مقاله علمی پژوهشی/نشریه پژوهش‌های علوم دامی/ جلد ۳۱ شماره ۲/ سال ۱۴۰۰/ صفحات ۲۹ تا ۴۳

DOI:10.22034/AS.2021.30510.1464

بررسی همبستگی‌های دامنه بلند DNA در ژن‌های مؤثر بر تولید شیر گاو

رکسانا آباده[۱]، مهدی امین افشار[۲]، مصطفی قادری زفره‌ای[۳*]، سید عباس محمدی[۴] و محمد چمنی[۵]

تاریخ دریافت: ۱۳۹۷/۰۹/۰۱ تاریخ پذیرش: ۱۳۹۹/۰۷/۰۸

[۱]دانشجوی دکتری، گروه علوم دامی، دانشکده علوم کشاورزی و صنایع غذایی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

[۲]استادیار، گروه علوم دامی، دانشکده علوم کشاورزی و صنایع غذایی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران،ایران

[۳]دانشیار، گروه علوم دامی، دانشکده علوم کشاورزی، دانشگاه یاسوج، یاسوج، ایران

[۴]دانشیار، گروه ریاضی، دانشکده علوم پایه، دانشگاه یاسوج، یاسوج، ایران

[۵]استاد، گروه علوم دامی، دانشکده علوم کشاورزی و صنایع غذایی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

* مسوول مکاتبه: Email: mghaderi@yu.ac.ir

چکیده

زمینه مطالعاتی: وجود همبستگی‌های دامنه بلند در ملکول DNA اشاره به وجود فرآیندهای بازآرایی یا مضاعف شدگی DNA دارد. این نوع پدیده‌ها کاربرد مستقیم در اصلاح نژاد ندارند و بیشتر در بررسی‌های تکاملی به کار می‌روند. **هدف**: در این پژوهش فرض شد که با استخراج همبستگی‌های دامنه بلند DNA بین تمامی نوکلئوتیدهای مختلف درون یک ژن می‌توان به درجه‌ای از ارتباط بین آن‌ها در وهله اول دست یافت و از این‌رو، ممکن است پژوهش‌های متکی به کشف SNP را بهتر می‌توان جهت دهی کرد. **روش کار**: ۲۴ ژن از ژن‌های موثر بر تولید شیر گاو در این پژوهش انتخاب شدند. توالی، طول، شماره دست یابی، تعداد و طول هر اگزون و جایگاه آن بر روی کروموزوم از بانک ژنی NCBI دریافت و توالی‌ها با فرمت FASTA ذخیره شدند. با استفاده از نرم افزاری که قبلا با زبان C# طراحی شده بود با توجه به خواسته پژوهش، شماره دسترسی ژن‌های مورد بررسی وارد گردید و خروجی مناسب به دست آمد. برای محاسبه همبستگی‌های دامنه بلند DNA ژن‌های مورد بررسی، از نرم افزار CorGen استفاده شد. **نتایج**: سطح معنی‌داری از همبستگی دامنه بلند در توالی DNA ژن هایی مانند *EZR, FGG, KRT6A, RAB1A, EIF3L, TBC1D20, ZNF419, S100A16, MRPL3, TPPP3, PHF10* وجود دارد. توان کاهشی حاصل از برازش تابع قانون توان روی همبستگی‌های دامنه بلند بدست آمده از ژن‌ها با طول‌های متفاوت، در دامنه ۰/۱٤٦و ۰/٦٤۳ قرار داشتند. **نتیجه گیری نهایی**: می‌توان نتیجه گرفت که کاهش میزان همبستگی‌های دامنه بلند با افزایش فاصله بین بازه‌های توالی DNA از روند تصادفی پیروی نمی‌کنند. بنابراین، هندسه فرکتال طبیعت نیز در این ژن‌ها دیده می‌شود. ژن‌های مورد بررسی پیچیدگی بالا و مقیاس ناوردایی را در DNA خود دارند. همچنین مشخص شد میزان بسامد حاصل از همبستگی‌های دامنه بلند در ژن‌ها متفاوت اما نزدیک به هم بود. پیشنهاد می‌شود این نواحی از نظر وجود عدم تعادل پیوستگی مورد کنکاش بیشتری قرار گیرند.

واژگان کلیدی: توان کاهشی، همبستگی‌های دامنه بلند، گاوشیری، هندسه فراکتال

مقدمه

ابع خودهمبستگی[1]به طور گسترده‌ایی در نظریه فرسته[2] و فیزیک به عنوان معیار وابستگی خطی و بسامد[3]به کار رفته است. کاربرد این روش در تحلیل توالی DNA در سال ۱۹۹۲ با پیدا کردن همبستگی‌های قانون توان[4]در توالی DNA اهمیت بیشتری پیدا کرد. وجود همبستگی‌های قانون توان نشان می‌دهد که پیچیدگی بالا و مقیاس ناوردا[5]در توالی‌های DNA موجود است (پنگ و همکاران۱۹۹۲، وس ۱۹۹۲). در گذشته نبود توالی‌های بلند مانعی برای تخمین مستقیم همبستگی در طول DNA محسوب می‌شد (برنائولا و همکاران ۲۰۰۲). در نتیجه کاربرد روش‌های غیر مستقیم منجر به نتایجی می‌شد که قابل مقایسه نبودند. امروزه با در دست بودن توالی‌های بلند DNA و حتی کل ژنوم یک سازواره در پایگاه‌های داده مثل NCBI؛ امکان محاسبه توابع خود همبستگی به طور مستقیم روی DNA وجود دارد. همبستگی‌های دامنه‌بلند در گروه توابع خود همبستگی، گروه‌بندی می‌شود. در کل با در دست داشتن توالی عددی= S $\{x_1,x_2,x_3,\dots x_n\}$ با واریانس $\sigma^2=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2$، تابع خود همبستگی در فاصله l به صورت ذیل تعریف می‌شود: $C(l)=\frac{1}{\sigma^2}\left[\frac{1}{\mathrm{N}-l}\sum_{i=1}^{N-l}x_ix_{i+l}-\frac{1}{(\mathrm{N}-l)^2}\sum_{i=1}^{N-l}x_i\sum_{i=1}^{N-l}x_ix_{i+l}\right]$. تابع خود همبستگی انحراف $<x_ix_{i+l}>$ از $><x_i><x_{i+l}>$ را اندازه می‌گیرد (برنائولا و همکاران ۲۰۰۲). اگر این انحراف برابر با صفر باشد نشان می‌دهد که ارتباطی بین جایگاه iام x_i با جایگاه lام x_{i+l} که در فاصله l از هم در توالی S وجود ندارد. هر چقدر که این انحراف بزرگتر باشد، همبستگی‌های خطی بین جایگاه‌هایی که با فاصله l از هم قرار دارند، قوی‌تر است. $C(l)>0$ نشان می‌دهد احتمال اینکه x_i و x_{i+l} به مقادیر مشابهی برسند نسبت به یک توالی تصادفی هم اندازه، بیشتر است؛ در حالی که $C(l)<0$ نشان می‌دهد که این احتمال نسبت به یک توالی تصادفی هم اندازه کمتر است. طبق تعریف وقتی $C(l=0)\equiv 0$ باشد بدان معنی است که درباره وجود همبستگی‌ها در توالی اطلاعی وجود ندارد. در $C(l)$ هیچگونه فرض کلی بر ایستا بودن ندارد؛ زیرا که واریانس کل توالی را در نظر می‌گیرد. با این حال امکان محاسبه $C(l)$ برای حالت‌های غیرایستا توالی نیز وجود دارد که می‌توان از آن اطلاعات مربوط به وجود ناهمگنی در توالی را استخراج کرد (برنائولا و همکاران ۲۰۰۲).

برای محاسبه تابع خود همبستگی روی یک توالی DNA، ابتدا باید نوکلوتیدهای (A,T,C,G) ملکول DNA به کمیت‌های عددی تبدیل شوند. در این تبدیل یا نگاشت باید دقت شود چرا که اگر هر مقدار عددی اختیاری به هر نوکلئوتید اختصاص داده شود، آنگاه همبستگی‌های حاصله، وابسته به مقدار اختصاص داده شده خواهد بود، یعنی این نوع نگاشت از نوکلئوتید به مقادیر عددی، می‌تواند مشکل‌ساز باشد. هرگاه نگاشت دو دویی باشد، مشکل یاد شده وجود نخواهد داشت؛ چرا که در آن صورت تابع خود همبستگی برای تمامی مقادیر عددی اختصاص یافته، یکسان خواهد بود. در کل هفت نوع قانون نگاشت دودویی وجود دارد (برنائولا و همکاران ۲۰۰۲) که معمولا در محاسبات DNA به کار می‌روند. جدول ۱ این قوانین نگاشتی را نشان می‌دهد.

Table 1- DNA bases mapping rules into binary sequences

Rule	Mapping
SW	C or G = 1 A or T = 0
RY	A or G = 1 C or T = 0
KM	G or T = 1 A or C = 0
A	A = 1 T, C or G = 0
T	T = 1 A, C or G = 0
C	C = 1 A, T or G = 0
G	G = 1 A, T or C = 0

[1] Autocorrelation function
[2] Signal theory
[3] Periodicity
[4] Power-law correlations
[5] Scale invariance

یکی از ویژگی‌های مهم ژنوم اغلب یوکاریوت‌ها، وجود همبستگی‌های دامنه بلند بین نوکلئوتیدهای ملکول DNA آن می‌باشد. همبستگی‌های دامنه بلند DNA معمولا همراه با قانون توان[1]و معمولا به صورت یک مفهوم واحد بیان می‌گردند. قانون توان یک رابطه خاص ریاضی به صورت $P(x) \propto x^{-\alpha}$ است که در آن α توان واپاشی یا فراسنجه مقیاس‌بندی[2]است که معمولا $2 < \alpha < 3$ قرار می‌گیرد (مهنتی و رائو ۲۰۰۲). این قانون ارتباط بین دو متغیر را نشان می‌دهد به طوری که یک متغیر به صورت توان متغیر دیگر، تغییر می‌کند. نکته قابل توجه این است که این تغییر، مستقل از مقادیر اولیه دو متغیر مورد نظر است. یکی از ساده‌ترین مثال‌های این مفهوم، ارتباط بین اندازه ضلع یک مربع با مساحت مربع یا ارتباط بین ضلع مکعب و مساحت آن است. اگر ضلع مربعی را دوبرابر کنیم مثل این است که مساحت آن مربع را در ٤ ضرب کرده‌ایم. آنچه اهمیت دارد این است که ضرایب یاد شده به اندازه اولیه مساحت‌ها، بستگی ندارد. در رایانش‌های DNA، معمولا پژوهشگر نیاز به یک مدل پایه یا فرض صفر دارد. این فرض صفر متکی بر ایجاد توالی تصادفی از DNA است. وجود همبستگی‌های دامنه بلند ممکن است ایجاد فرض صفر مبنی بر تصادفی بودن قرارگیری نوکلئوتیدها درDNA و عدم ارتباط بین آن‌ها را با مشکل روبرو کند. وجود همبستگی دامنه بلند DNA در بیوانفورماتیک گزارش شده است. مثلا در رایانش‌های همردیفیDNA که متکی به P-Value هستند، میزان P-Value نمره همردیفی را تحت تاثیر قرار می‌دهد و تشخیص فاکتورهای بیانی در توالی‌های ژنومی را نسبت به توالی‌های تصادفی مستقل کاهش می‌دهد. با نگاهی به اهداف فرضیه‌های متفاوت در پژوهش‌های انجام شده در سطح DNA، می‌توان مشاهده کرد که استخراج همبستگی‌های دامنه بلند روی توالیDNA در پژوهش‌های زیادی مورد بررسی قرار گرفته است (آرنئودو و همکاران ۱۹۹۸، آیودیت و همکاران ۲۰۰۱، چتزیدیمیتریو و همکاران ۱۹۹٤، لی ۱۹۹۷، مهنتی و رائو ۲۰۰۲، پنگ و همکاران ۱۹۹۲ و سوتهیباتپونگ ۲۰۱٦). از آنجایی که این پژوهش‌ها بیشتر درحوزه‌های غیر اصلاح نژادی صورت گرفته است، در این پژوهش برای اولین بار، ژن‌های موثر در تولید شیر گاو از نظر همبستگی‌های دامنه بلند مورد کنکاش قرار می‌گیرد. لذا با استفاده از روش ساده بین‌رشته‌ایی سعی می‌گردد دریافت‌ها و مفاهیم اصلاح نژادی از استخراج همبستگی‌های دامنه بلند DNA حاصل گردد. همچنین کمکی که استخراج همبستگی‌های دامنه بلند DNA می‌تواند به تشخیص علایم انتخاب تکاملی در سطح ژنوم داشته باشد، توضیح داده خواهد شد.

مواد و روش‌ها

گزارش شده است حدود ٦٨٧٥ ژن در بافت غدد پستانی گاو بیان می‌شوند که در تولید شیر موثر هستند (لمی و همکاران ۲۰۰۹). ژن‌های مورد پژوهش در این بررسی از گروه‌بندی‌هایی ژنی _ بر حسب حاشیه نویسی ژنومی هر ژن که لمی و همکاران (۲۰۰۹) ایجاد کرده بودند- انتخاب شدند. در این راستا از هر یک از شش گروه موجود، به طور تصادفی چهار ژن انتخاب شد. توالی و همچنین سایر اطلاعات ژن‌ها از جمله طول ژن، شماره دستیابی، تعداد و طول هر اگزون و جایگاه آن بر روی کروموزوم از بانک ژنیNCBI دریافت و با فرمت FASTA ذخیره شدند (جدول۲). به دلیل حجم زیاد اطلاعات ژن‌ها و اگزون‌های مربوط به آن، از نرم افزاری که قبلا با زبان C# طراحی شده بود،[3] برای آماده‌سازی اطلاعات استخراج شده استفاده شد. در این نرم افزار با توجه به خواسته پژوهش، شماره دسترسی ژن‌های

[1] Power law
[2] Scaling parameter
[3] پژوهشگران، در صورت نیاز، برای دستیابی به نرم افزار با نویسنده مسوول این مقاله مکاتبه کنند.

مورد بررسی به عنوانی ورودی به کار رفت و خروجی مناسب بدست آمد. برای محاسبه همبستگی‌های دامنه بلند DNA ژن‌های موثر روی تولید شیر از الگوریتم یا نرم افزار CorGen استفاده شد. به طور کلی این الگوریتم ابتدا میزان GC یک رشته DNA را محاسبه می‌کند، همبستگی دامنه بلند روی GC را بدست می‌آورد و در نهایت تابع قانون توان – که در بالا ذکر شد – را روی نتایج همبستگی دامنه بلند حاصل از GC برازش می‌دهد. این الگوریتم یک روش ساده پویا است و بر اساس مضاعف شدن تک نوکلوتیدی و فرایندهای جهشی در طول رشته DNA استوار است (هم مضاعف شدگی و هم فرآیندهای جهشی، از فرآیندهای شناخته شده تکامل ملکولی هستند). الگوریتم یاد شده که متکی به مدل جهش – مضاعف است، تمام مزایای ذیل را در بر دارد: نتایج دقیق برای تابع همبستگی روی توالی DNA را حاصل می‌کند، مدل یاد شده امکان بررسی هر ترکیبی از توان کاهشی[1] محتوای مطلوب GC و هر طولی از DNA را ارایه می‌دهد، بسامد[2]همبستگی محاسبه شده انقدر بزرگ است که بتواند با همبستگی‌های قوی ژنومی همآوردی کند و به راحتی می‌توان با هر مقداری که کاربر تعیین می‌کند، آن را کاهش داد و این الگورتیم را می‌توان با هر رایانه‌ای با زمان اجرای O (N) اجرا کرد.

نتایج و بحث

بیست و چهار ژن در این پژوهش مورد کنکاش قرار گرفت و اطلاعات حاصل از استخراج ویژگی‌های این ژن‌ها (جدول ۲) نشان داد که بین طول ژن‌ها اختلافات زیادی وجود دارد به طوری که بیشترین طول ژنی، به ترتیب مربوط به ژن‌های *CDYL* و *MRPL3* بود و کمترین طول ژنی مربوط به ژن‌های *YWHAH* و *APRT* بود. سایر مشخصات در جدول ۲ آمده است. این مشخصات، ناهمگنی زیاد داده‌های مورد بررسی را نشان می‌دهد. در کل داده‌های متکی به DNA ماهیتی ناهمگن دارند و این یکی از پیچیدگی‌های بررسی این نوع داده‌ها به شمار می‌رود. گروه بندی که به اسم "گروه" در ستون اول جدول ۲ آمده است نشان دهنده شش گروه ژنی است که از این گروه‌ها به طور تصادفی چهار ژن استخراج شده است.

Table 2- Results and characteristics of the genes studied in this study

GC Contents	Amplitude	Power decay	Total annotated spliced exon length	Map location chromosome	Exon count	length	Gene symbol	Protein	mRNA	Group
0.50	0.00662	0.540	2704	9	13	45027	*EZR*	NP_776642.1	NM_174217	Milk protein
0.38	0.00353	0.275	1611	17	9	7472	FGG	NP_776336.1	NM_173911	Milk protein
0.49	0.00683	0.311	2166	5	9	5254	*KRT6A*	NP_001076979.1	NM_001083510	Milk protein
0.36	0.00428	0.453	1328	11	5	28842	*RAB1A*	NP_001028800.1	NM_001033628	Milk protein
0.44	0.00591	0.643	1932	5	13	22981	*EIF3L*	NP_001030373.1	NM_001035296	Virgin
0.44	0.00563	0.622	3346	13	8	18982	*TBC1D20*	NP_001033118.1	NM_001038029	Virgin
0.51	0.00690	0.291	2137	18	5	7336	*ZNF419*	NP_001095402.1	NM_001101932	Virgin
0.56	0.00697	0.315	1061	3	3	6219	*S100A16*	NP_001068686.1	NM_001075218	Virgin
0.39	0.01080	0.622	1356	1	10	57201	*MRPL3*	NP_001073786.1	NM_001080317	Pregnancy

[1] Decay exponent

[2] Amplitude

.58	0.01204	0.518	1032	18	4	3580	*TPPP3*	NP_001029946.1	NM_001034774	Pregnancy
0.42	0.00614	0.431	1583	9	12	20137	*PHF10*	NP_001033141.1	NM_001038052	Pregnancy
0.50	0.00974	0.368	991	15	4	4245	*MRPL16*	NP_001029813.1	NM_001034641	Pregnancy
0.48	0.00324	0.174	1538	23	7	4374	*RRP36*	NP_001098949.1	NM_001105479	lactation
0.47	0.00485	0.146	2413	10	9	12595	*FAM161B*	NP_001019662.2	NM_001024491	lactation
0.51	0.00612	0.252	3068	8	1	3068	*SLC25A37*	NP_001096025.1	NM_001102555	lactation
0.48	0.01150	0.631	3298	23	7	97274	*CDYL*	NP_001095693.1	NM_001102223	lactation
0.45	0.00565	0.146	3474	X	5	4824	*ARMCX3*	NP_001179382.1	NM_001192453.2	Involution

Table 2- Results and characteristics of the genes studied in this study (Continued)

GC Contents	Amplitude	Power decay	Total annotated spliced exon length	Map location chromosome	Exon count	length	Gene symbol	Protein	mRNA	Group
0.47	0.01239	0.605	1329	22	9	35176	*SEC13*	NP_001069033.1	NM_001075565	Involution
0.63	0.00627	0.221	1256	5	7	3358	*TPI1*	NP_001013607.1	NM_001013589	Involution
0.59	0.00617	0.511	1852	19	6	4016	*ARHGDIA-15*	NP_788823.1	NM_001035401	Involution
0.46	0.00875	0.697	1054	2	2	4219	*LOC515042*	NP_001098840.2	NM_001105370	Mastitis
0.60	0.00689	0.148	978	18	5	2777	*APRT*	NP_001020505.1	NM_001025334	Mastitis
0.60	0.00732	0.451	1389	19	19	3834	*KRT19*	NP_001015600.1	NM_001015600	Mastitis
0.51	0.01972	0.569	1445	17	1	1445	*YWHAH*	NP_776917.2	NM_174492	Mastitis

به جهت اینکه همبستگی دامنه بلند قابل محاسبه باشد، بایستی طول آن به اندازه کافی بلند باشد (۱۰۰ نوکلئوتید) که در این پژوهش، اندازه ژن‌ها به قدر کافی بلند بودند (جدول ۲). بررسی نوسان‌های آماری در توالی‌های DNA می‌تواند اطلاعات ارزشمندی درباره سازماندهی و عملکرد ژنوم‌ها را ایجاد کند. اشکال ۱ تا ۷ پراکنش محتوای GC (نمودار بالا) و همبستگی‌های دامنه بلند (نمودار پایین) محاسبه شده را برای هر ژن نشان می‌دهد. در همبستگی‌های دامنه بلند (نمودار پایین) هر نگاره، تابع توان به رنگ سبز روی همبستگی دامنه بلند محاسبه شده، نشان داده شده است. نرم افزارCorGen از GC برای محاسبه همبستگی توالی دامنه بلند استفاده می‌کند. محتوی GC در کنکاش یک ژن بسیار مهم است. محتوی GC علاوه نوع خواص فیزیکی – شیمیایی که روی DNA اعمال می‌کند، یک ویژگی مهم در مدلسازی روی توالی DNA به شمار می‌رود (چتزی دیمیتریو و همکاران ۱۹۹٤). بر اساس نتایج بدست آمده در این پژوهش، ژن‌های مورد بررسی روند تقریبا تصادفی را

از نظر محتوای GC نشان دادند. به طوری که نمی‌توان دو ژن را یافت که تقریبا از یک روند محتوی GC پیروی کنند. به عنوان مثال، دو ژن *Mitochondrial ribosomal protein L3(MRPL3)* و *Armadillo repeat containing X-linked3*(ARMCX3) با افزایش طول DNA روند کاهشی در محتوی GC را از خود نشان دادند در حالی که ژن‌های *keratin 6A(KRT6A)* و *fibrinogen gamma chain(FGG)* با افزایش طول DNA، روند افزایشی در محتوی GC را نشان دادند. به عبارتی دیگر، این دو ژن در نواحی انتهایی ژن، از نظر محتوای GC غنی‌تر هستند. البته، این تفاوت در میزان همبستگی دامنه بلند محاسبه شده، بازتاب نیافته است. به طوری که در این ژن‌ها، میزان و شکل همبستگی دامنه بلند محاسبه شده، مشابه می‌باشد.

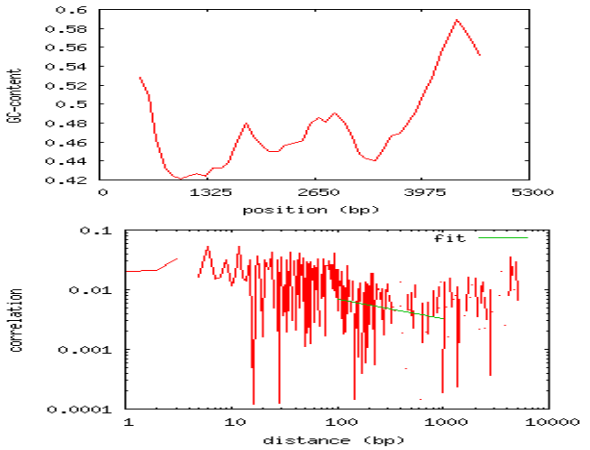


Figure 3- The above diagram, the distribution of the GC content of the *Keratin 6A (KRT6A)* gene and the below diagram of the amplitude correlation between nucleotides of the genes

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

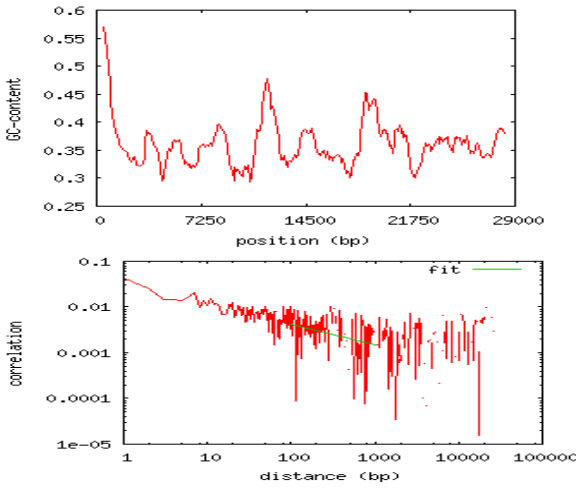


Figure 4- The above diagram, the distribution of the GC content of the *RAB1A member of the RAS oncogene family (RABIA)* gene, and the below diagram of the amplitude correlation between nucleotides of the gene
In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

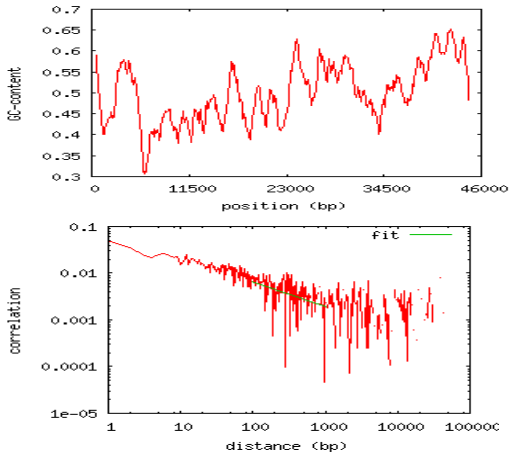


Figure 1- The above diagram, the distribution of the GC content of the *Ezrin (EZR)* gene and the below diagram of the amplitude correlation between nucleotides of the genes

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

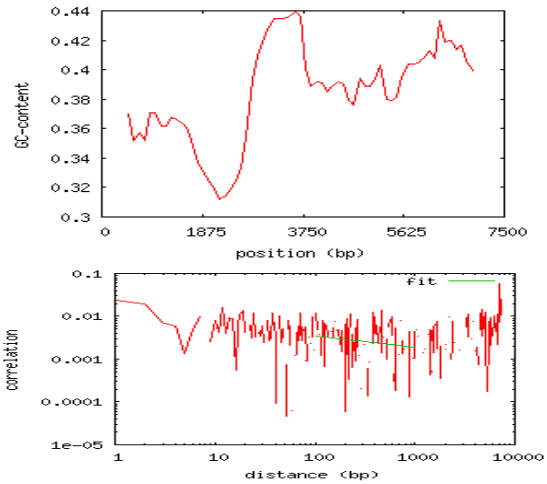


Figure 2- The above diagram, the distribution of the GC content of the *Fibrinogen gamma chain (FGG)* gene and the below diagram of the amplitude correlation between nucleotides of the gene

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

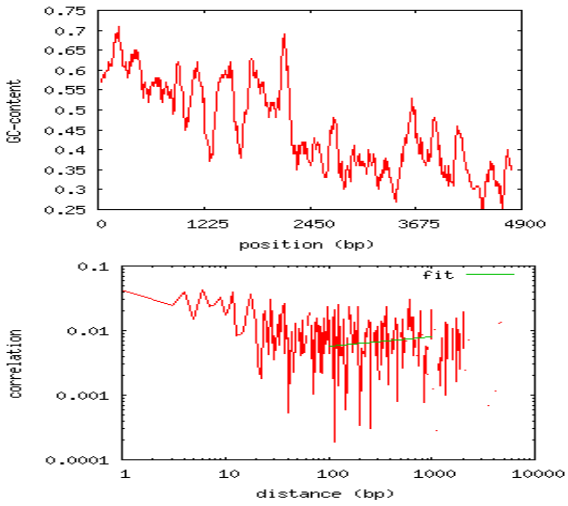
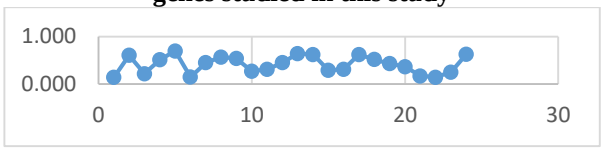


Figure 7- The above diagram, the distribution of the GC content of the *Armadillo repeat containing X-linked 3(ARMCX3)* gene and the below diagram of the amplitude correlation between nucleotides of the genes

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

Figure 8- Decay power distribution in different genes studied in this study



نکته مهم این است که $C(l)$ فقط همبستگی‌های خطی را اندازه می‌گیرد. برای محاسبه همبستگی‌های غیر خطی، نیاز است که عباراتی با درجات بالاتر، یعنی ضرب‌ها، که در آن توان‌های بالاتر x_i و x_{i+l} ظاهر می‌شوند، را لحاظ کنیم. از همین روی در مواردی به $C(l)$، تابع خود همبستگی درجه دوم نیز اطلاق می‌شود. با این وجود از دیرباز مشخص شده که همبستگی‌ها در DNA اساسا خطی هستند و بنابراین $C(l)$ ابزار خوبی برای تحلیل توالی DNA است (برنائولا و همکاران ۲۰۰۲). وجود همبستگی‌های بلند دامنه در DNA می‌تواند اساس بسیاری از محاسبات متکی به DNA را با چالش روبرو سازد. ثابت شده است که همبستگی‌های دامنه بلند DNA می‌تواند روی نمره حاصل از همترازی DNA اثر بگذارد (مسر و همکاران ۲۰۰۷). در چند دهه اخیر چندین روش

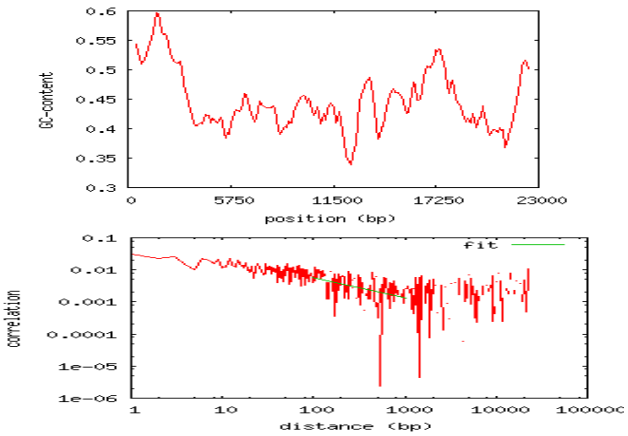


Figure 5- The above diagram, the distribution of the GC content of the *eukaryotic translation initiation factor 3 subunit L (EIF3L)* gene, and the below diagram of the amplitude correlation between nucleotides of the genes

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

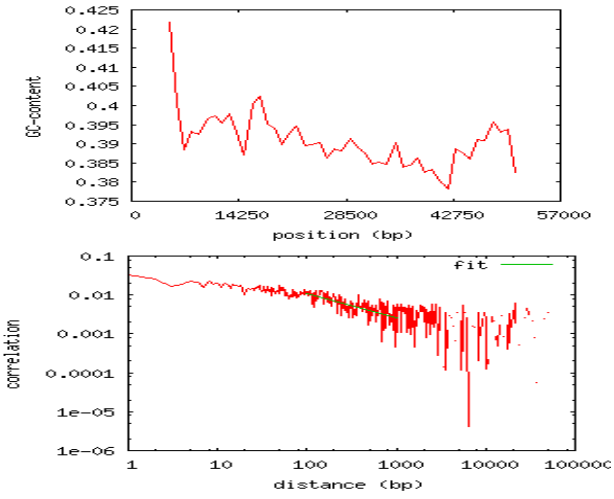


Figure 6- The above diagram, the distribution of the GC content of the *Mitochondrial ribosomal protein L3 (MRPL3)* gene and the below diagram of the amplitude correlation between the nucleotides of the genes

In the double logarithmic transformation, the correlation of the power function is indicated by its straight line (the green line in the below diagram).

برای خوشه‌بندی ژن‌ها و پروتئین‌ها پیشنهاد شده است، اغلب این روش‌ها بر اساس همترازی ژن‌ها بنا شده‌اند که با استفاده از سامانه‌های امتیازدهی بدست می‌آیند و تعدادی از فیلوژنی‌ها بوسیله این روش‌ها ساخته شده‌اند (ادگار و باتزوگلو ۲۰۰۶، کاتو و همکاران ۲۰۰۲ و لارکین و همکاران ۲۰۰۷). در این راستا همترازی چندگانه نقشی اساسی در مقایسه توالی‌ها بازی می‌کند و به صورت معمول برای خوشه‌بندی توالی‌های DNA و پروتئین استفاده می‌شود (وارنو ۲۰۱۳). اما همترازی چندگانه پیچیدگی محاسباتی و حافظه بالایی را برای توالی‌های با طول بزرگ طلب می‌کند (بلایسدل و همکاران ۱۹۸۶، ادگار و باتزوگلو ۲۰۰۶ و کمنا و نوتردام ۲۰۰۹). اخیرا برای اولین بار یک روش فیلوژنی آزاد از همترازی توالی DNA ارایه شده است. این روش بطور وسیعی اینک در آنالیز ژنومی به عنوان یک روش همترازی آزاد به کار می‌رود (کامین و ورزوتو ۲۰۱۲، جان و همکاران ۲۰۱۰، سیمز و همکاران ۲۰۰۹ و وینگا ۲۰۱۳). با توجه به اینکه در اغلب روش‌های فیلوژنی آزاد از همترازی، اطلاعات ساختاری و عملکردی توالی‌های DNA در نظر گرفته نمی‌شود، روش‌های مختلف و جدیدی برای ساخت فیلوژنی پیشنهاد گردیده است مثل آنالیز بر اساس مولفه‌ها[1] (ادواردز و همکاران ۲۰۰۲)، روش تجزیه مقادیر منفرد[2](استوارت و همکاران a۲۰۰۲ و استوارت و همکاران b۲۰۰۲)، روش دستوری دینامیک[3]و روش مدل مارکف[4] (یو و همکاران ۲۰۰۵) و روشهای فراکتال[5](یو و همکاران ۲۰۰۳، یو و همکاران ۲۰۰۵). بنابراین، با توجه به آنچه که آورده شد، توالی DNA ماده خامی است که با توجه به مفروضات مختلف، می‌تواند نتایج ویژه‌ایی را برای پژوهشگر حاصل کند. معمولا هر وقت توالی DNA یا هر توالی زیستی دیگر مورد پژوهش قرار گیرد، بایستی یک توالی تصادفی اما هم اندازه با توالی DNA مورد نظر تولید کرد. برای هر ژن مورد بررسی در این پژوهش، مقدار آلفا (تابع توان) و مدل برازش داده شده – که تابع طول DNA و مقدار آلفا هست – بدست آمد. شکل ۸، توزیع توان واپاشی حاصل از برازش تابع قانون توان روی همبستگی‌های دامنه بلند محاسبه شده روی ژن‌ها را نشان می‌دهد. از لحاظ قانون توان واپاشی و میزان مشابهت مقادیر این معیار یا قانون، شکل ۸ نشان می‌دهد بعضی از ژن‌ها را می‌توان در یک گروه(ماژول) قرار داد. به عبارتی دیگر، حالت هندسه فراکتال – که خاصیت تابع قانون توان به نحوی محسوب می‌شود – را می‌توان برای بعضی از ژن‌ها دید. دامنه پایین بسامد حاصله نشان می‌دهد که این خاصیت می‌تواند در طول کوتاهی از DNA ژن‌ها اتفاق بیفتد (جدول ۲). نقش احتمالی همبستگی‌های دامنه بلند ژنی در شکل گیری ماژول‌های ژنی در اینجا کمی آشکار می‌شود. ماهیت طبیعت زیستی در سازواره‌ها، حاکی از وجود ماژول‌های مختلف زیستی در سطوح مختلف زیست سامانه‌ها است. علم زیست سامانه‌ها تاکید تام بر اهمیت بالاتر ماژول‌های ژنی نسبت نقش تکین ژن‌ها در ایجاد و درمان بیماری‌ها، فنوتیپ‌ها و غیره دارد. خصوصیات یک ماژول ممکن است شامل تعداد زیادی از اتصالات نسبت به تعاملات باشد. با این ویژگی‌ها، ماژول‌ها واجد شرایط ساخت بلوک‌هایی از یک زیست سامانه‌ها می‌شوند، بنابراین، ممکن است بتوان ساختار سامانه‌ایی و پویایی یک زیست سامانه را با پژوهش ماژول‌ها به صورت جداگانه انجام داد و سپس کنکاش کل یک زیست سامانه پیچیده را از طریق تعاملات ماژول‌ها مورد بررسی قرار داد. تجزیه و تحلیل‌ها در سطح ماژول‌های ژنی ممکن است در ایجاد دیدگاهی نو به رفتارهای زیستی با تجزیه و تحلیل ویژگی‌های پویایی ماژول‌های ژنی یا خلاصه کردن یک زیست سامانه به ماژول‌ها کمک شایانی بنماید.

توزیع توان یک حالت خاصی از توزیع‌های احتمال است و به چندین روش می‌توان آن‌ها را نشان داد. گرچه

[1] Principal component analysis
[2] Singular value decomposition
[3] Method dynamical language
[4] Markov model method
[5] Fractal method

مقایسه بین ژن‌ها از نظر همبستگی‌های دامنه بلند محاسبه شده سخت است – که یک علت مهم آن به تفاوت میزان GC ژن و طول ژن‌ها برمی‌گردد– اما در کل می‌توان بین ژن‌های مورد بررسی در این پژوهش، دامنه از همبستگی دامنه بلند وجود دارد. اگرچه در پژوهش‌های اصلاح نژادی سابقه ندارد اما اخیرا محاسبه همبستگی‌های دامنه بلند روی توالی DNA بحث برانگیز شده است. احتمالا برای اولین بار ساختار همبستگی در توالی‌های اولیه DNA توسط لی و همکاران(۱۹۹۷) از جنبه‌های مختلف مانند تقارن توابع همبستگی ۱۶ نوکلئوتیدی، برآورد دقیق معیارهای همبستگی، رابطه بین 1/f و طیف لورنتسی، ناهمگونی در توالی‌های DNA، راهبردهای مدل سازی متفاوت ساختار همبستگی توالی‌های DNA، تفاوت ساختار همبستگی بین مناطق کد کننده و غیر کد کننده DNA بررسی گردید (لی ۱۹۹۷). آن‌ها نشان دادند که اگر چه بعضی از نتایج بحث برانگیز است، اما کار بر روی این موضوع نقطه شروع خوبی برای پژوهش‌های آینده است (لی ۱۹۹۷). در پژوهش دیگری روشی برای کنکاش خواص تصادفی توالی نوکلئوتیدی DNA، که امکان در نظرگرفتن همبستگی‌های دامنه کوتاه را اجازه می‌داد، توسعه یافت. در پژوهش یاد شده، خواص تصادفی نوکلئوتیدی DNA با نگاشت ۱:۱ این توالی روی گام، مفهوم گام DNA[1] پیشنهاد گردید. این روش امکان سنجش کمی همبستگی‌های دامنه بلند قانون توان که خاصیت ناوردای مقیاس DNA[2] جدیدی را به نمایش می‌گذاشت، فراهم گردید. با استفاده از روش یاد شده، نوع همبستگی دامنه بلندی در ژن‌های در بردارنده اینترون و توالی‌های غیر ترجمه‌ای تنظیمی DNA استخراج شد اما اینگونه همبستگی در توالی‌های مکمل DNA و ژن‌هایی که در بردارنده تعداد کمی اینترون بودند، استخراج نشدند (پنگ و همکاران ۱۹۹۲). نشان داده شده است که ترکیبی از مقیاس‌های مختلف DNA برای مولفه طیفی 1/f مشاهده شده در DNA دخیل هستند. یک راه برای بررسی همبستگی دامنه بلند در DNA با استفاده از پیچیدگی توالی‌های DNA، مدله کردن فرایندهای بازآرایی و مضاعف سازی ملکول DNA است. نشان داده شده است که یک فرایند مضاعف‌سازی DNA، نقش مهمی در مشاهده همبستگی‌های دامنه بلند روی پلمیر DNA دارد (لی و همکاران ۱۹۹٤). همچنین، نشان داده شده است که وجود همبستگی‌های دامنه بلند روی آماره نمره‌های همردیف سازی DNA اثر می‌گذارد (مسر و همکاران ۲۰۰۷). برای حل این معضل، روشی بر اساس توزیع مقدار حدی[3] ارایه شد تا نمره همردیفی تصحیح گردد. همچنین، خواص همبستگی‌های دامنه بلند توالی‌های DNA با استفاده از تجزیه واریانس توزیع چگال تکین یا گروهی از نوکلئوتیدها مورد بررسی قرار گرفته است (مهنتی و رائو ۲۰۰۲) و عدم تقارن در محتویات نوکلئوتیدی و یا ساختار طرح‌دار[4] به عنوان منشاء اصلی همبستگی‌های طولانی مدت نیز مورد بررسی قرار گرفته است. یافته‌های علمی روی همبستگی‌های دامنه بلند و فراکتال‌های یافت شده روی ژن‌های دربردارنده اینترون سازواره‌های زنده به صورت کمی و کیفی مورد بررسی قرار گرفته است. نشان داده شده است که این یافته‌ها با تغییرات ترکیب بازها در نواحی مختلف DNA تشابه و هم‌وردایی ناچیزی دارند (چتزی دیمیتریو و همکاران ۱۹۹٤). پژوهش‌ها حاکی از آن است که همبستگی‌های قانون توان مقیاس کوچک، مسئول اصلی پیچش DNA حول پروتئین‌های هیستونی و تشکیل ساختار نوکلئوزوم‌ها است(آیودیت و همکاران ۲۰۰۱). در این راستا نشان داده شده است که شکل گرفتن ساختارهای با مرتبه بالاتر و دینامک کروماتین‌ها، نشان دهنده وجود همبستگی‌های دامنه بلند در این ساختارهاست (آیودیت و همکاران ۲۰۰۱). اخیرا در پژوهشی به نقش با اهمیت

[1] DNA Walk
[2] Scale-Invariant Property of DNA
[3] Extreme Value Distribution
[4] Patchy Structure

همبستگی‌های دامنه بلند در مکانیک حلقه‌های DNA توجهی خاصی شده که می‌تواند در تنظیم بیان ژن نیز استفاده شود (سوتهیباتپونگ و همکاران ۲۰۱۶). در یک پژوهش که خواص فرکتال توالی‌های نواحی کد دهنده و غیرکد دهنده DNA انسان را با استفاده از تبدیلات ویولت بررسی کرد نشان داده شد که همبستگی‌های دامنه بلند در نواحی غیر کد کننده اینترون با محتوای گوانین – سیتوزین افزایش می‌یابد در صورتیکه این چنین همبستگی درهیچ درجه‌ایی از محتوای گوانین – سیتوزین بدست نیامد (سوتهیباتپونگ و همکاران ۲۰۱۶). در ادامه نشان داده شد که وجود همبستگی‌های دامنه بلند مشاهده شده احتمالا ارتباطی به سازوکارهای حذف – درج نوکلئوتید ندارد. همبستگی‌های دامنه بلند قانون توان در سامانه‌های گوناگونی کشف شده است (پنگ و همکاران ۱۹۹۵). از نظر فیزیکی؛ وجود همبستگی‌های دامنه بلند، یک واقعیت فیزیکی هستند که به نوبه خود به صورت هندسه فراکتال طبیعت گروه‌بندی و شناخته می‌شوند. وجود هبستگی دامنه بلند قانون توان در یک سامانه، می‌تواند درک بهتری از طبیعت آن سامانه را آشکار کند چرا که به محض اینکه این نوع همبستگی‌ها استخراج شوند، ما می‌توانیم آن سامانه را با توان بحرانی[1] بررسی کنیم (پنگ و همکاران ۱۹۹۵). بررسی این نوع هبستگی‌ها برای سامانه‌هایی که ظاهرا نامربوط هستند، امکان استخراج شباهت‌های سامانه‌های مختلف را نشان می‌دهد که به نوبه خود امکان اتحادسازهایی یا ماژول سازی‌هایی که در غیر اینصورت نادیده خواهد ماند را آشکار خواهد کرد. در پژوهشی که روابط درختچه‌های حیات بین ژن‌های گونه‌های مختلف را با استفاده از الگوهای همبستگی‌های دامنه بلند تحلیل کردند، نشان داده شد ژن‌هایی که ارتباط تکاملی مشابه دارند و

با هم در ارتباط هستند از نظر الگوهای همبستگی‌های دامنه بلند DNA نیز شبیه هم هستند. بنابراین، استخراج الگوهای همبستگی‌های دامنه بلند، می‌تواند درتحلیل‌های تکاملی و گونه‌زایی و درختچه‌های حیات استفاده شود (ناگار و سوخی ۲۰۰۸). دریک پژوهش فیزیک–ژنتیکی، که ارتباط بین همبستگی‌های دامنه بلند در DNA با انتقال بار الکتریکی در توالی‌های جانشینی DNA با استفاده از رهیافت ماتریس انتقالی[2] بررسی گردید. نشان داده شد که توالی‌های جانشینی DNA همبستگی‌های دامنه بلند را نشان می‌دهند و قدرت انتقال الکترون بهتری نسبت به توالی‌های تصادفی ناهمبسته را دارند (گائو ۲۰۰۷). از پژوهش‌های بالا می‌توان چنین نتیجه گرفت که استخراج همبستگی‌های دامنه بلند DNA، می‌تواند به سوال‌های زیستی زیادی پاسخ بدهد. برای ژن‌های مورد بررسی در این پژوهش، یک استدلال اصلاح نژادی را می‌توان بنا کرد و آن اینکه با توجه به پراکنش همبستگی‌های دامنه بلند برای هر ژن، آن نقطه‌ایی از توالی ژن که بیشترین میزان همبستگی را نشان می‌دهد، احتمالا درجه بالایی از تعادل عدم لینکاژ بین بازهای DNA را نشان می‌دهد. این مورد نیاز است که در پژوهش‌های دیگر بررسی شود. از این نظر که همبستگی‌های دامنه بلند DNA ورودی بسیاری از الگوریتم‌های تبدیلات فوریه است، نقش احتمالی تبدیلات فوریه گسسته نیز باید در متن داده‌های ژنی اصلاح نژادی بررسی شود. نتایج نشان داد که امکان استخراج ماژوهای ژنی در سطح DNA وجود دارد که این ماژول‌های ژنی بعدها می‌تواند در اصلاح نژاد به کار گرفته شود. با توجه به گستره کاربرد همبستگی‌های دامنه بلند DNA و سادگی انجام آن، می‌توان در تحلیل‌های تکاملی اصلاح نژادی نیز از نتایج این محاسبات استفاده کرد. نگاهی روی آزمون‌های ابداع شده برای بررسی فرض‌های خاص (عمدتا تکاملی) روی DNA نشان می‌دهد که روش‌های مختلف زیادی در این زمینه توسعه یافته است. این پژوهش‌ها امروزه پای خود را در محاسبات اصلاح نژادی نیز باز کرده است. در یکی از آخرین بررسی‌ها تعداد ۳۰۰ ژن کد کننده پروتئینی مورد انتخاب مثبت در اهلی سازی خوک *Meishan*

[1] Critical Exponent

[2] Transfer Matrix Approach

قرارگرفتند (ژائو و همکاران، ۲۰۱۸). در پژوهش جدید دیگری، با استفاده از روش‌های FST و هموزیگوسیتی‌هاپلوتیپ گسترش بسط داده شده (*EHH-Rsb*)، ژن‌های مرتبط با رشد / قد (*DOCK7*، *PLCB4*، *HS2ST1*، *FBP2* و *TG*)، کیفیت لاشه و گوشت (*TG*، *COL14A1*، *FBXO5*، *NR3C1*، *SNX7*، *ARHGAP26* و *DPYD*)، تعداد پستانک (*LOC100153159* و *LRRC1*)، رنگدانه (*MME*) و ریخت‌شناسی گوش (*SOX5*) در خوک‌ها بررسی شدند. این نتایج مبنایی برای بررسی جهش‌های مرتبط با تنوع فنوتیپ‌های مشاهده شده قرارگرفتند و چنین استنتاج شد که تحلیل پویش ژنومی، استفاده از برخی از این نشانگرها را در برنامه‌های ارزیابی ژنتیکی تسهیل می‌نماید.(ادی و همکاران ۲۰۱۷). به نظر می‌رسد که استخراج همبستگی‌های دامنه بلند DNA می‌تواند در پژوهش‌های تکاملی نیز به کار گرفته شود. شکی نیست که در این میان استفاده از شبیه سازی کننده‌های توالی DNA مثل INDELible (فلچر و یانگ ۲۰۰۹)می‌تواند کارگشا باشد.

نتیجه گیری

در این پژوهش، با استخراج همبستگی‌های دامنه بلند بر روی فهرستی از ژن‌های کاندیدای موثر بر تولید شیر در گاو، نشان داده شد که ژن‌های مختلف درجه خاصی از همبستگی دامنه‌بلند را نشان می‌دهند. این همبستگی‌ها می‌تواند، روی قدرت ماژول شدن با هم و همردیف سازی ژن‌ها اثر بگذارد. از طرفی، نشان داده شد که از نظر پیچیدگی آماری، ژن‌های مورد نظر، درجه بالایی از پیچیدگی دارند. نتایج نشان داد که استخراج همبستگی‌های دامنه بلند متکی به محتوی GC است. بنابراین، رابطه کل نوکلئوتیدها با هم در نظر گرفته نشده است. از طرفی ارتباط بین ژن‌ها از نظر میزان همبستگی دامنه بلند نیز بررسی نشد. لذا پیشنهاد می‌شود این ارتباط بین ژن‌ها از نظر همبستگی‌های دامنه بلند بین ژنی[۲] نیز بررسی شود. میزان بسامد حاصل از همبستگی‌های دامنه بلند در ژن‌ها متفاوت اما نزدیک به هم بود. پیشنهاد می‌شود این نواحی از نظر وجود عدم تعادل پیوستگی مورد کنکاش بیشتری قرار گیرند. پژوهش یاد شده اولین پژوهش در زمینه استفاده از همبستگی‌های دامنه بلند در حوزه ژنوم حیوانات اهلی بود. توصیه می‌شود در پژوهش‌های اصلاح نژادی، ابتدا ژن یا ژنوم مورد نظر با توجه به هزینه بر نبودن روش مورد بررسی در این مقاله و پیش آگهی‌دهنده بودن این روش از لحاظ آشکار کردن فرسته‌های ژنومی مورد استفاده قرار بگیرد و بودن یا نبود ماهیت فراکتال در آن‌ها بررسی شود. سپس روال ارزیابی معمول اصلاح نژادی متکی به SNP روی ژنوم مورد نظر انجام پذیرد تا میزان همپوشانی آن‌ها بدست اید.

تشکر و سپاسگزاری

از آقای دکتر هوشنگ دهقانزاده که در بالا بردن دقت استخراج ویژگی‌های ژن‌های مورد بررسی، کمک شایانی کردند، سپاسگزاری می‌گردد.

منابع مورد استفاده

Arneodo A, d'Aubenton-Carafa Y, Audit B, Bacry E, Muzy J and Thermes C, 1998. Nucleotide composition effects on the long-range correlations in human genes. The European Physical Journal B-Condensed Matter and Complex Systems 1: 259-263.

Audit B, Thermes C, Vaillant C, Aubenton-Carafa Y, Muzy JF and Arneodo A, 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. Physical Review Letters. 86:2471-2479.

Bernaola-Galván P, Carpena P, Román-Roldán R and Oliver J, 2002. Study of statistical correlations in DNA sequences. Gene 300: 105-115.

[1] Extended haplotype homozygosity

[2] Gene long-range cross-correlation

Blaisdell BE, 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences 83: 5155-5159.

Chatzidimitriou-Dreismann CA, Streffer R and Larhammar D, 1994. A quantitative test of long-range correlations and compositional fluctuations in DNA sequences. The FEBS Journal 224: 365-371.

Comin M and Verzotto D, 2012. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms for Molecular Biology 7: 34-41.

Edea Z, Hong JK, Jung JH, Kim DW, Kim YM, Kim ES, Shin SS, Jung YC, Kim KS, 2017. Detecting selection signatures between Duroc and Duroc synthetic pig populations using high-density SNP chip. Animal Genetics. 48(4):473-477.

Edgar RC and Batzoglou S, 2006. Multiple sequence alignment. current opinion in structural biology. Current Opinion in Structural Biology 16: 368-373.

Edwards SV, Fertil B, Giron A and Deschavanne PJ, 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Systematic Biology 51: 599-613.

Fletcher W, Yang Z, 2009. INDELible: a flexible simulator of biological sequence evolution. Molecular Biology and Evolution 26(8): 1879-88.

Guo AM, 2007. Long-range correlation and charge transfer efficiency in substitutional sequences of DNA molecules. Physical Review E 75:061915.

Jun SR, Sims GE, Wu GA and Kim SH, 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences 107: 133-138.

Katoh K, Misawa K, Kuma Ki and Miyata T, 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066.

Kemena C and Notredame C, 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25: 2455-2465.

Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A,Thompson RL, Gibson TJ and Higgins DG, 2007. Clustal W and clustal X version 2.0. Bioinformatics 23: 2947-2948.

Lemay DG, Lynn DJ, Martin WF, Neville MC and Casey TM, 2009. The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biology 10: R43.

Li W, 1997. The study of correlation structures of DNA sequences: a critical review. Computers & Chemistry 21: 257-271.

Li ,W., Marr TG and Kaneko K, 1994. Understanding long-range correlations in DNA sequences. Physica D: Nonlinear Phenomena 75: 392-416.

Messer PW, Bundschuh R, Vingron M and Arndt PF, 2007. Effects of long-range correlations in DNA on sequence alignment score statistics. Journal of Computational Biology 14: 655-668.

Mohanty A, and Rao A, 2002. Long range correlations in DNA sequences. arXiv preprint physics/0202075.

Nagar AK and Sokhi D, 2008. Phylogenetic comparison of genes using long range correlation patterns in DNA sequences. Proc. Computer Modeling and Simulation 2: 197-202:

Peng CK, Buldyrev S, Goldberger A, Havlin S and Mantegna R, 1995. Statistical properties of DNA sequences. Physica A: Statistical Mechanics and its Applications 221: 180-192.

Peng CK Buldyrev SV, Goldberger AL, Havlin S and Sciortino F, 1992. Long-range correlations in nucleotide sequences. Nature 356: 168-170.

Sims GE, Jun SR, Wu GA and Kim SH, 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences 106: 2677-2682.

Stuart GW, Moffett K and Baker S, 2002. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18: 100-108.

Stuart GW, Moffett K and Leader JJ, 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution 19: 554-562.

Sutthibutpong T, Matek C, Benham C, Slade GG and Noy A, 2016. Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. Nucleic Acids Research 44: 9121-9130.

Vinga S, 2013. Information theory applications for biological sequence analysis. Briefings in Bioinformatics 15: 376-389.

Voss RF, 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters 68(25): 3805-3808.

Warnow T, 2013. Large-scale multiple sequence alignment and phylogeny estimation in models and algorithms for genome evolution. pp. 85-146, Springer Londin.

Yu ZG, Anh V and Lau KS, 2003. Multifractal and correlation analyses of protein sequences from complete genomes. Physical Review E 68: 021913.

Yu ZG, Anh V and Zhou LQ, 2005. Fractal and dynamical language methods to construct phylogenetic tree based on protein sequences from complete genomes. In International Conference on Natural Computation (pp. 337-347). Springer, Berlin, Heidelberg.

Yu ZG, Zhou LQ, Anh VV, Chu KH, Long SC and Deng JQ, 2005. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. Journal of Molecular Evolution 60: 538-545.

Zhao P, Yu Y, Feng W, Du H, Yu J, Kang H, Zheng X, Wang Z, Liu G, Ernst CW, Ran X, Wang Jand Liu J, 2018. Evidence of evolutionary history and selective sweeps in the genome of Meishan pig reveals its genetic and phenotypic characterization. Gigascience 7(5): 1-12

Study of long-range DNA correlations in milk yield affecting genes of dairy cow

Roxana Abadeh[1], Mehdi Aminafshar[2], Mostafa Ghaderi-Zefrehei[3*], Seyyed Abbas Mohammadi[4] and Mohammad Chamani[5]

Received: November 22, 2018 Accepted: September 29, 2020

[1]PhD Student, Department of Animal Science, Faculty of Agricultural Sciences and Food Industry, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Assistant Professor, Department of Animal Science, Faculty of Agricultural Sciences and Food Industry, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3]Associate Professor, Department of Animal Science, Faculty of Agricultural Sciences, Yasouj University, Yasouj, Iran
[4]Associate Professor, Department of Mathematical, Faculty of Basic Sciences, Yasouj University, Yasouj, Iran
[5]Professor, Department of Animal Science, Faculty of Agricultural Sciences and Food Industry, Science and Research Branch, Islamic Azad University, Tehran, Iran
*Corresponding Author`s Email: mghaderi@yu.ac.ir

پژوهش‌های علوم دامی Animal Science Research

Journal of Animal Science/vol.31 No.2/ 2021/pp 29-43
https://animalscience.tabrizu.ac.ir

OPEN ACCESS

DOI: 10.22034/AS.2021.30510.1464

Introduction: For mathematically oriented investigators, DNA is a string. A DNA sequence is considered as a string of symbols and correlation of its structure can almost completely be characterized by base-base correlation functions at any range, short, long and/or their corresponding power spectra. Long-range correlations between bases in the DNA sequence are a statistical feature found in the genome of many eukaryotes. The existence of long-range DNA correlations indicates the existence of DNA rearrangement or duplication processes. The phenomena is not directly applicable to breeding and is mostly used in evolutionary studies. Our basic assumption in this study was that by extracting long-range DNA correlations between all the different nucleotides within a gene, it is possible to achieve a degree of correlation between them in the first place and possibly better run SNP-based research. Due to various issues, not all investigations of a complete characterization of long-scale correlation structure of DNA sequences were motivated by biology arena. Rather, many such investigations were motivated by the issues of mathematical modeling, cryptography language code detections, dynamical systems, stochastic processes, and noise detections. Perhaps due to this reason, long-scale correlation structure has not yet become part of the toolbox in the "mainstream" DNA sequence analysis in human genetics and breeding settings. Prediction of DNA correlations from a sequence with finite length could be done with frequency-count estimator, indirect and direct Bayesian estimator. In this study, we followed CorGen theory.
Material and methods: In this study, 24 dairy cow milk yield affecting genes were investigated. The number, length and length of each exon as well as its position on the chromosome were obtained from the NCBI gene bank and the sequences were saved in FASTA format. According to the research request, the accession numbers of the studied genes were plugged in a previously designed program (by #C language) and the appropriate output was obtained. CorGen software was used to calculate the long-range DNA correlations of the genes involved in milk production. The objectives of this study were:1) there has been discordant on the result of correlation structure in DNA sequences. Due to this matter of what the actual result is, some researches still believe that DNA sequences do not

exhibit any feature long-range DNA correlation which cannot be explained by the basic known stochastic processes such as random sequence or Markov chain. Resolving this disagreement can be straightforward once everybody agrees to use the same measure of correlation, estimator, and apply this estimator of the correlation to the same sequence, 2) to highlight more biologically-motivated study of correlation structure of long range DNA sequences especially in animal breeding setting.
Results and discussion: The results showed that there is a significant level of long-term correlation in DNA sequence of several genes such as *EZR, FGG, KRT6A, RAB1A, EIF3L, TBC1D20, ZNF419, S100A16, MRPL3, TPPP3, PHF10*. The reduction power of the fitting function of the power function was based on the long-range correlations obtained from genes of different lengths, in the range of 0.146 and 0.643. Hence it can be concluded that reducing the range of long-range correlations by increasing the interval between DNA sequence intervals does not follow a random process. Accordingly, the fractal geometry of nature was observed in these genes. In this study, we attempted to address long-DNA correlation in dairy cattle genes. Although this research does not accomplish this task, the intention was to at least put forward the issue. Most of the current studies of correlation (especially the long range one) in DNA sequences are based-base base statistical correlations. This base-base correlation won't be a powerful tool to reveal the correlation on a global scale or between larger blocks in DNA-sequences. The genes studied have been shown to have high complexity and mode of invariant on their DNA. This type of analysis can be generalized to the work of breeding setting. A more complete characterization of long-range correlation between base pairs at both short and long distances became possible only as long DNA sequences became more commonly available. Nowadays, due to significant growth of DNA generating technologies, almost the whole genome of an organism can cost- and time-effectively be sequenced. Therefore, a raw data shall be available to researchers interested in checking new DNA correlation hypotheses in handy DNA sequences. The claim of DNA base-base statistical correlation at long distances in DNA sequences is sought to be still a few steps away from finding a simple organization principle of the genome.

Keywords: Decay exponent, Long range correlation, Dairy cow, Fractal geometry