دانشکده ریاضی و علوم کامپیوتر

عنوان : گزارشی بر مقاله ی Ficket

نگارش : علیرضا محمدزاده

استاد راهنما: دکتر فاطمه زارع

مقدمه

سوال زیستی

تقاضای یک روش قابل اعتماد برای تفکیک میان قطعات کدینگ (در تمام این مقاله و گزارش منظور از کدینگ، coding for protein هست.)وغیر کدینگ DNA یک نیازمندی ثابت در بیوانفورماتیک بوده است.

چالش های آزمایشگاهیی

گاهی اوقات محل آغاز ژن تنها به صورت حدودی مشخص است و این باعث می شود که سکانس چندین کاندید برای ORF داشته باشد.حتی بعد از اینکه محل دقیق ژن مشخص شد،سکانس های اطراف ممکن است شامل ORF های دیگر باشند به طوری که function این ORF ها نامشخص هست.وجود یک روش شناسایی مناطق کدینگ از غیر کدینگ یک ابزار بسیار قدرتمند برای کشف پروتئین ها محسوب می شود.

سوال محاسباتی

ساخت یک روش تصمیم گیری برای یک توالی DNA که به آن عنوان کدینگ یا غیر کدینگ را بدهد.

روش های حل مسئله

دو روش به طور کلی برای حل این مسئله وجود دارند:

۱ جست و جو برای transcription or translation initiation signals

۲ روش های آماری (توجه به ویژگی های آماری سکانس ها)

چالش های هر روش

روش ۱)جست و جو برای transcription or translation initiation signals:

مشکل۱)سیگنال های شروع ممکن است،این اتفاق زمانی می افتد که انتهای 5’ ، ORF که در حال مطالعه

ی آن هستیم،در توالی وجود ندارد.گاهی اوقات نیز PCS که به دنبال آن هستیم،هیچ سیگنال آغازی ندارد.به عنوان مثال ژن لیزیس phage MS2 تنها در صورت عبور از کدون توقف قبلی ترجمه می شود.
مشکل۲) توصیف دقیق اینکه چه چیزی یک سیگنال شروع است،هنوز مشکل به نظر می رسد.

روش ۲) روش های آماری (توجه به ویژگی های آماری سکانس ها):
برای بررسی چالش ها و مشکلات روش های آماری به بیان برخی از مقالات ارائه شده در این حوزه ها می پردازیم و آن ها را نقد می کنیم.
۱) شوالمن و همکارانش الگوهایی را برای مناطق کدینگ دو فاژ پیدا کردند که به یک کد سه حرفی اشاره داشت اما نمونه ای که آنها به کار بردند بسیار کوچک بود و آنها پیش بینی خودشان را برای نمونه های دیگر بررسی نکردند.
۲) شپرد برای پیدا کردن منشا کد ژن دوره های تناوبی را در توالی ها پیدا کرد که شامل باز های منفرد و جفت باز ها در DNA بود و از این موضوع برای پیدا کردن مناطق کدینگ کمک گرفت.با اینکه الگو های جالبی پیدا شد اما هیچ تستی برای کدینگ/غیر کدینگ ارائه نشد. همین طور هیچ مدرکی که این موضوع را اثبات کند که مناطق غیر کدینگ چنین الگو هایی را ندارند ارائه نشد.
۳) استادن و مک لاکلن یک برنامه ی کامپیوتری برای مپ کردن مناطق کدینگ یک توالی ارائه کردند.این برنامه به این شکل عمل می کرد که ابتدا میزان شباهت codon usage یک PCS شناخته شده را با ORF که در حال مطالعه هستیم محاسبه می کرد و بر اساس این شباهت قضاوت را انجام می داد.این متود نیاز دارد که PCS شناخته شده از نظر codon usage به PCS، ORF که در حال مطالعه ی آن هستیم،نزدیک باشد.این موضوع طرح را بسیار وابسته به قضاوت یوزر می کند و این باعث می شود که این تست غیر کاربردی شود.

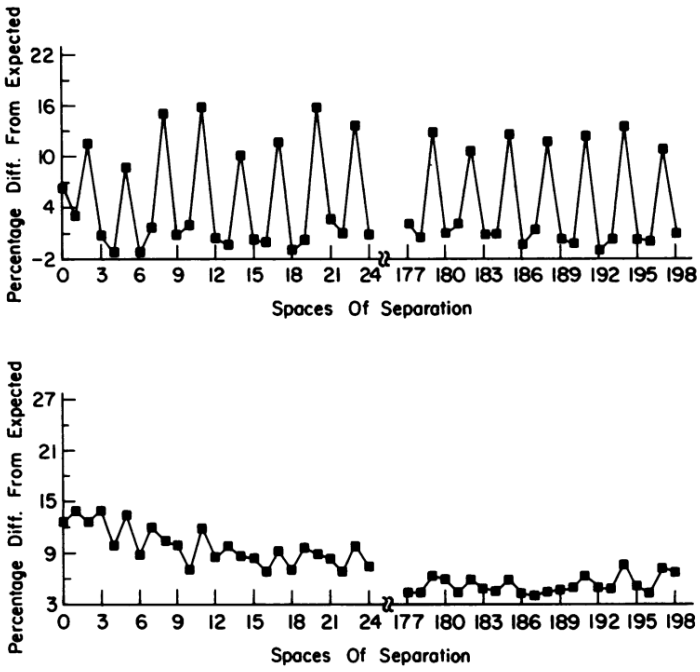
مقاله جاری چه چالشی را حل کرده

مقاله هایی که بررسی کردیم همگی بر اساس روش آماری بودند که برای مشخص کردن مناطق کدینگ و غیر کدینگ استفاده می شوند.تمام این الگوها دارای پتانسیل تست کدینگ/غیر کدینگ را دارند اما ما معتقدیم که مقاله ی ما اولین مقاله ای است که یک تست کاملا مشخص و عینی بوده که روی تعداد زیادی توالی بررسی شده است تا قابل اطمینان باشد.

ما همچنین تلاش کردیم که تستی را پیدا کنیم که یک جواب ساده ی کدینگ یا غیرکدینگ به یک منطقه ی به خصوص بدهد و نه اینکه کل توالی را به صورت کدینگ یا غیر کدینگ مپ کنیم.این کار انجام آزمون قابل اعتماد برای مقیاس بزرگ را تسهیل می کند.

متود

زیربنای تمام مشاهدات مربوط به ترتیب آماری در PCS این است که کدون ها با فرکانس های متفاوت استفاده می شوند.یکی از پیامد های این موضوع این است که نوکلوتید ها تمایل دارند که در تناوب های ۳ تایی در PCS تکرار شوند.نمودار شماره ۱میزان همبستگی تیمین را در بخش های کدینگ و غیر کدینگ نشان می دهد.این شکل این موضوع را نمایش می دهد که تعداد باز های تیمین در جایگاه های 2,5,8,11,... یعنی (3n+2) بیشتر از (3n) و (3n+1) هست.اما نمودار شماره ۲ نشان می دهد که چنین موضوعی برای بخش های غیر کدینگ وجود ندارد.



شکل شماره 1

نمودار های همبستگی بالا برای تیمین(T) ترسیم شده اند.این دیتا از ۳۲۱ قطعه(fragment) کدینگ و همین طور ۲۴۹ قطعه ی (هر قطعه طول ۲۰۰ باز دارد)غیر کدینگ کتابخانه توالی Los Alamos بدست آمده

است.به ازای هرفاصله ی ممکنk تعداد باز های تیمین را که با فاصله ی k در توالی های کدینگ ظاهر شده اند را شمارش کردیم و آن را با تعداد مورد انتظار مقایسه کردیم.منظور از تعداد مورد انتظار یعنی مدلی که باز ها به صورت مستقل انتخاب شده اند.درصد تفاوت در بین kهای بین ۰ تا ۲۴ و همین طورk های بین ۱۷۷ و ۱۹۸ نشان داده شده است.در نمودار پایین نیز همین موضوع برای ناحیه های غیر کدینگ تکرار شده است.همانطور که مشاهده می کنید این موجی که برای که برای بخش های کدینگ وجود دارد ، برای بخش های غیر کدینگ وجود ندارد.این موضوع به طور تقریبی برای باز های دیگر نیز وجود دارد.

حال به سراغ تعریف ۸ پارامتر عددی می رویم که این پارامتر ها مناطق کدینگ را از غیر کدینگ متمایز می کنند.چهار پارامتر اول با انگیزه ی تفاوت توزیع هر باز در بین ۳ جایگاه کدون شکل گرفتند.

A_1, A_2, A_3 به شکل زیر تعریف می شوند.

A_1 = Number of A's in positions 1,4,7,10,...

A_2 = Number of A's in positions 2,5,8,11,...

A_3 = Number of A's in positions 3,6,9,12,...

پارامتر A-Position را به شکل زیر تعریف می کنیم.

$$\text{A-Position} = \frac{\text{MAX}(A_1, A_2, A_3)}{\text{MIN}(A_1, A_2, A_3)+1}$$

این پارامتر را برای باز های دیگر نیز تعریف می کنیم.(C,G,T)

A-C-G-T-Position میزان درجه یی که هر باز علاقه دارد که در کدام پوزیشن بیشتر علاقه دارد، کدون قرار بگیرد را مشخص می کند.۴ پارامتر بعدی که تعریف می شوند،A-C-G-T-Content هستند که در واقع این پارامتر فرکانس هر کدام از این باز هارا مشخص می کند.

$$X^S - Content = \frac{R_X^S}{n},\ X \in \{A, C, G, T\}.$$

توزیع نسبی این ۸ پارامتر در بین مناطق کدینگ و غیر کدینگ در جدول ۱ مشخص شده است.این ۸ پارامتر در مرحله های بعدی برای مشخص کردن مناطق کدینگ و غیر کدینگ استفاده شده اند اما توجه داشته باشیم که به هر کدام از این پارامتر ها اگر به صورت تنها نیز توجه کنیم باز هم تفاوت آنها را میان کدینگ و غیر

کدینگ ها می توانیم متوجه شویم.به عنوان مثال برای قطعه هایی که T-Position آنها کمتر از 1.2 هست،تنها احتمال 9% وجود دارد که آن قطعه کدینگ باشد.در حالی که برای قطعه هایی که T-Position آنها بیشتر از 1.7 هست احتمال کدینگ بیش از 90% هست.

Position Parameter			Probability of Coding			
0.0	to	1.1	A: .22	C: .23	G: .08	T: .09
1.1		1.2	.20	.30	.08	.09
1.2		1.3	.34	.33	.16	.20
1.3		1.4	.45	.51	.27	.54
1.4		1.5	.68	.48	.48	.44
1.5		1.6	.58	.66	.53	.69
1.6		1.7	.93	.81	.64	.68
1.7		1.8	.84	.70	.74	.91
1.8		1.9	.68	.70	.88	.97
1.9		2.0+	.94	.80	.90	.97

Content Parameter			Probability of Coding			
.00	to	.17	A: .21	C: .31	G: .29	T: .58
.17		.19	.81	.39	.33	.51
.19		.21	.65	.44	.41	.69
.21		.23	.67	.43	.41	.56
.23		.25	.49	.59	.73	.75
.25		.27	.62	.59	.64	.55
.27		.29	.55	.64	.64	.40
.29		.31	.44	.51	.47	.39
.31		.33	.49	.64	.54	.24
.33		.99	.28	.82	.40	.28

جدول۱

اطلاعات جدول بالا بر اساس توالی های دیتابیسی که داریم بدست آمده است.محدوده ی هر پارامتر به ده بازه تقسیم شده است.برای هر بازه احتمال کدینگ محاسبه شده است و عدد آن در جدول قرار گرفته است.این یعنی برای هر بازه آن قطعه هایی که پارامترشان در آن بازه قرار می گرفت احتمال کدینگ محاسبه شده است.(تعداد کدینگ ها تقسیم بر کل قطعه ها)

چگونه کدینگ را از غیر کدینگ تشخیص دهیم

در این قسمت قصد داریم که بر هر کدام از ۸ پارامتری که در قسمت قبل محاسبه کردیم،وزن بدهیم.از جدول شماره ۱ می توانیم این نتیجه را بگیریم که اطلاعات T-Position می تواند اطلاعات بیشتری نسبت به A-Content به ما بدهد.برای اینکه بتوانیم عددی را پیدا کنیم که برای ما مشخص کند که هر پارامتر چه سهمی باید در تست ما داشته باشد، این فرآیند را طی می کنیم. هر بازه که با احتمال بیشتر از 50% (احتمال

رندوم)کدینگ را به درستی پیش بینی کند،را در نظر می گیریم و احتمال آن را منهای 0.5 می کنیم و میانگین گیری می کنیم. به این شکل اطلاعات جدول ۲ بدست می آید.

جدول 1

	Position	Content
A	.26	.11
C	.18	.12
G	.31	.15
T	.33	.14

حال بعد از اینکه وزن هر پارامتر را مشخص کردیم،می توانیم TESTCODE را معرفی کنیم.TESTCODE معیاری برای مشخص کردن این موضوع است که قطعه ی DNA کدینگ است و یا غیر کدینگ.به ازای هر قطعه ای A_i, C_i, G_i, T_i آن را به ازای $i = \{1,2,3\}$ محاسبه می کنیم.بعد از آن ۸ پارامتر A-C-G-T-Position و A-C-G-T-Content را محاسبه می کینم و اعداد متناظر آنها را در جدول ۱ پیدا می کنیم و مقدار آنها را بر می داریم.$(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$بعد از آن وزن مربوط به هر پارامتر را از جدول ۲ برداشته $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)$ و TESTCODE را به این شکل محاسبه می کنیم.

$$TESTCODE = \sum_{i=1}^{8}(w_i p_i)$$

در جدول ۳ اطلاعات مربوط به TESTCODE و بازه بندی شده و در هر بازه احتمال کد شدن مشخص شده و همین طور پیش بینی TESTCODEمشخص شده است.

TESTCODE Indicator	Probability of Coding	Prediction
0.32 to 0.43	0.00	Noncoding
0.43 0.53	0.04	Noncoding
0.53 0.64	0.07	Noncoding
0.64 0.74	0.29	Noncoding
0.74 0.84	0.40	No Opinion
0.84 0.95	0.77	No Opinion
0.95 1.05	0.92	Coding
1.05 1.16	0.98	Coding
1.16 1.26	1.00	Coding
1.26 1.37	1.00	Coding

جدول۳

دیتابیس

تمام مطالعات گزارش شده در اینجا بر اساس داده های توالی ذخیره شده در کتابخانه توالی Los Alamos بوده که یک بانک اطلاعات عمومی در رایانه‌های CDC 7600 در آزمایشگاه ملی Los Alamos که در حال حاضر 486000 باز را در 320 توالی فهرست می کند.

هر توالی در کتابخانه به مناطق کدینگ و غیر کدینگ تقسیم شده است که این بخش بندی بر اساس شواهد تجربی گزارش شده انجام شده است.برای داده های اولیه ما 321 قطعه ی کدینگ شامل 230877 باز و 249 قطعه ی غیر کدینگ شامل 158987 باز را در نظر گرفتیم که هر کدام از قطعه ها حداقل طول ۲۰۰ را دارند.

برای پیاده سازی از توالی ژنوم یک قارچ به نام "Tilletia controversa" استفاده کردم که اگزون و اینترون آن مشخص شده است.

لینک دیتابیس: https://www.ncbi.nlm.nih.gov/nuccore/CAJHJB010000001.1

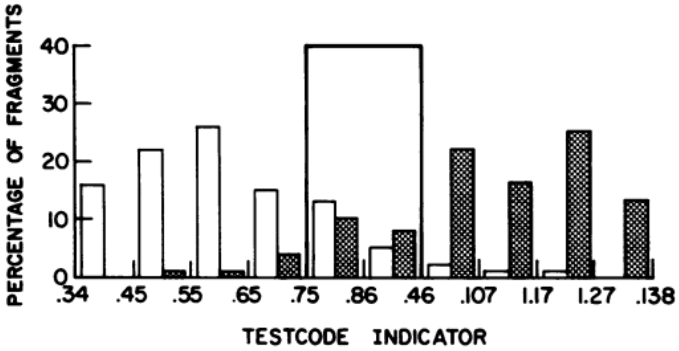
صحت

از جدول ۳ مشخص است که TESTCODE به درستی کد شدن قطعات را در بیشتر مواقع به جز تعداد خیلی محدودی به درستی پیش بینی می کند اما از آنجایی که ما از قطعات مشابهی هم برای محاسبه ی توزیع پارامتر ها و هم TESTCODE استفاده کرده ایم،الگوریتم احتمالا دچار remembering شده است.(Remembering لفظی است که در مقاله برای overfitting استفاده شده است.) برای حل این موضوع ما دیتابیسی را که در اختیار داشیم به دو بخش تقسیم کردیم،یک بخش برای محاسبه ی پارامتر ها و از اطلاعات بدست آمده از این بخش،بخش دوم را پیش بینی کردیم(train,test). نتیجه ی این کار به این شکل بود که تنها 5% این پیش بینی ها به کمک TESTCODE اشتباه بود و این موضوع تاکیدی بر این موضوع است که TESTCODE برای پیش بینی کد شدن قطعه ها به خوبی عمل می کند و می توانیم برای دیگر دیتابیس ها نیز از این الگوریتم بهره ببریم.

جزئیات بیشتر نحوه ی test,train :

ما ۳۲۹ قطعه ی کدینگ و همینطور ۲۴۹ قطعه ی غیر کدینگ داریم.برای محاسبه ی پارامتر های داخل جدول ۱ و همین طور وزن های جدول دوم از قطعه های با شماره ی فرد استفاده کردیم.بعد از آن قطعه های با شماره ی زوج را TESTCODE آنها را با استفاده از توزیع هایی که از بخش قبل بدست آمده بود محاسبه کردیم.نتیجه ی پیش بینی TESTCODE به این شکل بود که در رابطه با 18% داده ها "No Opinion" و 6% کدینگ ها را به اشتباه غیر کدینگ و 3% غیر کدینگ ها را به اشتباه کدینگ پیش بینی کرد.اطلاعات بیشتر را می توانیم در شکل ۳ ببینیم. ستون های سفید غیر کدینگ ها بوده و ستون های هاشوری کدینگ ها.ناحیه مستطیل میانی نیز "No Opinion"

شکل 2



یکی دیگر از موضوعاتی که خوب است به آن توجه کنیم ، این است که بهتر است در صورت وجود دیتابیس از دیتابیس ها متنوعی برای تست استفاده کنیم.برای مثال ما از مهره داران استفاده کردیم که شامل ۸۲ قطعه ی کدینگ و ۱۰۲ قطعه ی غیر کدینگ بوده و نتایج آن به این شکل بود که 12% قطعه ها را "No Opinion" و3% را به اشتباه کلاس بندی کرد.

در طول این آزمایش ما توجهمان را به قطعه های با طول حداقل ۲۰۰ محدود کردیم.متوجه شدیم که TESTCODE برای قطعات کمتر از ۲۰۰ غیر قابل اطمینان هست.زمانی که ما تست را بر روی ۵۷ قطعه ی غیر کدینگ و ۱۵۹ قطعه ی کدینگ با طول ۱۰۰ تا ۱۹۹ باز اجرا کردیم،13% اشتباه پیش بینی کرده و 29% "No Opinion".به این نتیجه رسیدیم که ۲۰۰ باز تعداد مینیمم مناسبی هست چون زمانی که پیش بینی ها برای طول های بازه های 299-200 و 399-300 و 499-400 و 599-500 و +600 انجام شد، اندازه ی error نزدیک به 5% بود.تفاوت اصلی که برای طول های بیش از ۲۰۰ بود،در رابطه با "No Opinion" بوده که برای 299-200 ، 24% و برای بیشتر از آن 15% بوده.

جمع بندی

ما از تفاوت های اساسی میان مناطق کدینگ و غیر کدینگ برای تولید یک الگوریتم ساده به نام TESTCODE برای تمایز کدینگ و غیر کدینگ با قاطعیت زیاد استفاده کردیم.زمانی که این الگوریتم را برای دیتابیس کتابخانه توالی Los Alamos استفاده کردیم ، به صورتی که نیمی از دیتا را به عنوان دیتای آموزش و نیم دیگر را به عنوان دیتای تست در نظر گرفتیم ، نتایج به این شکل بود که در رابطه با 18% از مناطق تست شده نتیجه ی No Opinion را دریافت کردیم و error rate ، 5% داشتیم.

روش تشخیص مناطق کدینگ از غیر کدینگ استفاده های بالقوه ی زیادی دارد از جمله ی آنها این است که زمانی که قطعه ای از DNA مشخص شد که دارای ژن، برای پروتئین خاصی است،ابتدا ایزوله می شود و سپس سکانس می شود.

یک مثال اخیر آن جست و جوی ژن E. coli trpR توسط Singleton و همکارانش است.نویسندگان مقاله سه ORF احتمالی را در نظر گرفتند و ORF درست را به کمک mutation analysis پیدا کردند. TESTCODE نیز تنها ORF درست را از میان دیگر ORF ها پیش بینی کرده بود.بنابراین TESTCODE و الگوریتم های مشابه ممکن است بتوانند آزمایش های زیست شناسان را کاهش دهند.

دومین کاربرد این متود در زمان هایی است که سکانس DNA به تازگی کشف شده باشد و شامل ORF های متفاوتی باشد.TESTCODE می تواند برای پیش بینی اینکه آیا این ORF ها توانایی کدکردن پروتئین جدید را دارد و یا نه تصمیم بگیرد و این می تواند یک تکنیک قدرتمند برای کشف پروتئین های جدید باشد.

سومین کاربرد می تواند به کار گیری TESTCODE برای بررسی صحت داده های زیست شناسان باشد. به عنوان مثال ما توانستیم به کمک TESTCODE چندین خطا را در کتابخانه ی توالی Lso Alamos بیابیم.

تصور ما بر این است که TESTCODE هم از نظر آزمایشگاهی و هم از نظر تئوری بسیار مفید است.در رابطه با آزمایشگاه ها،TESTCODE در مراحل ابتدایی آنالیز داده های توالی ها نقش ایفا می کند و از نظر تئوری ،TESTCODE برای بدست آوردن تفاوت های میان کدینگ و غیر کدینگ نقش ایفا می کند.

برخی از مشکلات این روش

TETSCODE روی توالی های کمتر از ۲۰۰ باز به خوبی عمل نمی کند.

TETSCODE برای مناطقی که بخشی از آنها کدینگ و بخشی دیگر غیر کدینگ است،خوب عمل نمی کند، بنابراین باید برای توالی هایی استفاده شود که یا به طور کامل کدینگ هستند و یا غیر کدینگ.

TESTCODE بخش زیادی از داده های تست را به عنوان No Opinion دسته بندی می کند و این موضوع برای توالی های با طول کمتر از ۲۰۰ باز بسیار بیشتر دیده می شود.