فصلنامه

دانش میکروب شناسی

سال اول، شماره ۱، زمستان ۱۳۸۷

کاربرد نظریه اطلاع در توالی های نوکلئوتیدی *E.coli*

عین اله پاشا ۱، حمیدرضا مصطفایی*[۲] و شیما حسینی[۳]

۱- گروه ریاضی دانشگاه تربیت معلم تهران ایران ۲- گروه آمار دانشگاه آزاد اسلامی واحد تهران شمال (عهده دار مکاتبات) ۳- کارشناسی ارشد آمار دانشگاه آزاد اسلامی واحد تهران شمال

چکیده

پروتئین ها مکان های خاصی از توالی های نوکلئیک اسید در ژنوم را تعیین و به آنها متصل می شوند. بنابراین شناسایی جایگاه اتصال از اهمیت به سزایی برخوردار است. در این تحقیق برای تشخیص جایگاه های اتصال پروتئین ها به توالی هایی از نوکلئیک اسیدها از معیار آنتروپی شانون استفاده شده است. این اندازه $R_{seqence}$ نامیده شده که چگونگی توزیع اطلاع در سراسر محل های این جایگاه ها را نشان می دهد. با استفاده از این اندازه نشان داده شده است که محتوای اطلاع در جایگاه های اتصال پروتئین ها به توالی های نوکلئیک اسیدها دارای مقدار بزرگتر از صفر و در خارج این جایگاه ها به سمت صفر کاهش می یابد. همچنین اندازه دیگری از اطلاع با عنوان $R_{frequency}$ محاسبه شده است که مقدار اطلاع لازم برای پیدا کردن جایگاه اتصال پروتئین را نشان می دهد. در اغلب جایگاه ها این دو مقدار از اطلاع، نزدیک به یکدیگر می باشند. در بعضی موارد این جایگاه ها حاوی اطلاع بیشتری از مقدار اطلاع لازم برای شناسایی آنها هستند که این مقدار اطلاع اضافی می تواند دلیلی برای اتصال پروتئین یا پروتئین های دیگری به این مکان ها باشد. مقایسه بین این دو اندازه $R_{sequence}$ و $R_{frequency}$ بیانگر این مطلب است که اطلاع در جایگاه های اتصال، برای پیدا کردن آنها توسط پروتئین ها در ژنوم کافی می باشد.

واژگان کلیدی: احتمال، نظریه اطلاع، DN و RNA .

مقدمه

قریب به پنج دهه از گسترش نظریه اطلاع Shanon می گذرد، کاربرد نظریه اطلاع در ژنتیک ابتدا توسط اشنایدر (۱۹۸۶) مطرح گردید. وی تحقیق خود با عنوان محتوای اطلاع جایگاه های اتصال در توالی های نوکلئیک اسیدی را منتشر کرد و کاربرد اطلاع را برای تشخیص جایگاه های اتصال پروتئین ها به DNA یا RNA بیان نمود.

ترکیب خطی یک زنجیره ی DNA، چگونگی استفاده از اطلاع ژنتیکی را تفهیم می کند. ترکیب خطی چهار نوکلئوتید حاوی یک پیام ژنتیکی برای تولید پروتئین می باشد. حال سوال اساسی که قابل طرح است این می باشد که چقدر اطلاع در این آرایش ها نهفته است و چگونه اندازه گیری می شود؟ شانون (۱۹۸۴) در مقاله مشهورش تحت عنوان ((نظریه ریاضی ارتباطات)) راهی برای اندازه گیری مقدار اطلاع بنا نهاد. این مقاله با در نظر

گرفتن تعریف شانون از اندازه دارای ساختار بدین شرح است که ابتدا در بخش دوم کلیاتی از مفهوم آنتروپی و نظریه اطلاع را معرفی نموده در بخش سوم رابطه ژنتیک و آنتروپی مورد بررسی قرار گرفته و با شبیه سازی نشان داده شده است که نواحی اتصال در کدام قسمت طول ژنوم قرار می گیرد و اینکه این جایگاه ها دارای اطلاع کافی برای پیدا شدن آن توسط پروتئین LexA هستند و مقدار اطلاع لازم برای پیدا کردن این جایگاه یعنی $R_{frequency}$ محاسبه شده است.

آنتروپی و نظریه اطلاع

اصطلاح آنتروپی اولین بار توسط Clausius در اواسط قرن نوزدهم به کار برده شد. در این زمینه Boltzman تعبیر احتمالاتی آنرا در زمینه مکانیک آماری مطرح نمود لکن رابطه صریح بین آنتروپی و احتمال به سال ۱۹۰۶ توسط Planck به ثبت رسید. هر چند در این خصوص ریاضیدانان مشهوری همچون Neumann، Nyquist و Hartley مطالعات و تحقیقاتی زیادی به انجام رسانده اند.

به طور کلی، در هر فرآیند ارتباطی، یک سلسله وقایعی روی می دهد که سبب می شود چیزی به نام اطلاعات از جائی (مبداء) به جایی دیگر (مقصد) انتقال یابد، اطلاعات از منبعی به منبع دیگر زمانی جریان می یابد که منبع ارسال کننده در قیاس با منبع دریافت کننده در سطح بالاتری از اطلاع قرار داشته تا از توانایی دگرگونی در ساختار دریافت کننده برخوردار باشد.

اطلاع هارتلی: هارتلی اولین فردی است که اندازه اطلاع را تعریف کرد. او فرض کرد که هر نماد یک پیام را بتوان به S طریق انتخاب کرد، لذا با در نظر گرفتن پیام های L نمادی می توان S^L پیام متمایز تشخیص داد. هارتلی اطلاع را به صورت لگاریتم تعداد پیام های قابل تشخیص تعریف نمود بنابراین در حالتی که پیام ها با طول L باشند $H_H(S^L) = los(S^L) = L\log(S)$ و برای پیام با طول یک $H_H(S^1) = los(S)$ لذا اطلاع پیام بطول L ، L برابر اطلاع پیامی بطول یک می باشد.

در تعریف هارتلی هیچ فرضی در این خصوص که امکان دارد S نماد با شانس های نا برابر رخ دهند در نظر گرفته نشده است.

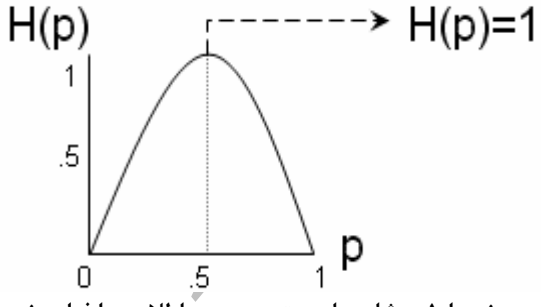
اطلاع شانون: دستاورد بزرگ شانون این است که او نظریه نیکوسیست و هارتلی را توسعه داد و نظریه اطلاع امروزی با مفهوم احتمال را پایه گذاری کرد. در حالت کلی شانون اندازه اطلاع را بر مبنای مفهوم اطلاع معرفی نمود که اندازه هارتلی را به عنوان حالت خاصی شامل می شود. درتعریف شانون، آنتروپی بعنوان یک اندازه ایی از عدم قطعیت تعریف می شود که با معکوس احتمال رخ دادن آن پیشامد مرتبط است (۳).

آنتروپی متغییر تصادفی: فرض کنید متغییر تصادفی X مقادیر $x_1, x_2, ..., x_n$ را با احتمال های $p_1, p_2, ..., p_n$ انتخاب کند. بنابر تعریف شانون $\log p_n$ میزان عدم قطعیت حاصل از پیشامد $\{X = X_1\}$ است. متوسط اندازه عدم قطعیت از اطلاع X برابر است با $H(X) = -\sum_{i=1}^{n} p_i \log p_i$ که آنتروپی متغییر تصادفی X نامیده می شود.

مثال: فرض کنید آزمایشی دارای دو برآمد با احتمال های $p_1 = p, p_2 = 1-p$ باشد،

$$H(p) = -p\log(p) - (1-p)\log(1-p)$$

نمودار ۱ نشان می دهد که $H(p)$ چگونه به عنوان تابعی از p عمل می کند. در این نمودار ملاحظه می شود که اگر یک برآمد با احتمال یک رخ دهد اندازه اطلاع آن صفر است.



نمودار۱. پیشامدهای حتمی – هیچ اطلاعی را فراهم نمی کنند(۳).

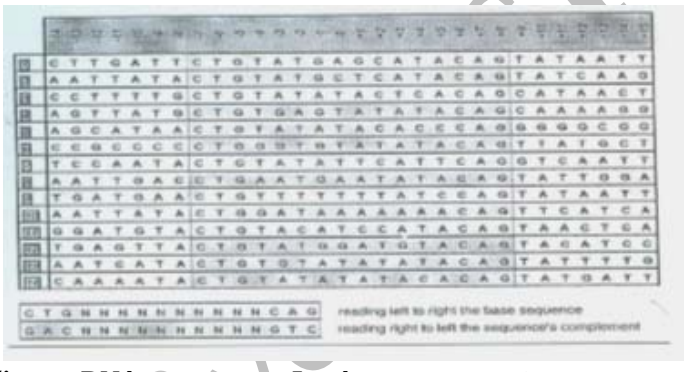
ژنتیک و آنتروپی

پروتئین های زیادی اعم از رپرسورها، پلی مرازها، ریبوزوم ها و سایر ماکرومولکولها مکان های خاصی از توالی های نوکلئیک اسید در ژنوم را تعیین و به آنها متصل می شوند. این پروتئین های باعث روشن و خاموش شدن ژن ها می شوند. این مکان ها را که یک تشخیص

دهنده اعم از پروتئین ماکروملکول به آنها متصل می شوند جایگاه اتصال می نامند. تشخیص دهنده های شناخته شده بسیاری وجود دارد. به عنوان مثال یک مولکول به نام RNA پلی مراز به توالی هایی به نام آغازگر نزدیک محل شروع ژن متصل می شود، سپس RNA پلی مراز در طول DNA حرکت و نسخه برداری از ژن را برای تشکیل یک رشته RNA شروع می نماید. تشخیص دهنده دیگر بنام ریبوزوم به توالی کوتاهی از نوکلئوتیدها در قسمت بالا دست ژن به نان جایگاه اتصال ریبوزوم متصل شده و آنگاه پس از حرکت در طول RNA آنرا به پروتئین ترجمه می کند . لذا شناسایی جایگاه های اتصال بسیار حائز اهمیت است (۷ و ۸).

محتوی اطلاع جایگاه های اتصال در باکتری *E. coli*

حال تعداد توالیهای باز در DNA باکتری *E.coli* را بررسی می شود. جدول.۱ یک مثال از تعداد توالی های جایگاه اتصال پروتئین LexA به توالی هایی از DNA باکتری *E.coli* را نشان می دهد. مختصات هر باز در بالای آن مشخص شده است. اگر به دقت به تمامی ۱۴ توالی دقت شود، الگوی یکسان GTGNNNNNNNNNCAG را که از جایگاه ۷- شروع می شود، مشاهده می گردد، که N، بیانگر هر باز می باشد. رشته ها در جدول۱ به صورت توالی ها و مکمل های آن ها از چپ به راست نوشته شده است. برای مثال در جایگاه ۱۵ از توالی اول یک باز T وجود دارد که مکمل آن باز جایگاه ۱۴- از توالی دوم یعنی A می باشد.



جدول۱. تعداد توالی های جایگاه اتصال پروتئین LexA به توالی هایی از DNA باکتری *E.coli*

بررسی توالی ها در جدول۱ نشان می دهد که در جایگاه ۵- همواره یک G است بنابراین بر طبق تعریف شانون عدم قطعیت در مورد اینکه چه بازی انتظار می رود که در جایگاه اتصال دیگری در همان مکان قرار گیرد صفر بیت است. ولی جایگاه ۴- با پیشامدهای A,C,G,T دارای عدم قطعیت بیشتری می باشند.

اگر $P(B,L)$ احتمال قرار گرفتن باز B در جایگاه L باشد یعنی $P(B,L)=\frac{n(B,L)}{n(L)}$ که $n(B,L)$ تعداد بازهای $B\in\{A,C,G,T\}$ در جایگاه L در یک مجموعه از توالی های جایگاه اتصال است و تعداد کل توالی ها در یک جایگاه $n(L)=\sum_{B=A}^{T} n(B,L)$ می باشد. لذا بنابر تعریف شانون عدم قطعیت حاصل برابر است با $H_s(L)=-\sum_{B=A}^{T} p(B,L)\log(B,L)$ (۵و۷و۸و۹) لذا بر این اصل نتایج ذیل حاصل می شود.

اگر تنها یک باز مانند A در توالی ها ظاهر شود آنگاه $p(B,L)=1$ در حالی که سایر احتمال ها مساوی صفر می باشد. بنابراین $H_s(L)$ صفر بیت می شود.

۱- اگر فقط دو باز با فراوانی یکسان ظاهر شوند مانند $p(T,L)=0, p(G,L)=0.5, p(C,L)=0, p(A,L)=0.5$ آنگاه عدم قطعیت یک بیت خواهد شد.

۲- اگر تمامی چهار باز با فراوانی یکسان ظاهر شوند آن گاه $p(B,L)=0.25$ و عدم قطعیت ۲بیت می شود.

مقدار اطلاع از تفاضل عدم قطعیت (H) قبل از دریافت اطلاع و بعد ازآن بدست می آید یعنی $R=H_{befor}-H_{after}$ در ابتدا و قبل از اتصال یک پروتئین به توالی از DNA بدلیل احتمال قرار گرفتن چهار باز A,C,G,T در هر جایگاه ۲ بیت عدم قطعیت وجود دارد بنابراین بیت۲ = H_{befor} است ولکن بعداز اتصال عدم قطعیت برای هر جایگاه L در یک مجموعه از

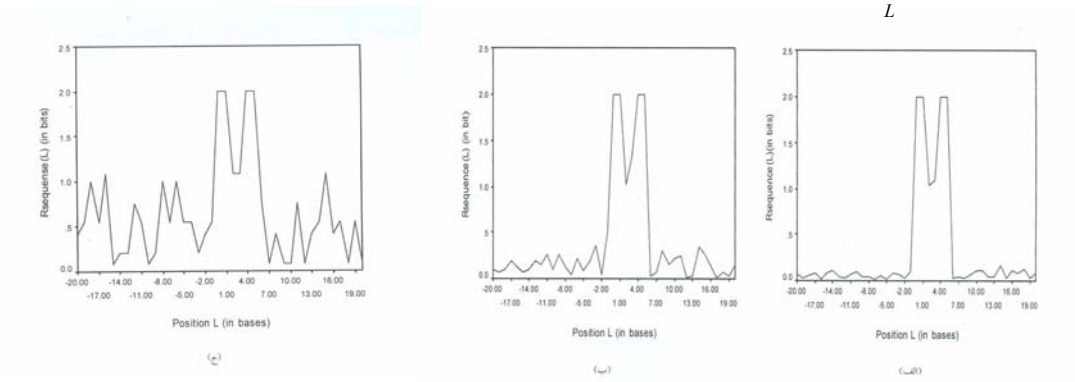
توالی های به صورت

$$H_{after} = H_s(L) = -\sum_{B=A}^{T} p(B,L)\log_2 p(B,L)$$

در نتیجه اطلاع در هر جایگاه L برابر است

با $R_{sequence}(L) = 2 - H_s(L)$ و اطلاع کل مجموعه اطلاع بر روی تمامی مکانهای یک جایگاه اتصال می باشد $R_{sequence} = \sum_L R_{sequence}(L)$ (۵ و ۱۱).

نمودار۲ منحنی $R_{sequence}(L)$ به ترتیب برای الف ۶۱، ب ۷۱ و ج ۶ جایگاه اتصال *Hinc II* انتخاب شده از قسمت انتهایی چپ باکتریوناژ T_7 را نشان می دهد. $R_{sequence}(L)$ برای ۲۰ باز محاسبه و با استفاده از نرم افزار SPSS نمودار $R_{sequence}(L)$ برای هر جایگاه رسم شده است (۸).



نمودار۲. $R_{sequence}(L)$ به ترتیب برای الف ۶۱، ب ۷۱ و ج ۶ جایگاه اتصال *Hinc II* انتخاب شده از قسمت انتهایی چپ باکتریوناژ T_7 را نشان می دهد.

با لحاظ کردن تصحیح خطای نمونه گیری در معادله تعریف شانون عدم قطعیت وجود دارد.

$$H_s(L) = -\sum_{B=A}^{T} p(B,L)\log(B,L) + e(n(L))$$

مقدار اطلاع لازم برای پیدا کردن جایگاه های اتصال:

اگر ژنومی دارای G جایگاه باشد که یک تشخیص دهنده می تواند به آن متصل شود و همه جایگاه ها دارای احتمال دسترسی برابر باشند آن گاه انتخاب یک جایگاه نیازمند به $\log_2 G$ بیت اطلاع می باشد. حال اگر این ژنوم دارای γ جایگاه اتصال باشد عدم قطعیت باقی مانده پس از مشخص شدن جایگاه ها $\log_2 \gamma$ بیت می باشد. بنابراین مقدار اطلاع مورد نیاز برای انتخاب یک جایگاه اتصال در ژنوم از رابطه زیربه دست می آید:

$$R = H_{befor} - H_{after}$$

$$R_{frequency} = \log_2 G - \log_2 \gamma = \log_2 \frac{G}{\gamma} = -\log \frac{\gamma}{G}$$

زمانی که تعداد جایگاه های اتصال در ژنوم افزایش یابد مقدار اطلاع مورد نیاز برای پیدا کردن یک جایگاه کاهش می یابد. از آنجایی که γ به تعداد جایگاه شناخته شده محدود شده و یک مقدار برآوردی نمی باشد بنابراین رابطه بالا یک کران بالا از $R_{frequency}$ را می دهد زیرا امکان دارد جایگاه های ناشناخته ای هم وجود داشته باشند (۵و۷و۸).

جدول۲. مقادیرمحاسبه شده $R_{sequence}$ و $R_{frequency}$ را برای تشخیص دهنده ها و ارگانیسم های متفاوت نشان می دهد.

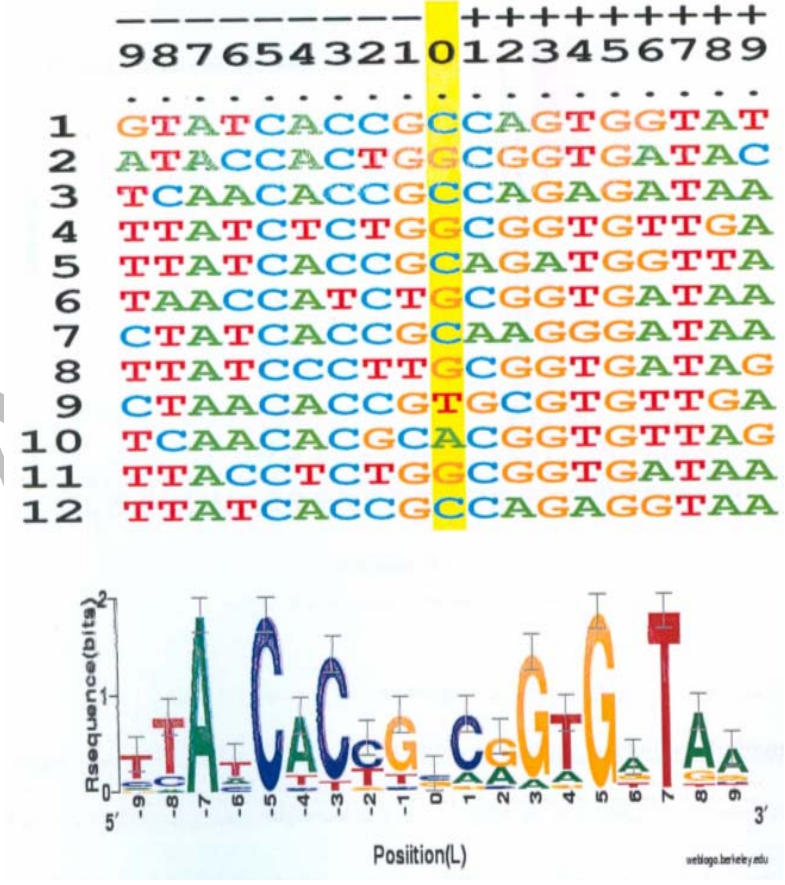
Organism	Recognizer	n	Range	$R_{sequence}$	S.D.	γ	$G\times10^{-6}$	$R_{frequency}$	$R_{sequence}/R_{frequency}$
E.coli	Ribosome	149	-26 to 18	11	0.1	2574	3.9	10.6	1.0
E.coli	LexA	14	-9 to 10	21.1	0.6	11	3.9	18.4	1.1
E.coli	TrpR	6	-18 to 19	23.4	1.9	3	3.9	20.3	1.1
E.coli	LacI	2	-21 to 21	19.2	2.8	2	7.8	21.9	0.9
E.coli	ArgR	16	-9 to 10	16.4	0.5	22	7.8	18.4	0.9
T7	RNA Polymerase	17	-29 to 12	35.4	0.7	83	7.8	16.5	2.1
T7	Symmetry	34	-6 to 7	16.4	0.2	34	7.8	17.8	0.9

با تقسیم این دو مقدار نسبت $R_{sequence}$ به $R_{frequency}$ را محاسبه نموده و مشاهده می شود این نسبت به ریبوزوم، LexA, TrpR, lacI, ArgR نزدیک ۱ می باشد. بنابراین این مطلب نشان دهنده آنست که محتوای اطلاع در جایگاه های اتصال، برای پیدا کردن آنها توسط تشخیص دهنده ها کافی می باشد.

اما یک استثنا" وجود دارد و آن اینست که در جایگاه اتصال RNA پلی مراز در باکتریوفاژ T7 حاوی دو برابر اطلاعی است که برای پیدا کردن آن ها در ژنوم لازم می باشد و این سئوال که چرا این اطلاع اضافی در جایگاه اتصال RNA پلی مراز وجود دارد؟ آیا RNA پلی مراز از همه اطلاع در جایگاه های اتصال استفاده می کند؟

زیست شناسان بر اساس تجربه بر این عقیده بودند که دو پروتئین به این توالی متصل می شوند و آن مقدار اطلاع اضافی برای اتصال پروتئین دوم به این جایگاه ها می باشد. سال ها پس از این عقیده اشنایدار با کمک نت هرمن ناحیه دیگری به نام *F plasmid in CD* را بررسی کردند. این جایگاه شامل ۶۰/۲ بیت اطلاع است ولی تنها ۲۰ بیت اطلاع را پیدا کردن آن ها لازم می باشد به عبارت دیگر $R_{sequence}$ =۶۰/۲ و $R_{frequency}$= ۱۹/۶ بیت است. بنابراین نسبت $R_{sequence}$ به $R_{frequency}$ برابر ۳/۰۷ می باشد. آنها بر این گمان بودند که ۳ پروتئین به این جایگاه ها متصل می شوند و با نشان دادند این که سه پروتئین KDa-33, KDa- 37, KDa – 75 به این توالی ها متصل می شوند شگفتی ساز شدند. آنها به این باور رسیدند که می توان از طریق مقایسه ای $R_{sequence}$ و $R_{frequency}$ می نیمم تعداد پروتئین هایی را که به توالی ها در DNA متصل می شوند را پیشگویی کرد (۵ و۷ و۸ و۱۰).

اشنایدر و استفنز در سال ۱۹۹۹ یک روش گرافیکی به نام نمودار توالی را معرفی کردند که با استفاده از این تکنیک قوی ، تحلیل توالی ها را در هر مجموعه توالی از DNA و RNA با پروتئین نمایش داده می شود. در روش نمودار توالی محور افقی جایگاه هر باز و محور عمودی مقدار اطلاع را نشان می دهد.



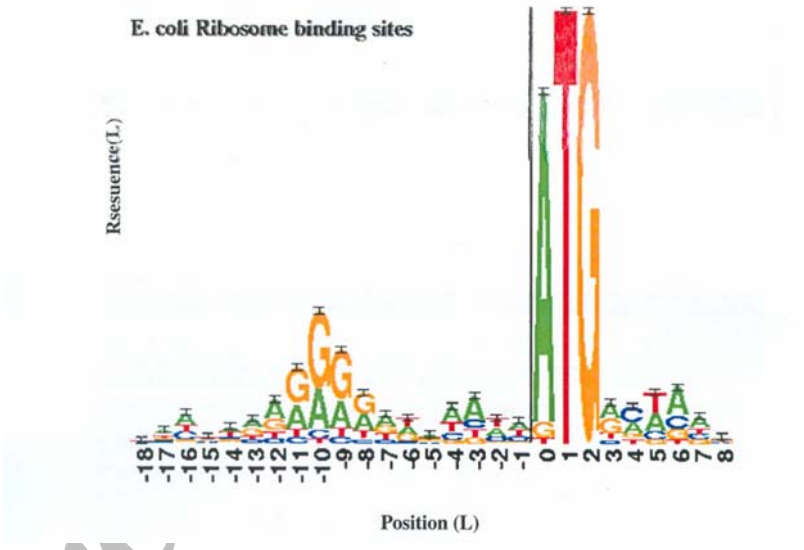
نمودار۳. نمودار توالی

وقتی که یک حرف T با ارتفاع ۲ بیت اطلاع در جایگاه ۷+ وجود دارد، بدان معنی است که تقریبا در تمامی این توالی ها در این جایگاه باز T قرار گرفته است. در این نمودار حروف در هر جایگاه به ترتیب اهمیت شان بروی هم چیده شده اند. بالاترین حرف در هر جایگاه توالی هماهنگ را نشان می دهد .

ارتفاع باز B در جایگاه L $= p(B,L)R_{sequence}(L)$

همانطور که در نمودار۳ مشاهده می شود در جایگاه ۳+ بیشترین تعداد باز مربوط به باز گوانین می باشد اما یک مورد باز آدنین هم وجود دارد بنابراین حرف G هم بلندتر و هم در بالای A قرار گرفته است. حال اگر به ستون ۹- نگاه کنیم حروف GATTTTCTCTTT راکه برای محاسبه آنتروپی از آن استفاده کرده ایم، خواهیم دید. اطلاع در جایگاه ۹- برابر ۰/۵۸ بیت می باشد که با روش تصحیح خطای نمونه گیری به ۰/۳۸ کاهش می یابد. این روش، روش ایده آلی برای تحلیل جایگاه های اتصال در DNA و RNA و پروتئین است.توالی های هماهنگی معرف نوکلئوتیدی است که در اغلب موارد در توالی ها به همان شکل مشاهده می شوند.در حقیقت توالی های هماهنگ توالی های نوکلئوتیدی هستند که در ژنوم دسته وسیعی از موجودات، دیده می شوند. با استفاده از این نمودار شناسایی توالی های هماهنگ براحتی و برابر با بازی است که بیشترین ارتفاع را دارد.

درنمودار۴ نمودار توالی جایگاه های اتصال ریبوزوم در توالی هایی از باکتری *E. coli* را نشان می دهد (۶و۱۱).

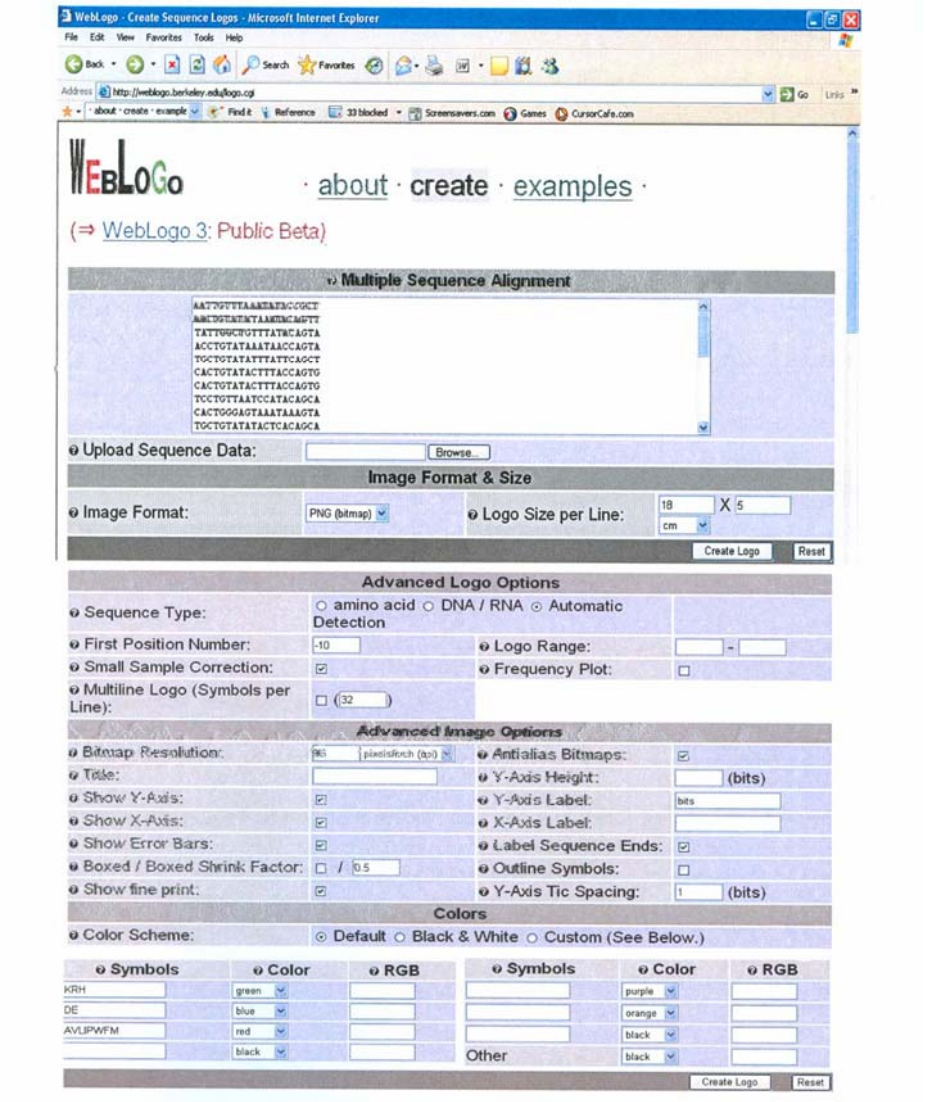


نمودار ۴. توالی جایگاه های اتصال ریبوزوم

این نمودارها از طریق برنامه weblogo قابل ترسیم و از طریق سایت اینترنتی

http://weblogo.berkeley.edu/logo.cgi

قابل دسترسی می باشند. داده های حاصل از توالی های نوکلئیک اسید در قمست *Multiple Sequence Alignment* وارد می شود، سپس پس از انتخاب کردن آیتم مورد نظر از قبیل تعیین محدوده، عنوان، برچسب محورهای x و y، سایز، رنگ نمودار گزینه *create logo* را انتخاب می گردد (نمودار ۵).



نمودار ۵.

تعیین اندازه جایگاه های اتصال

پروتئین ها توالی های خاصی از نوکلئیک ها اسیدها را تعیین و برای انجام عملکردهای به آن ها متصل می شوند. حال این سئوال مطرح است که نحوه شناسایی این محل ها چگونه است؟

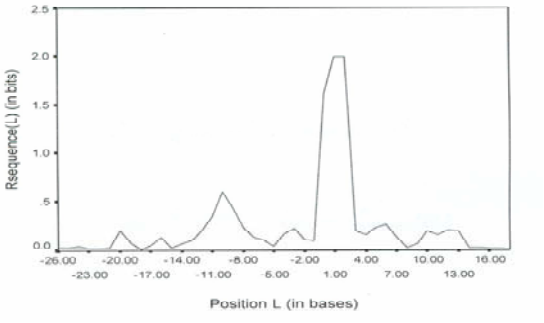
تعیین نواحی این جایگاه ها مشکل است. زیست شناسان از روش های تجربی از قبیل روش ردپا استفاده می کنند. که این روش ناحیه دقیق تماس پروتئین با نوکلئیک اسیدها را تعیین نمی کند. این روش برای تعیین محل نوکلئوتیدهایی که به پروتئین متصل می شوند کاربرد دارد. اشنایدر با استفاده از اندازه اطلاع معرفی شده روشی کمی برای تعیین جایگاه های اتصال معرفی کرد. او به این نتیجه مهم رسید که ((محتوی اطلاع در جایگاه هایی اتصال پروتئین ها به توالی های نوکلئیک اسیدها دارای مقدار بزرگتر از صفر و در خارج از این جایگاه ها به سمت صفر کاهش می یابد))(۸).

جدول ۱، ۱۴ توالی از جایگاههای اتصال پروتئین LexA به توالی هایی از باکتری *E. coli* را نشان می دهد. این توالی ها که مربوط به ژن های SOS می باشند در حالت عادی توسط مهار کننده LexA پوشانیده و این مهار کننده از بیان ژن های SOS جلوگیری می نماید. به طور طبیعی LexA با جعبه های SOS پیوند برقرار نموده و رونویسی را کاهش و یا متوقف می نماید. با غیر فعال شدن LexA، جعبه های SOS از قید سد کننده ها رها می شوند و رونویسی از ژن ها شروع می شود. انجام عمل رونویسی منجر به سنتز انواع

آنزیم های مورد نیاز در سیستم های ترمیم می شود با ختم و ترمیم و از بین رفتن محل آسیب دیدگی، recA غیر فعال شده و این امر سبب سنتز LexA و مهار مجدد جعبه هایSOS می شود (۴).

نتیجه گیری

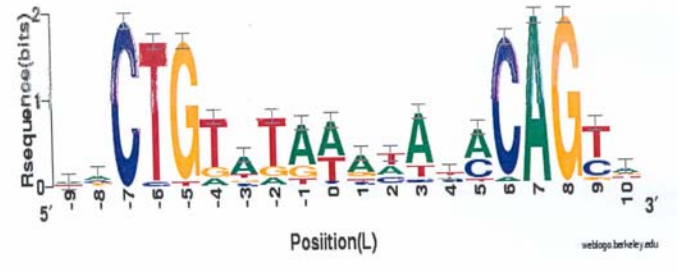
بنابراین پروتئین LexA به دلایلی که در بالا شرح آن گفته شده به این توالی متصل می شود حال برای بررسی این جایگاه ۵ باز به دو انتهای ناحیه پیشنهاد شده بوسیله روش تجربی اضافه می کنیم. محتوای اطلاع برای هر محل بر اساس رابطه $R_{sequence}(L)=2-H_s(L)$ محاسبه شده است. نمودار۶ منحنی $R_{sequence}(L)$ را نشان می دهد.



نمودار۶. منحنی $R_{sequence}(L)$

محدوده اتصال جایگاه اتصال پروتئین LexA به این توالی ها از ۹- تا ۱۰+ می باشد با جمع کردن مقادیر $R_{sequence}(L)$ در محدوده اتصال پروتئین، اندازه کل اطلاع یعنی $R_{sequence}$ حاصل می شود. که برای پروتئین LexA برابر ۲۱/۱ بیت می باشد. با محاسبه $R_{frequency}$ می بینیم که این جایگاه دارای اطلاع کافی برای پیدا شدن آنها توسط پروتئین LexA می باشند. برای محاسبه این مقدار همانطور که در جدول۲ ملاحظه می شود دو مقدار γ , G لازم می باشد. یک برآورد خوب برای اندازه ژنوم باکتری *E. coli* ۳/۹ × ۱۰^{۶} جفت باز می باشد. حداقل ۱۱ ژن تحت کنترل پروتئین LexA می باشند. می بینیم مقدار برای γ، ۱۱ و کران بالا برای $R_{frequency}$، ۱۸/۴ بیت است و نسبت $R_{frequency}$ به $R_{sequence}$ برابر ۱/۱ می باشد. بنابراین محتوی اطلاع در این جایگاه ها برای شناسایی آن ها توسط تشخیص دهنده ی LexA کافی می باشد.

لذا در این تحقیق نشان داده شد ازآنجایی که پروتئین های زیادی برای انجام عملکردهای گوناگون از قبیل روشن و خاموش کردن ژن ها، رونویسی، ترجمه مکان های خاصی از توالی های نوکلئیک اسید در ژنوم را تعیین و به آنها متصل می شود، نواحی اتصال در کدام قسمت طول ژنوم قرار گرفته و نحوه شناسایی این جایگاه ها که موسوم به اتصال است می تواند بر اساس معیار آنتروپی و نظریه اطلاع مشخص گردیده و با استفاده از نظریه اطلاع شانون راهکاری دقیق تر برای تشخیص نواحی ارئه شده است. داده های جایگاه اتصال پروتئین LexA در توالی های DNA باکتری اشرشیاکلای از سایت http://predoric.tu-bs.de/inden.php?iden انتخاب شده است، که ابتدا احتمال قرار گرفتن هر باز را در جایگاه محاسبه سپس با استفاده از رابطه مقدار عدم قطعیت (آنتروپی) و با محاسبه $\frac{\sigma}{\sqrt{n}}$ مقدار خطای حاصل از نمونه گیری را محاسبه می کنیم. مقدار انحراف استاندارد حاصل از ۲۵ جایگاه اتصال پروتئین LexA برابر ۰/۵۷ و بنابراین خطای نمونه گیری برابر ۰/۱۱ بیت شده است. با جمع کردن این دو مقدار با عدم قطعیت بدست آمده از رابطه $R=H_{befor}-H_{after}$ این مقدار برابر ۱۷/۳۹ می شود. با رسم نمودار توالی از برنامه Weblogo نمودار۷ نحوه قرار گرفتن بازها نشان داده می شود.



نمودار۷. نحوه قرار گرفتن بازها

در این نمودار در جایگاه ۷- و ۶- و ۵- به ترتیب باز G,T,C قرار گرفته شده است. درنتیجه بنابر تعریف شانون، عدم قطعیت در مورد این که چه بازی انتظار می رود در جایگاه اتصال دیگری از این پروتئین در باکتری *E. coli* در این مکان ها قرار گیرد صفر است. البته این قطعیت در مورد مکان های ۶+ و ۷+ و ۸+ نیز دیده می شود. ولی در مکان های دیگر این جایگاه با احتمال قرار گرفتن بازهای مختلف عدم قطعیت بیشتری وجود دارد.

با استفاده از این نمودار توالی ها هماهنگ CTG, CAG که توسط زیست شناسان بطور تقریبی تشخیص داده می شد قابل رویت است.

حال برای بررسی اینکه این جایگاه دارای اطلاع کافی برای پیدا شدن آن توسط پروتئین LexA هستند مقدار اطلاع لازم برای پیدا کردن این جایگاه یعنی $R_{frequency}$ را محاسبه می کنیم. مقدار G,γ مورد نیاز برای محاسبه در باکتری *E. coli* به ترتیب برابر ۳۹۰۰۰۰۰ و ۲۵۴۷ و مقدار $R_{frequency}$ برابر ۱۸/۴ بیت می باشد.

با محاسبه نسبت $R_{sequence}$ به $R_{frequency}$ این مقدار برابر ۰/۹۵ بدست می آید که نزدیک به یک است، لذا این جایگاه دارای اطلاع کافی برای پیدا شدن آنها توسط پروتئین LexA می باشد.

منابع

۱- امتیازی، گ. کریمی، م. (۱۳۸۰) مبانی زیست مولکولی و مهندسی ژنتیک

۲- ارقامی ن ، پور عبدالله نژاد م .ع (۱۳۷۷) نظریه مقدماتی اطلاع مرکز نشر دانشگاهی

۳- آذرنوش ح (۱۳۸۰) نظریه اطلاع انتشارات دانشگاه فردوسی مشهد

۴- سیدنا (۱۳۸۰) ژنتیک موسسه انتشارات امید

5- Multihac R; Cicuttin A; Multihac R. C, (2001). Entropic Approach to Information Coding in DNA molecules, Materials Science and Engineering C.**18**:51-60.

6- Schneider T. D; Stephens R. M, (1990) Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. **18**: 6097-6100.

7- Schneider T.D, (1988). Information and entropy of patterns in genetic switches, in Erickson, Smith, maximum entropy and Bayesian methods in science and engineering applications , vol.2, kluwer Academic Publishers, Dordrecht,147-154.

8- Schneider T. D; Stormo G. D; Gold L, (1986). Information content of binding sites on nucleotide sequence J. Mol. Biol., **188**:415-431.

9- Schneider T. D. Some lessons for molecular biology from information theory, in entropy measure, maximum entropy principle and emerging applications, Springer- Verlag, New York. 229-237

10- Schneider, T. D, (2000). New approach in mathematical biology: information theory and molecular machines a brief 9 page introduction to molecular information theory. Oct 13 Shannon, C. E., Bell. System tech. j., **27**: 379-423,623-656

11- Shaner M. C; Blair I. M; Schneider T. D. Sequence logos a powerful, yet simple tool proceeding of the twenty – sixth annual Hawaii international conference on system sciences, volume 1, architecture and biotechnology computing, 813-821