی آزاد شده از $d' =< i', j', k', l' > \in C[i]$ به صورت زیر محاسبه می‌گردد:

$$\textbf{MFE}\ (d') = \sum_{\substack{(r_p, r_q), (r_{p+1}, r_{q-1}) \in R \\ i' \le p \le k' \\ j' \le q \le l'}} e(r_{p,p+1}, r_{q-1,q})$$

که $e(r_{p,p+1}, r_{q-1,q})$ انرژی آزاد شده از دو جفت‌باز مجاور (r_p, r_q) و (r_{p+1}, r_{q-1}) می‌باشد و R نشان‌دهنده مجموعه‌ای از جفت‌بازها در زیرقطر است. حداقل انرژی آزاد شده از تمام دو جفت‌بازهای مجاور که تشکیل ساقه می‌دهند، در جدول۱ نشان داده شده است. انرژی آزاد حلقه‌های هیرپین، بالج و داخلی مجموع دو مقدار زیر می باشد (۲۰):

۳) حداقل انرژی آزاد شده از این نوع حلقه‌ها براساس اندازه حلقه در جدول۲ مشخص شده است. برای حلقه-های با طول بیشتر از ۳۰، حداقل انرژی آزاد به صورت زیر محاسبه می‌شود:

$$\textbf{MFE}(l) = \textbf{MFE}(30) + 1.75 * RT * ln\ (size/30)$$

به طوری که R ثابت جهانی گاز، T دمای خالص و $size$ اندازه حلقه است.

جدول ۱ - حداقل انرژی آزاد شده از تمام دو جفت‌بازهای مجاور در ساقه.

5'->3'	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-0.9
AC	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-2.2	.
AG	.	.	.	.	.	.	.	.	.	.	.	.	.	-2.1	.	-0.6
AU	.	.	.	.	.	.	.	.	.	.	.	.	-1.1	.	-1.4	.
CA	.	.	.	.	.	.	.	.	.	.	.	-2.1	.	.	.	.
CC	.	.	.	.	.	.	.	.	.	.	-3.3	.	.	.	.	.
CG	.	.	.	.	.	.	.	.	.	-2.4	.	-1.4	.	.	.	.
CU	.	.	.	.	.	.	.	.	-2.1	.	-2.1	.	.	.	.	.
GA	.	.	.	.	.	.	.	-2.4	.	.	.	.	.	.	.	-1.3
GC	.	.	.	.	.	.	-3.4	.	.	.	.	.	.	.	-2.5	.
GG	.	.	.	.	.	-3.3	.	-1.5	.	.	.	.	.	-2.1	.	-0.5
GU	.	.	.	.	-2.2	.	-2.5	.	.	.	.	.	-1.4	.	1.3	.
UA	.	.	.	-1.3	.	.	.	.	.	.	.	-1	.	.	.	.
UC	.	.	-2.4	.	.	.	.	.	.	.	-1.5	.	.	.	.	.
UG	.	-2.1	.	-1	.	.	.	.	.	-1.4	.	0.3	.	.	.	.
UU	-0.9	.	-1.3	.	.	.	.	.	-0.6	.	-0.5	.	.	.	.	.

جدول ۲- حداقل انرژی آزاد شده از حلقه‌های داخلی، بالج و هیرپین براساس اندازه حلقه.

هیرپین	بالج	داخلی	اندازه
.	۳.۸	.	۱
.	۲.۸	.	۲
۵.۴	۳.۲	.	۳
۵.۶	۳.۶	۱.۱	۴
۵.۷	۴	۲.۱	۵
۵.۴	۴.۴	۱.۹	۶
۶	۴.۶	۲	۷
۵.۵	۴.۷	۲.۲	۸
۶.۴	۴.۸	۲.۳	۹
۶.۵	۴.۹	۲.۴	۱۰
۶.۶	۵	۲.۵	۱۱
۶.۷	۵.۱	۲.۶	۱۲
۶.۸	۵.۲	۲.۷	۱۳
۶.۹	۵.۳	۲.۸	۱۴
۶.۹	۵.۴	۲.۸	۱۵
۷	۵.۴	۲.۹	۱۶
۷.۱	۵.۵	۳	۱۷
۷.۱	۵.۵	۳	۱۸
۷.۲	۵.۶	۳.۱	۱۹
۷.۲	۵.۷	۳.۲	۲۰
۷.۳	۵.۷	۳.۲	۲۱
۷.۳	۵.۸	۳.۳	۲۲
۷.۴	۵.۸	۳.۳	۲۳
۷.۴	۵.۸	۳.۴	۲۴
۷.۵	۵.۹	۳.۴	۲۵
۷.۵	۵.۹	۳.۴	۲۶
۷.۵	۶	۳.۵	۲۷
۷.۶	۶	۳.۵	۲۸
۷.۶	۶	۳.۶	۲۹
۷.۷	۶.۱	۳.۶	۳۰

۴) تعدادی از حلقه‌ها انرژی مازادی براساس نوکلئوتیدهای حلقه دارند. این نوع حلقه‌ها و انرژی مازاد آنها در جدول ۳ نشان داده شده است.

توجه کنید که انرژی دیگری بین جفت‌باز انتهایی حلقه‌ها و دو باز جفت‌نشده مجاور آن وجود دارد که در اینجا به آن پرداخته نشده است.

در انتها، افراد براساس مقدار برازندگی حداقل انرژی آزاد به‌ترتیب صعودی مرتب می‌شوند.

جدول ۳- حداقل انرژی آزاد شده از حلقه‌ها با توجه به نوکلئوتیدهای حلقه.

انرژی	دنباله	اندازه
۶.۸	CAACG	۵
۶.۹	GUUAC	۵
۲.۸	CUACGG	۶
۲.۷	CUCCGG	۶
۳.۷	CUUCGG	۶
۳.۳	CCAAGG	۶
۳.۴	CCCAGG	۶
۳.۵	CCGAGG	۶
۳.۷	CCUAGG	۶
۳.۷	CCACGG	۶
۳.۶	CCGCGG	۶
۲.۵	CCUCGG	۶
۳.۶	CUAAGG	۶
۳.۷	CUCAGG	۶
۳.۵	CUUAGG	۶
۲.۸	CUGCGG	۶
۵.۵	CAACGG	۶
۲.۹	ACAGUGCU	۸
۳.۶	ACAGUGAU	۸
۱.۸	ACAGUUCU	۸
۲.۸	ACAGUACU	۸

عملگر تقاطع: عملگر تقاطع با نرخ ۰/۹ روی افراد انجام می‌گیرد، به این صورت که ابتدا ۵ درصد از بهترین افراد و ۵ درصد از افراد با برازندگی متوسط به نسل بعد منتقل می‌شوند. باقی‌مانده افراد، به‌ترتیب میزان برازندگی دو به دو برای جفت‌گیری انتخاب می‌گردند. به عبارت دیگر دو فرد i و $i+1$ از جمعیت به عنوان والدین در نظر گرفته می‌شوند، سپس از یک موقعیت تصادفی جفت‌گیری می‌کنند و در پایان دو فرزند ایجاد می‌گردد. با توجه به اینکه

طول هر فرد n است، احتمال انتخاب هر موقعیت تصادفی $1/n$ می‌باشد. در این موقعیت، والدین به دو بخش تقسیم شده و فرزند اول و دوم به‌ترتیب زیرقطرهای موجود در طرف چپ والدین اول و دوم را به خود اختصاص می دهند. هریک از زیرقطرهای طرف راست والدین اول و دوم در صورتی به‌ترتیب در فرزندان دوم و اول قرار می-گیرند که با هیچ یک از زیرقطرهای موجود در فرزند همپوشانی نداشته باشند. اگر همپوشانی وجود داشته باشد، در فرزند قرار داده نمی‌شود یا در صورت امکان بخشی که همپوشانی ندارد انتخاب شده و در فرزند قرار می‌گیرد. این فرآیند نسل بعد را با فرزندان جدید تشکیل می‌دهد.

عملگر جهش: با توجه به اینکه نرخ جهش ۰/۱ است، ۱۰ درصد از ضعیف‌ترین افراد جمعیت گزینش می‌گردند تا عملگر جهش روی آنها اجرا شود. برای این افراد، یک زیرقطر تصادفی از ماتریس نقطه‌ای انتخاب شده و تنها در صورتی با یک زیرقطر تصادفی از فرد جایگزین می‌گردد که با هیچ یک از زیرقطرهای موجود در فرد (جز زیرقطر انتخابی از فرد) همپوشانی نداشته باشد. اگر همپوشانی وجود داشته باشد، انتخاب زیرقطر تصادفی از ماتریس - نقطه‌ای ادامه می‌یابد تا زمانی که همپوشانی موجود نباشد یا زمان خاتمه یابد.

خاتمه الگوریتم: فرآیند تولید نسل هنگامی متوقف می‌شود که شرط $Fitness(C[1]) \geq 2 * Fitness(C\,[\lfloor |C|/2 \rfloor])$ برقرار گردد. در انتها فردی جهت تشکیل ساختار دوم گزینش می‌شود که انرژی آزاد کمتری داشته باشد.

نتایج

الگوریتم پیشنهادی، RNAG، روی تعدادی از RNAها جهت پیشگویی ساختار دوم آنها اجرا شده است. مجموعه داده‌ها شامل *CopA*، *CopT*، *R1inv*، *R2inv*، *Tar*، *Tar**، *DIS*، $IncRNA_{54}$ و *RepZ* است. به‌عنوان مثال *CopA* را در نظر بگیرید. شکل۱ ساختار دوم پیشگویی شده *CopA* را نشان می‌دهد که به ساختار واقعی بسیار نزدیک است. برای ارزیابی دقت پیشگویی RNAG از دو معیار حساسیت و برجستگی ویژه استفاده می‌شود که به صورت زیر محاسبه می‌گردند:

$$\text{حساسیت} = \frac{\text{تعداد جفت بازهای به طور صحیح پیشگویی شده}}{\text{تعداد جفت بازها در ساختار مرجع}} \quad (۱)$$

$$\text{برجستگی ویژه} = \frac{\text{تعداد جفت بازهای به طور صحیح پیشگویی شده}}{\text{تعداد جفت بازها ی پیشگویی شده}} \quad (۲)$$

معیارF با در نظر گرفتن هر دو مقدار حساسیت و برجستگی ویژه به صورت زیر تعیین می‌شود:

$$\text{معیارF} = \frac{\text{برجستگی ویژه}*\text{حساسیت}*۲}{(\text{برجستگی ویژه}+\text{حساسیت})} \quad (۳)$$

5'-GUGGGCCCCGGUAAUCUUUUCGUACUCGCCAAAGUUGAAGAAGAUUAUCGGGGUUU-3'

. . . . .((((((((((((((. . . . . .)))))))))))))). . .

ساختار دوم واقعی CopA

5'-GUGGGCCCCGGUAAUCUUUUCGUACUCGCCAAAGUUGAAGAAGAUUAUCGGGGUUU-3'

. . . . .(((((.(((((((((. . . . . .)))))))).))))). . .

ساختار دوم پیشگویی شده CopA

شکل۱

جدول ۴ - دقت پیشگویی RNAG روی مجموعه‌ای از RNAها.

دنباله RNA	طول	(٪) حساسیت	(٪) برجستگی ویژه	(٪) معیار F
Tar	۱۶	۱۰۰,۰۰	۱۰۰,۰۰	۱۰۰,۰۰
Tar*	۱۶	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
R1inv	۲۱	۱۰۰,۰۰	۱۰۰,۰۰	۱۰۰,۰۰
R2inv	۱۹	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
DIS	۳۵	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
CopA	۵۶	۹۲,۳۰	۱۰۰,۰۰	۹۶,۰۰
CopT	۵۷	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
$IncRNA_{54}$	۵۴	۱۰۰.۰۰	۷۳,۳۳	۸۴,۶۱
RepZ	۶۱	۶۸.۱۸	۷۸.۹۵	۷۳,۱۷
Average		۹۵.۶۱	۹۴.۷۰	۹۵.۱۵

جدول ۵ - مقایسه حساسیت RNAG با تعدادی از رویکردها.

توالی RNA	RNAG	RNAFold	MFold
Tar	۱۰۰,۰۰	۱۰۰,۰۰	۱۰۰,۰۰
Tar*	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
R1inv	۱۰۰,۰۰	۱۰۰,۰۰	۱۰۰,۰۰
R2inv	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
DIS	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
CopA	۹۲,۳۰	۱۰۰,۰۰	۱۰۰,۰۰
CopT	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
$IncRNA_{54}$	۱۰۰.۰۰	۱۰۰,۰۰	۱۰۰,۰۰
RepZ	۶۸.۱۸	۱۰۰,۰۰	۶۸,۱۸
Average	۹۵.۶۱	۱۰۰,۰۰	۹۶,۴۶

جدول۶- مقایسه برجستگی ویژه RNAG با تعدادی از رویکردها.

توالی RNA	RNAG	RNAFold	MFold
Tar	۱۰۰,۰۰	۱۰۰,۰۰	۸۳,۳۳
Tar*	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
R1inv	۱۰۰,۰۰	۷۷,۷۸	۷۷,۷۸
R2inv	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
DIS	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
CopA	۱۰۰,۰۰	۶۱,۹۰	۷۲,۲۲
CopT	۱۰۰.۰۰	۶۶,۶۷	۶۶,۶۷
$IncRNA_{54}$	۷۳,۳۳	۶۴,۷۰	۵۷,۸۹
RepZ	۷۸,۹۵	۹۰,۹۰	۷۸,۹۵
Average	۹۴.۷۰	۸۴,۶۶	۸۱,۸۷

جدول۷- مقایسه معیارF روش RNAG با تعدادی از رویکردها.

RNA توالی	RNAG	RNAFold	MFold
Tar	۱۰۰,۰۰	۱۰۰,۰۰	۹۰,۹۱
Tar*	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
R1inv	۱۰۰,۰۰	۸۷,۵۰	۸۷,۵۰
R2inv	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
DIS	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰,۰۰
CopA	۹۶,۰۰	۷۶,۴۷	۸۳,۸۷
CopT	۱۰۰.۰۰	۸۰,۰۰	۸۰,۰۰
$IncRNA_{54}$	۸۴,۶۱	۷۸,۵۷	۷۳,۳۳
RepZ	۷۳,۱۷	۹۵,۲۳	۷۳.۱۷
Average	۹۵,۱۵	۹۱,۶۹	۸۸,۵۷

جدول ۴ دقت پیشگویی RNAG را در حساسیت، برجستگی ویژه و معیارF روی داده‌های آزمایشی نشان می‌دهد. برای RNAهای *R1inv*، *Tar*، *Tar**، *R2inv*، *DIS* و *CopT* دقت پیشگویی ۱۰۰ درصد در هر سه معیار مذکور حاصل شده است. دقت پیشگویی *CopA* در حساسیت، برجستگی ویژه و معیارF به‌ترتیب ۹۲,۳، ۱۰۰ و ۹۶ درصد است. برای $IncRNA_{54}$ و *RepZ* معیارF به‌ترتیب ۸۴,۶۱ و ۷۳,۱۷ درصد حاصل شده است. همان‌طور که مشاهده می شود، دقت متوسط الگوریتم پیشنهادی روی مجموعه داده‌ها به‌ترتیب ۹۵,۶۱، ۹۴,۷ و ۹۵,۱۵ درصد در حساسیت، برجستگی ویژه و معیارF است.

بحث

در این مقاله، یک روش ژنتیک جهت پیشگویی ساختار دوم RNA معرفی شد. در این روش یک ماتریس نقطه‌ای نشان‌دهنده تمام جفت‌بازهای ممکن RNA ایجاد می‌گردد و زیرقطرهای آن که به‌عنوان مناطق ممکن برای تشکیل ساقه در نظر گرفته می‌شوند، استخراج می‌گردند. هر فرد در این الگوریتم شامل یک زیرمجموعه تصادفی از زیرقطرهای غیرهمپوشان است. در ادامه مقدار برازندگی حداقل انرژی آزاد برای هریک از افراد محاسبه شده و افراد به‌ترتیب صعودی مقدار برازندگی مرتب می‌گردند. عملگر تقاطع با نرخ ۰/۹ روی افراد انجام می‌شود. در این عمل، فرزندان از ترکیب والدین در یک موقعیت تصادفی ساخته می‌شوند. پس از آن جهش با نرخ ۰/۱ انجام می‌پذیرد و به این ترتیب نسل بعد ایجاد می‌گردد. اگر مقدار برازندگی فردی مناسب باشد، آن فرد برای تشکیل ساختار دوم گزینش می‌شود، در غیر این صورت نسل بعد تشکیل می گردد. الگوریتم پیشنهادی روی تعدادی از RNAها مانند *CopA*، *CopT*، *R1inv*، *R2inv*، *Tar*، *Tar**، *DIS*، $IncRNA_{54}$ و *RepZ* اجرا شده است.

جدولهای ۵، ۶ و ۷ به‌ترتیب میزان حساسیت، برجستگی ویژه و معیارF روشهای مختلف، RNAFold (۳) و MFold (۲۱)، را در مقایسه با RNAG نشان می‌دهند. همان‌طور که مشاهده می‌شود حساسیت روش پیشنهادی از روشهای RNAFold و MFold کمتر است اما مقدار برجستگی ویژه، و معیارF که به‌عنوان میانگین همساز حساسیت و برجستگی ویژه در نظر گرفته می‌شود از روشهای مذکور بیشتر است. متوسط معیارF روشهای RNAG، RNAFold و MFold روی داده‌های آزمایشی به‌ترتیب ۹۵,۱۵، ۹۱,۶۹ و ۸۸,۵۷ درصد حاصل شده است. بنابراین روش پیشنهادی به کارآیی روشهای دیگر در حساسیت، برجستگی ویژه و معیارF است.

تشکر و قدردانی

لازم است از جناب آقای دکتر محمد گنج تابش به دلیل ارائه جدول حداقل انرژی آزاد جفت بازهای مجاور (جدول۱) تشکر و قدردانی شود.

منابع

۳) قربانی، ا.، چینی کار، ص.، و بهمنی، م.خ.، (۱۳۸۸)، بررسی مولکولی و تعیین توالی بخش RNA s ژنوم ویروس تب کریمه- کنگو (CCHF) در ایران، مجله زیست شناسی ایران، ج ۲۲، ش ۴، ص ۷۰۴-۷۱۰.

۴) مرادی، ا.، شریفی، م.، و موسوی، ا.، (۱۳۹۰)، بررسی بیان ژن H6H و ایزوفرمهای PMT تحت تأثیر غلظتهای مختلف سالیسیلیک اسید در ریشه های مویی و اندامهای مختلف شابیزک، مجله زیست شناسی ایران، ج ۲۴، ش ۳، ص ۳۶۶-۳۷۲.

3) Hofacker, I.L., (2003), Vienna RNA secondary structure server, Nucleic Acids Research, 31(13): 3429–31.

4) Kato, Y., Akutsu, T., and Seki, H., (2009), A grammatical approach to RNA–RNA interaction prediction, Pattern recognition, 42: 531-538.

5) Mathews, D.H., and Turner, D.H., (2006), Prediction of RNA secondary structure by free energy minimization, Vol 16, 3: 270-278.

6) Mathews, D.H., Sabina, J., Zuker, M., and Turner D.H., (1999), Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, Journal of Molecular Biology, 288: 911-940.

7) McCaskill, J.S., (1990), The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers, 29: 1105-1119.

8) Meyer, I.M., (2008), Predicting novel RNA-RNA interactions, Current opinion in structural biology, 18: 387-393.

9) Nussinov, R. and Jacobson, A.B., (1980), Fast algorithm for predicting the secondary structure of single-stranded RNA, In Proceedings of the National Academy of Sciences of the United States of American, Vol 77: 6309-6313.

10) Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J., (1978), Algorithms for loop matching, SIAM J.Appl.Math, 35: 68-82.

11) Puerta-Fernandez, E., Romero-Lpez, C., Barroso-delJesus A., and Berzal-Herranz, A., (2003), Ribozymes: recent advances in the development of RNA tools, FEMS Microbiology Reviews, 27: 75–97.

12) Sakakibara, Y., Brown, M., Hughey, R., Mian I.S., Sjolander K., Underwood R.C. and Hussler D., (1999), Stochastic context-free grammars for tRNA modeling, Nucleic Acids Res, 22: 5112-5120.

13) Salari, R., Backofen, R., and Sahinalp, S.C., (2010), Fast prediction of RNA-RNA interaction, Algorithms for molecular Biology, 5: 5-15.

14) Simons, R.W., and Grunberg-Manago, M., (1998), RNA structure and function, Cold Spring Harbor Laboratory Press.

15) Turner, D.H., Sugimoto, N., Jaeger, J.A., Longfellow, C.E., Freier, S.M., and Kierzek, R., (1987), Improved parameters for prediction of RNA structure, Cold Spring Harb. Symp. Quant. Biol., 52:123-133.

16) Zvelebil, M., and Baum, J.O., (2008), Understanding Bioinformatics, Garland Science. 461-514.

17) Zuker, M., Mathews, D.H., and Turner, D.H., (1999), Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide, In RNA Biochemistry and Biotechnology.

18) Zuker, M., and Sankoff, M., (1984), RNA secondary structures and their prediction, Blletin of Mathematical of biology, Vol 46, 4: 591-621.

19) Zuker, M., (1994), Prediction of RNA secondary structure by energy minimization, Method in Molecular Biology, 25: 267–94.

20) Zuker, M. and Stiegler P., (1981), Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic Acids Res, 9(1): 133-48.

21) Zuker, M. ,(2003), Mfold web server for nucleic acid folding and hybridization

prediction, Nucleic Acids Res. 31(13): 3406- 3415.

A genetic approach to accurately predict RNA secondary structure

Montaseri S.[1], Moghadam-Charkari N.[2] and Zare-Mirakabad F.[3]

[1] Computer Sciences Dept., Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, I.R. of Iran

[2] Faculty of Electrical & Computer Engineering, Tarbiat Modares University, Tehran, I.R. of Iran

[3] Faculty of Mathematics & Computer Science, Amirkabir University of Technology, Tehran, I.R. of Iran

Abstract

RNA molecule plays important and fundamental roles in many biological processes. In the most times, activities of RNAs are determined by their structures. In notice to complexity and costly of laboratory methods to predict RNAs structure, computational approaches are used. There are variety of algorithms to predict RNA secondary structure. In this paper, a genetic algorithm called RNAG is presented to predict the RNA secondary structure based on minimum free energy (MFE). In this algorithm, each individual of population includes some stems. The individuals are increasingly ranked based on fitness value of MFE from stems and loops, and in the follow, crossover and mutation operations are done on individuals to make a new population, respectively. Process of population generation continues until an individual with proper MFE is produced. Finally, this individual is selected as an optimal RNA secondary structure. The proposed algorithm is performed on some RNAs in the *bacteria*. Results of the paper show that RNAG algorithm has a high accuracy in comparison with the other related methods.

Key words: minimum free energy, stem, fitness value.