

## پیشگویی دقیق ساختار دوم RNA مبتنی بر الگوریتم ژنتیک

سهیلا منتصری<sup>۱</sup>، نصرالله مقدم-چرکری<sup>۲\*</sup> و فاطمه زارع میرک آباد<sup>۳</sup>

<sup>۱</sup> تهران، دانشگاه تربیت مدرس، دانشکده علوم ریاضی، گروه علوم کامپیوتر

<sup>۲</sup> تهران، دانشگاه تربیت مدرس، دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر

<sup>۳</sup> 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 به ترتیب صعودی رتبه‌بندی شده و در ادامه به ترتیب عملگرهای تقاطع و جهش روی آنها برای ایجاد نسل بعد اجرا می‌گردد. 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.

واژه‌های کلیدی: حداقل انرژی آزاد، ساقه، مقدار برازندگی.

\* نویسنده مسئول، تلفن: ۰۲۸۸۳۳۰۱، پست الکترونیکی: charkari@modares.ac.ir

### مقدمه

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 پیشینه 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

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، ۲ و ۱۱). 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

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*, *IncRNA54* و *RepZ* هستند (۴).

**تعاریف پایه:** دنباله RNA از چهار نوع نوکلئوتید تشکیل می شود که شامل آدنین (A)، گوانین (G)، سیتوزین (C) و یوراسیل (U) است. هر RNA دارای دو انتهای مجزا است که به عنوان ۳' و ۵' شناخته می شوند. یک دنباله RNA به نام  $R$ ،  $|R| = n$  در جهت ۵' به ۳' به صورت زیر تعریف می گردد:

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

معکوس  $R$  با  $r_1, r_2, \dots, r_n$  در جهت ۳' به ۵' مشخص می شود. بنابراین  $r_{ij} = r_i r_{i+1} \dots r_j$  زیردنباله ای از  $R$  است که از موقعیت  $i$  شروع شده و به موقعیت  $j$  ختم می گردد. دنباله RNA با تشکیل پیوند هیدروژنی بین بازهای آن تشکیل ساختار می دهد. بیشتر پیوندها بین بازهای مکمل واتسون-کریک روی می دهند که در آنها  $G$  با  $C$  و  $A$  با  $U$  جفت می شوند و برعکس. این پیوندها می توانند ساختار دوم RNA را تشکیل دهند.

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

$$i < i' < j' < j$$

$$i' < i < j < j'$$

به فرض اینکه  $r_{k1}$  و  $r_{ij}$  دو زیردنباله از RNA باشند که تشکیل ساقه می دهند. بنابراین زیردنباله  $r_{ij}$  به معکوس  $r_{k1}$  متصل می شود. به عبارت دیگر، بین هر یک از بازهای  $r_{ij}$  و معکوس  $r_{k1}$  به ترتیب پیوند هیدروژنی برقرار می گردد.

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

در این مقاله، یک الگوریتم ژنتیک به نام RNAG جهت پیشگویی دقیق ساختار دوم مولکول RNA ارائه می شود. در این الگوریتم، یک ماتریس نقطه ای ایجاد می گردد که نشان دهنده تمام جفت بازهای ممکن در RNA است. هر زیرقطر در ماتریس نقطه ای را می توان به عنوان یک ساقه در نظر گرفت. سپس جمعیتی از ساقه هایی که به طور تصادفی انتخاب می شوند، ایجاد شده و مقدار برازندگی (Fitness value) حداقل انرژی آزاد شده از ساقه ها و حلقه ها برای هر فرد موجود در جمعیت محاسبه می گردد. برای ایجاد نسل جدید، عملگرهای تقاطع (Crossover) و جهش (Mutation) به ترتیب روی تعدادی از افراد نسل جاری انجام می شود. فرآیند تولید نسل ادامه می یابد تا زمانی که انرژی آزاد فردی به حد مطلوب برسد. در نهایت، این فرد با حداقل انرژی آزاد جهت تشکیل ساختار دوم RNA انتخاب می شود. الگوریتم پیشنهادی روی تعدادی از داده ها شامل *CopA*, *CopT*, *R1inv*, *R2inv*, *Tar*, *Tar\**, *DIS*, *IncRNA54* و *RepZ* در باکتریها به کار رفته است. نتایج حاصل از این تحقیق نشان می دهد که



انرژی آزاد شده از  $d' = \langle i', j', k', l' \rangle \in C[i]$  به صورت زیر محاسبه می‌گردد:

$$MFE(d') = \sum_{(r_p, r_q), (r_{p+1}, r_{q-1}) \in R} e(r_{p,p+1}, r_{q-1, q})$$

$$i' \leq p \leq k'$$

$$j' \leq q \leq l'$$

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 جدول ۱ نشان داده شده است. انرژی آزاد حلقه‌های هیرپین، بالچ و داخلی مجموع دو مقدار زیر می‌باشد (۲۰):

۳) حداقل انرژی آزاد شده از این نوع حلقه‌ها براساس اندازه حلقه در جدول ۲ مشخص شده است. برای حلقه‌های با طول بیشتر از ۳۰، حداقل انرژی آزاد به صورت زیر محاسبه می‌شود:

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

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

الف)  $C[i], C[i] \subseteq D^R$  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  $d_1 = \langle i_1, j_1, k_1, l_1 \rangle$  و  $d_2 = \langle i_2, j_2, k_2, l_2 \rangle$  همپوشانی دو زیرقطر  $d_1$  و  $d_2$  به صورت زیر تعریف می‌گردد:

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

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

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

به طوری که  $d'$  نشان‌دهنده یک زیرقطر در مجموعه  $C[i]$  است و  $l$  حلقه‌ای در مجموعه حلقه‌های هیرپین، بالچ، داخلی و چندحلقه‌ای در فرد را نشان می‌دهد. حداقل

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، بالچ و هیرپین براساس اندازه حلقه.

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

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

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

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

**عملگر تقاطع:** عملگر تقاطع با نرخ  $0/9$  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، به ترتیب 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

(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.

**خاتمه الگوریتم:** فرآیند تولید نسل هنگامی متوقف می‌شود که شرط  $Fitness(C[1]) \geq 2 * Fitness(C[|C|/2])$  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  $CopA$ ،  $CopT$ ،  $R1inv$ ،  $R2inv$ ،  $Tar^*$ ،  $DIS$ ،  $IncRNA_{54}$  و  $RepZ$  است. به‌عنوان مثال  $CopA$  را در نظر بگیرید. شکل ۱ 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:

$$(1) \quad \text{حساسیت} = \frac{\text{تعداد جفت بازهای به طور صحیح پیشگویی شده}}{\text{تعداد جفت بازها در ساختار مرجع}}$$

$$(2) \quad \text{برجستگی ویژه} = \frac{\text{تعداد جفت بازهای به طور صحیح پیشگویی شده}}{\text{تعداد جفت بازهای پیشگویی شده}}$$

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

$$(3) \quad \text{معیار F} = \frac{\text{برجستگی ویژه} * \text{حساسیت}}{(\text{برجستگی ویژه} + \text{حساسیت})}$$

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 به ترتیب زیرقطرهای موجود در طرف چپ والدین اول و دوم را به خود اختصاص می‌دهند. هریک از زیرقطرهای طرف راست والدین اول و دوم در صورتی به ترتیب در فرزندان دوم و اول قرار می‌گیرند 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  $0/1$  است،  $10\%$  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.

5'-GUGGGCCCCGGUAAUCUUUCGUACUCGCCAAAGUUGAAGAAGAUUAUCGGGGUUU-3'  
 .....((((((((((((((((.....)))))))))))))).....)

ساختار دوم واقعی CopA

5'-GUGGGCCCCGGUAAUCUUUCGUACUCGCCAAAGUUGAAGAAGAUUAUCGGGGUUU-3'  
 .....((((((((((((((((.....)))))))))))))).....)

ساختار دوم پیشگویی شده CopA

شکل ۱

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 <sub>54</sub>	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 <sub>54</sub>	۱۰۰,۰۰	۱۰۰,۰۰	۱۰۰,۰۰
RepZ	۶۸,۱۸	۱۰۰,۰۰	۶۸,۱۸
Average	۹۵,۶۱	۱۰۰,۰۰	۹۶,۴۶

جدول ۶ - 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 <sub>54</sub>	۷۳,۳۳	۶۴,۷۰	۵۷,۱۹
RepZ	۷۸,۹۵	۹۰,۹۰	۷۸,۹۵
Average	۹۴,۷۰	۸۴,۶۶	۸۱,۸۷

جدول ۷- مقایسه معیار F روش RNAG با تعدادی از رویکردها.

توالی RNA	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 <sub>54</sub>	۸۴,۶۱	۷۸,۵۷	۷۳,۳۳
RepZ	۷۳,۱۷	۹۵,۲۳	۷۳,۱۷
Average	۹۵,۱۵	۹۱,۶۹	۸۸,۵۷

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 می‌گردد. الگوریتم پیشنهادی روی تعدادی از RNAها مانند *DIS*، *Tar\**، *Tar*، *R2inv*، *R1inv*، *CopT*، *CopA* و *IncRNA<sub>54</sub>* اجرا شده است.

جدولهای ۵، ۶ و ۷ به ترتیب میزان حساسیت، برجستگی ویژه و معیار F 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 معیار F که به‌عنوان میانگین همساز حساسیت و برجستگی ویژه در نظر گرفته می‌شود از روشهای مذکور بیشتر است. متوسط معیار F روشهای RNAG، RNAFold و MFold روی داده‌های آزمایشی به ترتیب ۹۵,۱۵، ۹۱,۶۹ و ۸۸,۵۷ درصد حاصل شده است. بنابراین روش پیشنهادی به کارایی روشهای دیگر در حساسیت، برجستگی ویژه و معیار F است.

جدول ۴ دقت پیشگویی RNAG را در حساسیت، برجستگی ویژه و معیار F روی داده‌های آزمایشی نشان می‌دهد. برای RNAهای *Tar\**، *Tar*، *R1inv* و *DIS* دقت پیشگویی ۱۰۰ درصد در هر سه معیار مذکور حاصل شده است. دقت پیشگویی *CopA* در حساسیت، برجستگی ویژه و معیار F به ترتیب ۹۲,۳، ۱۰۰ و ۹۶ درصد است. برای *RepZ* و *IncRNA<sub>54</sub>* معیار F به ترتیب ۸۴,۶۱ و ۷۳,۱۷ درصد حاصل شده است. همان‌طور که مشاهده می‌شود، دقت متوسط الگوریتم پیشنهادی روی مجموعه داده‌ها به ترتیب ۹۵,۶۱، ۹۴,۷ و ۹۵,۱۵ درصد در حساسیت، برجستگی ویژه و معیار F 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 هر یک از افراد محاسبه شده و افراد به ترتیب صعودی مقدار برازندگی مرتب می‌گردند. عملگر



## تشکر و قدردانی

لازم است از جناب آقای دکتر محمد گنج تابش به دلیل

## 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 زیست‌شناسی 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 ریشه‌های مویبی و اندام‌های مختلف شایبک، مجله زیست‌شناسی 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, *Bulletin 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.<sup>1</sup>, Moghadam-Charkari N.<sup>2</sup> and Zare-Mirakabad F.<sup>3</sup>

<sup>1</sup> Computer Sciences Dept., Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, I.R. of Iran

<sup>2</sup> Faculty of Electrical & Computer Engineering, Tarbiat Modares University, Tehran, I.R. of Iran

<sup>3</sup> 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.