

کشف رابطه میان تنظیم بیان ژنها و تغییرات هیستون استیلاسیون با استفاده از شبکه عصبی

نفیسه بنی‌رضی مطلق و فاطمه زارع میرک‌آباد*

ایران، تهران، دانشگاه صنعتی امیرکبیر، دانشکده ریاضی و علوم کامپیوتر

تاریخ پذیرش: ۹۶/۶/۲۵

تاریخ دریافت: ۹۵/۲/۷

چکیده

استیلاسیون پروتئین‌های هیستونی یکی از مهم‌ترین فرآیندهای اپی‌ژنتیکی است که به منظور تنظیم بیان ژنها رخ می‌دهد. کروماتین به واسطه اتصال گروه استیل به دنباله هیستونی نوکلئوزوم‌هایش، رشته DNA را در دسترس فاکتورهای رونویسی و دیگر پروتئین‌های تنظیم‌کننده بیان ژن قرار می‌دهد. مطالعات نشان داده که نوع استیلاسیون نوکلئوزوم‌ها می‌تواند در شناسایی جایگاه پیوند فاکتورهای رونویسی یک سیگنال مهم باشد. در این تحقیق هدف یافتن یک روش محاسباتی برای پیشگویی جایگاه فاکتورهای رونویسی براساس الگوی نوع پراکندگی ۱۸ هیستون استیلاسیون است. در این راستا، الگوی پراکندگی ۱۸ هیستون استیلاسیون در سلول CD4+T انسان که اطراف فاکتور رونویسی SPI قرار گرفته‌اند مورد تحلیل قرار گرفت. نتایج نشان داد که از ۱۸ هیستون استیلاسیون ۱۲ نوع از آنها در شناسایی جایگاه فاکتور رونویسی SPI موثرند. سپس به وسیله تکنیک یادگیری با نظارت، یک شبکه چند لایه پرسپترون براساس جایگاه‌های پیوند فاکتور رونویسی SPI (استخراج شده از کروموزوم ۱ انسانی) و الگوی پراکندگی ۱۲ هیستون در اطراف آنها، آموزش داده شد. در نهایت از این شبکه برای پیشگویی جایگاه‌های فاکتور رونویسی SPI بر روی کروموزوم ۲ انسانی و جایگاه‌های ۱۲ فاکتور رونویسی دیگر بر روی کروموزوم‌های ۱ و ۲ استفاده گردید.

واژه‌های کلیدی: جایگاه پیوند فاکتور رونویسی، هیستون استیلاسیون، شبکه چند لایه پرسپترون، الگوریتم انتشار به عقب.

* نویسنده مسئول، تلفن: ۰۲۱۶۶۴۶۰۹۴۸، پست الکترونیکی: f.zare@aut.ac.ir

مقدمه

می‌دهند که ساختار کروموزوم می‌تواند به تعیین جایگاه پیوند فاکتور رونویسی کمک کند. به عبارت دیگر، کروموزوم‌ها برای اجرای عملکردهای مختلف سلول از قبیل همانندسازی، ترمیم، نوترکیبی و رونویسی DNA، متحمل تغییرات فراوانی می‌شوند. در هر لحظه قسمتی از ژنوم که نیاز به رونویسی دارد، در دسترس فاکتورهای رونویسی قرار می‌گیرد؛ در حالی که بقیه ژنوم به دقت محافظت می‌شود. این عملکرد در صورتی امکان‌پذیر است که ساختار کروماتین در شرایط گوناگون و در پاسخ به سیگنال‌هایی که دریافت می‌کند، به درستی تغییر نماید. از عوامل تغییر کروماتین در فرآیند تنظیم بیان ژنها می‌توان به تغییرات متیلاسیون و استیلاسیون دم هیستونی اشاره نمود

یکی از مسائلی که دانشمندان علم ژنتیک در سالهای اخیر در پی شناخت آن بوده‌اند، مکانیسم‌های گوناگون درون سلولی است که با همکاری یکدیگر، بیان ژنها را تنظیم می‌کنند. فاکتورهای رونویسی یکی از مهم‌ترین عوامل دخیل در تنظیم بیان ژنی هستند. این پروتئین‌ها در آغاز رونویسی به محل پروموتور ژن متصل می‌شوند و عمل رونویسی را فعال یا سرکوب می‌کنند (۳). بنابراین مطالعه جایگاه‌های اتصال فاکتورهای رونویسی به DNA می‌تواند نقش عمده‌ای در شناخت فرآیندهای تنظیم بیان ژنها داشته باشد. پژوهشگران در سه دهه اخیر روش‌های متنوعی برای کشف این جایگاه‌ها ارائه داده‌اند؛ اگرچه اکثر آنها دقت کافی برای پیشگویی ندارند (۱۴). مطالعات اخیر نشان

متفاوت است. به عبارت دیگر، برای بیان همه ژنها ترکیب یکسانی از تغییرات هیستونی نیاز نیست، بلکه هر ژن الگوی خاصی از این نوع تغییرات را برای بیان نیاز دارد (۷). رسمی در سال ۲۰۱۳، اثرگذاری هیستون استیلاسیونها را در تشخیص جایگاه پیوند فاکتورهای رونویسی بر روی ژنوم انسان آنالیز نمود (۱۲).

هدف در این مقاله ارائه یک روش محاسباتی برای پیشگویی مکان پیوند فاکتورهای رونویسی براساس الگوی تغییرات هیستون استیلاسیونهاست. برای دستیابی به این منظور، در مرحله اول به بررسی اثر هر یک از ۱۸ نوع هیستون استیلاسیون موجود در ژنوم انسان، برای شناسایی جایگاه عوامل الگوبرداری پرداخته شد. آنالیز داده‌ها نشان می‌دهد که ۱۲ نوع از ۱۸ نوع هیستون استیلاسیون در شناسایی جایگاههای پیوند اثر چشم‌گیری دارد. سپس با استفاده از یک مدل "دسته‌بندی (Classification)", جایگاههای پیوند فاکتورهای رونویسی براساس الگوی ۱۲ نوع هیستون استیلاسیون، پیشگویی می‌گردد. در این راستا، سه روش "رگرسیون لجستیک (Logistic regression)", "ماشین بردار پشتیبان (Support Vector Machine)", "شبکه پرسپترون چند لایه (Multilayer Perceptron)" (MLP) مورد مطالعه قرار گرفت. در این میان اگرچه LR از مزیت سادگی پیاده‌سازی برخوردار است، اما دو روش دیگر در حل مسائل با تعداد ویژگیهای متعدد و روابط پیچیده غیرخطی بین آنها موفق‌تر هستند (۹). نتایج حاصل از مطالعاتی که بر کارایی هر یک از روشهای SVM و MLP صورت گرفته است، نشان می‌دهد SVM در مسائلی با داده‌های کلان از سرعت بالای یادگیری و توانایی حصول تقریب محلی (Local approximation) برخوردار است. از سوی دیگر، گرچه MLP به زمان بیشتری برای آموزش نیاز دارد، اما می‌تواند به تقریب کلی (Global approximation) بهتری برسد؛ و در نتیجه برای این مسئله مناسب‌تر است (۱۰). این تحقیق یک شبکه MLP (Multilayer Perceptron) (پرسپترون

که با جا به جایی بار الکتریکی نوکلئوزوم، سبب کاهش نیروی جاذبه میان نوکلئوزوم و DNA پیچیده شده در اطراف آن می‌شود (۱ و ۱۱).

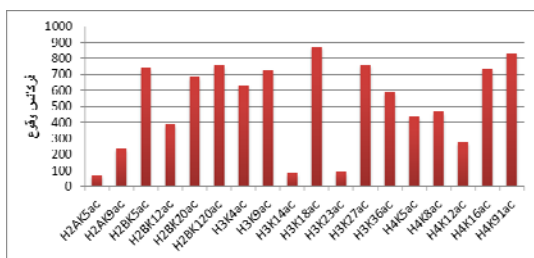
در زمینه تغییرات هیستونی متیلاسیون، طالبزاده در سال ۲۰۱۳ نشان داد که می‌توان با رگرسیون لجستیک از نواحی متیله شده کروماتین، به عنوان مارکری برای شناسایی جایگاه پیوند فاکتورهای رونویسی در انسان بهره برد (۱۴). همچنین طی مطالعه‌ای در سال ۲۰۱۴، مشخص شد که متیلاسیون در نواحی پروموتوری ژنهای سرکوبگر سرطان، منجر به بروز سرطان می‌شود (۲). در ضمن، ارتباط هیستون استیلاسیونها و تنظیم رونویسی در ژنوم مخمر مورد بررسی قرار گرفت که منجر به طرح دو فرضیه زیر گشت:

۱- فاکتورهای رونویسی به کروموزوم متصل می‌شوند و سپس استیل‌ترانسفرازها و داستیل‌ترانسفرازها را به کار می‌گیرند تا هیستونها را دچار تغییر استیلاسیونی کنند (۸).

۲- وقوع هر فاکتور رونویسی خاص وابسته به الگوی منحصر به فردی از تغییرات هیستونی است که رونویسی ژن را ممکن می‌سازد (۴، ۵ و ۱۳). این فرضیه با نام فرضیه "کد هیستونی" شناخته می‌شود.

نخستین بار، در سال ۲۰۰۶ ارتباط میان هیستون استیلاسیونها و فاکتورهای رونویسی مخمر به وسیله روش خوشه‌بندی مورد مطالعه قرار گرفت (۶). دو سال بعد براساس اولین فرضیه، یک شبکه تنظیم بیان ژنی براساس رگرسیون خطی برای مخمر ساخته شد (۱۱). ژیبین ونگ در سال ۲۰۰۸ الگوی تغییرات هیستونی در اطراف پروموتور ژنهای یوکاریوتی را بررسی نمود و نشان داد الگوهای ترکیبی هیستون استیلاسیونها باعث تنظیم بیان ژنهای خاصی در ژنوم انسان می‌گردد (۱۵). نتایج بیانگر این بود که میزان بالای یک تغییر خاص در اطراف برخی ژنها و از طرفی میزان بسیار کم آن در برخی دیگر، نشان دهنده تفاوت در الگوهای تغییرات هیستونی در اطراف ژنهای

یک فاکتور عمومی است که در اکثر پروموتورها رخ می‌دهد. مطالعات نشان می‌دهد که در بیشتر موارد در فاصله کمتر از ۱۰ جفت نوکلئوتیدی اطراف SP1، هیستونها دچار تغییرات استیلاسیون می‌شوند. نمودار ۱ فرکانس رخ داد هر نوع هیستون استیلاسیون در فاصله ۱۰ جفت نوکلئوتیدی جایگاه پیوند فاکتور رونویسی SP1 را نشان می‌دهد. ۱۲ عدد از هیستون استیلاسیونها که تعداد جایگاههای پیوند SP1 در فاصله ۱۰ جفت نوکلئوتیدی از آنها بیشتر از ۴۰۰ است، برای پیشگویی محل پیوند فاکتور رونویسی SP1 مورد استفاده قرار می‌گیرد (جدول ۱).



نمودار ۱- نمودار فرکانس وقوع ۱۸ هیستون استیلاسیون در فاصله کمتر یا مساوی ۱۰ جفت نوکلئوتید از جایگاه پیوند فاکتور رونویسی SP1 بر روی کروموزوم یک انسانی

جدول ۱- دوازده نوع هیستون استیلاسیون با فرکانس رخداد بالا (بیش از ۴۰۰) در فاصله ۱۰ جفت نوکلئوتیدی از جایگاههای پیوند فاکتور

رونویسی SP1

400 < Frequency		
H2BK5ac	H3K9ac	H4K5ac
H2BK20ac	H3K27ac	H4K8ac
H2BK120ac	H3K18ac	H4K16ac
H3K4ac	H3K36ac	H4K91ac

شبکه MLP برای تشخیص جایگاه پیوند فاکتور رونویسی: همان‌طور که نتیجه تحلیلها در بخش قبل نشان می‌دهد، در میان ۱۸ نوع هیستون استیلاسیون، ۱۲ نوع از آنها در ایجاد تمایز میان جایگاه پیوند فاکتورهای رونویسی از دیگر جایگاهها بیشتر حائز اهمیت است. در این بخش هدف این مطالعه ارائه یک شبکه MLP با معماری ۴ لایه است که بتواند جایگاه پیوند فاکتورهای رونویسی را براساس فاصله نزدیک‌ترین هیستون استیلاسیون از آنها

چندلایه) را بر اساس موقعیت هیستون استیلاسیونهای اطراف فاکتورهای رونویسی SP1 کروموزوم ۱ سلول CD4+T آموزش می‌داد. سپس از این شبکه برای پیشگویی جایگاههای پیوند فاکتور رونویسی SP1 بر روی کروموزوم ۱ و ۱۲ نوع فاکتور رونویسی دیگر بر روی هر دو کروموزوم ۱ و ۲ سلول انسانی استفاده گردید. نتایج مشابه در پیشگویی دیگر فاکتورهای رونویسی نشان می‌دهد که تغییرات هیستون استیلاسیون در اطراف این فاکتورهای رونویسی از الگوی یکسانی برخوردار است.

مواد و روشها

در این تحقیق مکان نوکلئوزومهای حاوی ۱۸ نوع هیستون استیلاسیون H2BK5ac, H2AK9ac, H2AK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K9ac, H3K18ac, H3K14ac, H3K27ac, H3K23ac, H3K18ac, H3K14ac, H3K9ac و H4K16ac, H4K12ac, H4K8ac, H4K5ac, H3K36ac و H4K91ac بر روی کروموزومهای ۱ و ۲ انسانی، مورد بررسی قرار می‌گیرد. این داده‌ها توسط ژیبین ونگ (۱۵) از سلول CD4+T که نقشی اساسی در سیستم ایمنی بدن انسان دارد، به وسیله تکنیک ChIP-Seq استخراج شده است. همچنین جایگاه پیوند فاکتورهای رونویسی در کروموزومها از پایگاه داده فانتوم به‌دست آمده است (۱۶).

تعیین هیستون استیلاسیونهای مؤثر در پیوند فاکتور رونویسی SP1: به منظور بررسی تأثیر هیستون استیلاسیونها در پیشگویی، نزدیک‌ترین نوکلئوزوم به جایگاه پیوند فاکتور رونویسی SP1 که دچار تغییرات استیلاسیونی شده است، برای هر نوع استیلاسیون نشانه گذاری شد. سپس فاصله این نوکلئوزومها را بر حسب جفت نوکلئوتید (Base Pair) با هریک از فاکتورهای رونویسی SP1 متناظر گردید.

علت انتخاب فاکتور رونویسی SP1 برای بررسی اثر هیستون استیلاسیونها این است که این فاکتور رونویسی

لایه ورودی شبکه، یک بردار ۱۲ نورونی است که به ازای موقعیت J از ژنوم به صورت زیر ساخته می‌شود:

$$\delta^J = [I_1^J, I_2^J, \dots, I_{12}^J],$$

$$i \in \{H2BK5ac, H2BK20ac, H2BK120ac, H3K4ac, H3K9ac, H3K18ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac, H4K16ac, H4K91ac\},$$

ورودی هر نورون از یک لایه به نورونی در لایه بعدی، از تابع سیگموئید $\varphi(x_1, x_2, \dots, x_s)$ استفاده می‌شود (۹):

$$\varphi(X) = \frac{1}{1 + e^{-(W \times X)}},$$

به طوری که s تعداد ورودیهای نورون لایه بعدی،

$$X = [1, x_1, x_2, \dots, x_s]$$

$$W = [w_0, w_1, w_2, \dots, w_s]$$

$$W \times X = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_s x_s$$

این دو بردار است. در نخستین مرحله از آموزش، مقادیر اولیه بردار W به طور تصادفی انتخاب می‌شود؛ و مقدار هزینه شبکه براساس رابطه زیر محاسبه می‌گردد (۱۲):

$$Cost(W) = \frac{-1}{2n} \sum_{k=1}^{2n} r^k \log(y^k) + (1 - r^k) \log(1 - y^k),$$

به طوری که y^k خروجی تولید شده توسط شبکه می‌باشد. به منظور کاهش هزینه از روش "انتشار به عقب (Backpropagation)" استفاده می‌گردد.

پس از انجام فرآیند یادگیری، اگر یک محل مشخص از دنباله DNA به عنوان ورودی به این MLP اضافه گردد، انتظار می‌رود که با در نظر گرفتن موقعیت ۱۲ هیستون استیلاسیون منتخب، تشخیص دهد که آیا آن ناحیه، جایگاه پیوند فاکتور رونویسی هست یا خیر.

نتایج

نتایج پیشگویی جایگاه پیوند فاکتور رونویسی SP1: برای یادگیری شبکه، ۱۹۰۳ جایگاه پیوند فاکتور رونویسی SP1 را از کروموزوم ۱ انسانی استخراج نموده و مجموعه داده مثبت (P) طبق تعریف آن در بخش ۴، ساخته می‌

پیشگویی کند. در ادامه معماری این شبکه توضیح داده می‌شود.

به طوری که I_i^J نشان دهنده تعداد جفت نوکلئوتیدهای بین نزدیک‌ترین هیستون استیلاسیون J به موقعیت J از ژنوم است. آخرین لایه شبکه دربردارنده یک نورون به نام y است و مقدار خروجی آن ۰ یا ۱ می‌باشد. برای یادگیری شبکه دو مجموعه مثبت و منفی به صورت زیر تعریف می‌گردد:

۱- با فرض اینکه جایگاه پیوند فاکتور رونویسی x در موقعیت J از ژنوم رخ داده باشد، بردار δ^J را ساخته می‌شود. سپس براساس n جایگاه پیوند فاکتور رونویسی x در ژنوم مجموعه داده مثبت $P = \{\delta^J_1, \dots, \delta^J_n\}$ حاصل می‌گردد.

۲- بردار δ^J ، با فرض اینکه موقعیت J از ژنوم، جایگاه پیوند هیچیک از فاکتورهای رونویسی نیست، ساخته می‌شود. سپس براساس n موقعیت از ژنوم که دارای این شرایط است، مجموعه داده منفی $N = \{\delta^J_1, \dots, \delta^J_n\}$ تولید می‌گردد.

همچنین بردار $R = [r^J_1, \dots, r^J_n, r^J_1, \dots, r^J_n]$ ، متناظر با اجتماع دو مجموعه مثبت و منفی، به صورت زیر تعریف می‌شود:

$$r^k = \begin{cases} 1 & k \in P, \\ 0 & k \in N. \end{cases}$$

شبکه MLP با استفاده از روش یادگیری "با نظارت (Supervised)"، تمایز بین نواحی جایگاه پیوند فاکتور رونویسی (مجموعه مثبت) و نواحی غیر پیوندی (مجموعه منفی) را آموزش می‌بیند. در هنگام آموزش، برای انتقال

می‌دهد که خطای پیش بینی براساس داده های آموزشی و آزمایشی مشابه است.

جدول ۲- پیشگویی داده‌های آموزشی و آزمایشی. (الف) نتیجه

پیشگویی داده‌های آموزشی واقع بر روی کروموزوم ۱.

(ب) نتیجه پیشگویی داده‌های آزمایشی واقع بر روی کروموزوم ۲.

الف	Positive Set	Negative Set
True Prediction	1661	1363
False Prediction	242	540

ب	Positive Set	Negative Set
True Prediction	851	915
False Prediction	240	394

شود. برای ساخت مجموعه داده منفی (N) نیز ۱۹۰۳ موقعیت که تا فاصله ۴۵ جفت بازی از آنها هیچ پیوندی با انواع فاکتورهای رونویسی رخ نداده باشد، در نظر گرفته می‌شود.

نتایج حاصل از آموزش و آزمایش این شبکه برای تشخیص مکان پیوند SP1 برای هر دو کروموزوم ۱ و ۲ انسانی در جدول ۲ قابل رؤیت است. در جدول ۳ نیز صحت پیشگویی با معیارهای گوناگون در مجموعه داده‌های آموزشی و آزمایشی قابل مشاهده می‌باشد. مقایسه ستونهای اول و دوم جدول (از راست به چپ) نشان

جدول ۳- روش محاسبه معیارهای متفاوت نرخ صحت پیشگویی و مقدار آنها در تشخیص جایگاه پیوند SP1 در کروموزومهای ۱ و ۲

		Chromosome 1	Chromosome 2
TPR	$\frac{TP}{(TP + FN)}$	0.87	0.78
TNR	$\frac{TN}{(TN + FP)}$	0.71	0.69
PPV	$\frac{TP}{(TP + FP)}$	0.75	0.68
F1	$\frac{2 \times PPV \times TPR}{(PPV + TPR)}$	0.8	0.72
ACC	$\frac{(TP + TN)}{(TP + TN + FP + FN)}$	0.79	0.73

نتیجه‌گیری

در این مقاله، تأثیر هیستون استیلاسیونها در تنظیم بیان ژنها مورد بررسی قرار گرفت. بر اساس نتایج به دست آمده، از میان ۱۸ هیستون استیلاسیون موجود در سلول انسان، ۱۲ نوع از آنها در وقوع پیوند فاکتور رونویسی SP1 مؤثرند. بر اساس این ۱۲ نوع هیستون استیلاسیون، محل پیوند فاکتور SP1 با 79 درصد و محل پیوند ۱۲ فاکتور رونویسی دیگر با میانگین TPR ۷۶ درصد در کروموزومهای ۱ و ۲ قابل پیشگویی است.

پیش‌گویی دیگر فاکتورهای رونویسی: در این بخش، هدف ارزیابی روش پیشگویی فاکتور رونویسی SP1، برای تشخیص دیگر فاکتورهای رونویسی است. به این منظور موقعیت هیستون استیلاسیونها در اطراف ۱۲ نوع فاکتور رونویسی موجود در کروموزومهای ۱ و ۲ به عنوان ورودی به شبکه عصبی داده شد و میزان TPR محاسبه گردید. جدول ۴ نتایج این پیشگویی را نشان می‌دهد. همانطور که قابل ملاحظه است میزان TPR فقط در مورد فاکتورهای رونویسی TFAP2A و ELF1 بر روی کروموزوم ۲ پایین‌تر از ۰/۷ است؛ در حالی که این مقدار برای همه فاکتورهای رونویسی دیگر بیشتر از ۰/۷ می‌باشد.

جدول ۴- نرخ صحت پیشگویی دیگر فاکتورهای رونویسی

TF	TPR in Chromosome 1	TPR	TPR in Chromosome 2	TPR
POL011	$\frac{380}{450}$	0.84	$\frac{212}{294}$	0.72
NFYA	$\frac{358}{423}$	0.84	$\frac{186}{239}$	0.77
TFDP1	$\frac{226}{268}$	0.84	$\frac{115}{138}$	0.83
ZNF278	$\frac{436}{529}$	0.82	$\frac{239}{278}$	0.85
MAZ	$\frac{812}{992}$	0.81	$\frac{497}{632}$	0.78
POL013	$\frac{242}{298}$	0.81	$\frac{140}{188}$	0.74
NRF1	$\frac{271}{341}$	0.79	$\frac{190}{229}$	0.82
TFAP2A	$\frac{244}{319}$	0.76	$\frac{123}{178}$	0.69
TFAP2B	$\frac{222}{290}$	0.76	$\frac{112}{154}$	0.72
GTF2I	$\frac{182}{237}$	0.76	$\frac{123}{161}$	0.76
bHLH	$\frac{155}{211}$	0.73	$\frac{80}{109}$	0.73
ELF1	$\frac{341}{473}$	0.72	$\frac{211}{310}$	0.68

علاوه بر این، در این مقاله فقط رخداد یا عدم رخداد پیوند فاکتورهای رونویسی با ژنوم تشخیص داده شد. در حالی که می‌توان اطلاعات مربوط به نوع فاکتور رونویسی را نیز در فرآیند پیشگویی دخیل کرد.

معیار پیشگویی در این تحقیق، فاصله نزدیک‌ترین هیستون استیلاسیون در بالادست یا پایین‌دست فاکتور رونویسی در نظر گرفته شد؛ پیشنهاد این تحقیق در جهت بهبود نتایج در تحقیقات آتی این است که فاصله نزدیک‌ترین هیستون استیلاسیون در پروموتور، ملاک ارزیابی قرار داده شود.

منابع

- ۱- حسین پور، م.، پژوهنده، م. و محمودی کردی، ف. ۱۳۹۵. شناسایی یک میانکنش مولکولی در کنترل اپی ژنتیکی بیان ژن FLC در گیاه اراییدوپسیس با استفاده از تکنیک دوپل هیبرید مخمر. مجله پژوهش‌های سلولی و مولکولی (مجله زیست‌شناسی ایران)، جلد ۲۹ شماره ۱ صفحه ۵۹-۷۱.
- ۲- محمدگنجی، ش.، مهدوی، ص. و رستگار جزی، ف. ۱۳۹۴. متیلاسیون پروموتور ژنهای p15، p14 و p16 در مبتلایان ایرانی سرطان بافت سنگفرشی مری. مجله پژوهش‌های سلولی و مولکولی (مجله زیست‌شناسی ایران)، جلد ۲۸ شماره ۳ صفحه ۴۰۳-۴۱۲.
- 3- Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2010. Essential Cell Biology, Garland Science.
- 4- Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarval, A.K. and Zhou, M.M. 1999. Structure and ligand of a histone acetyltransferase brodomain. Nature, 399(6735): 491-496.

- 5- Edmondson, D.G., Smith, M.M. and Roth, S.Y. 1996. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev.*, 10(10): 1247-1259.
- 6- Guo, X., Tatsuoka, K. and Liu, R. 2006. Histone acetylation and transcriptional regulation in the genome of *Saccharomyces cerevisiae*. *Bioinformatics*, 22(4): 392-399.
- 7- Horikoshi, N., Kumar, P., Sharma, G.G., Chen, M., Hunt, C.R., Westover, K., Chowdhury, S. and Pandita, T.K. 2013. Genome-wide distribution of histone H4 Lysine 16 acetylation sites and their relationship to gene expression. *Genome Integrity*. 4(1): 3.
- 8- Kuo, M.H. and Allias, C.D. 1998. Roles of histone acetyltransferases and deacetylases in gene regulation. *Bioessay*, 20(8): 615-626.
- 9- Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill Science.
- 10- Osowski, S., Siwek, K. and Markiewicz, T. 2004. MLP and SVM Networks. *NORSIG*.
- 11- Pham, H., Ferrari, R., Cokus, S.J., Kurdastani S.K. and Pellegrini, M. 2007. Modeling the regulatory network of histone acetylation in *Saccharomyces cerevisiae*. *Molecular Systems Biology*, 3:153.
- 12- Ramsey, S.A., Knijnenburg, T.A., Kennedy, K.A., Zak, D.E., Gilchrist, M., Gold, E.S., Johnson, CD., Lampano, A.E., Litvak, V., Navarro, G., Stolyar, T., Aderem, A. and Shmulevich, I. 2010. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 26 (17): 2071–2075.
- 13- Strahl, B.D. and Allis, C.D. 2000. The language of covalent histone modifications. *Nature*, 403: 41-45.
- 14- Talebzadeh, M. and Zare-Mirakabad, F. 2013. Transcription Factor Binding Sites Prediction Based on Modified Nucleosomes. *PLOS ONE*, 9(2): 1-10.
- 15- Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. and Zhao, K. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet.*, 40(7): 1-14.
- 16- http://fantom.gsc.riken.jp/4/download/GenomeBrowser/hg18/TFBS_CAGE

Transcription Factor Binding Sites Prediction Based on Histone Acetylations using Neural Networks

Banirazi Motlagh N. and Zare-Mirakabad F.

Dept. of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, I.R. of Iran

Abstract

DNA to transcription factors and gene regulators by histone tail acetylation in nucleosomes. There are some studies to show the relation between gene regulation and histone acetylation. In this paper, our main goal is to propose a computational method for transcription factor binding site prediction based on a pattern of 18 types of histone acetylations. In this regard, we analyze 18 types of histone acetylations near SP1 binding sites on Chromosome 1 in human CD4+T cells. The results show that 12 out of 18 marks are strongly correlated with transcription factor binding sites. Then, we implement a multilayer perceptron neural network with supervised training. This network is trained using binding sites of various transcription factors of SP1 in chromosome 1 and 18 types of histone acetylations near them. Finally, this network is applied for predicting binding sites of various transcription factors on chromosomes 1 and 2.

Key words: Transcription Factor Binding Site, Histone acetylation, Multi-layer Perceptron, Backpropagation Algorithm