۳ رویکرد

روشهای پیشین محاسباتی برای به دست آوردن اثرات ترکیبی درجه بالاتر در بین تغییرات هیستون موفق نبودند. این روش ها از استراتژی های مرتبط با بین استفاده می کردند که نمی توانستند نشان دهنده همسایگی بین ها باشند، یا این استراتژی ها متکی به روش های متعدد برای جدا کردن پیش بینی و آنالیز ترکیبی هستند. ما از یک مدل شبکه عصبی کانولوشن عمیق برای پیش بینی بیان ژن از داده های تغییرات هیستون استفاده می کنیم. شبکه به طور خودکار هم تعامل ترکیبی و هم طبقه بندی را به طور مشترک در یک چارچوب واحد افتراقی یکپارچه می کند، و نیاز به تلاش انسان در مهندسی ویژگی ها را از بین می برد. از آنجا که اثرات ترکیبی به طور خودکار از طریق چندین لایه از ویژگی ها آموخته می شوند، تکنیک تجسم را برای استخراج این تعاملات ارائه می دهیم و از این طریق مدل را قابل تفسیر می کنیم.

۳,۱ تولید ورودی

با هدف درک سیستماتیک رابطه بین تنظیم ژن و تغییرات هیستون، ما ۱۰،۰۰۰ جفت پایه ناحیه DNA (bp) را در اطراف محل شروع رونویسی (TSS) هر ژن به بین های با طول bp ۱۰۰ تقسیم کردیم. هر بین حاوی مقادیر bp ۱۰۰ در مجاورت طول TSS یک ژن است. در مجموع، پنج علامت تغیرات هیستون اصلی را از پایگاه داده REMC در نظر می گیریم، این داده ها در جدول ۲ خلاصه شده اند (کونداج و همکاران، ۲۰۱۵). این پنج تغییرات هیستون انتخاب شده اند زیرا آنها در همه انواع سلول هایی که در این مطالعه مورد بررسی قرار گرفته اند، یکپارچه هستند. این باعث می شود ورودی برای هر ژن ماتریس ۵ × ۱۰۰ باشد، که در آن ستون ها نشان دهنده بین های مختلف و ردیف نشان می دهد تغییرات هیستون هستند. برای هر بین، مقدار سیگنال هیستون ۵ را به عنوان ویژگی های ورودی برای آن بین گزارش می کنیم (شکل ۱). پیش بینی بیان ژن را به عنوان یک تسک دسته بندی باینری فرمول بندی می کنیم. به طور خاص، خروجی DeepChrome برچسب ها + ۱ و -۱ است، که بیانگر سطح بیان ژن به شکل زیاد و کم است. براساس نتایج چنگ و همکاران (۲۰۱۱)، ما از بیان ژن متوسط در تمام ژن ها وجود در یک نوع سلولی خاص به عنوان آستانه ای برای تفسیر هدف بیان ژن استفاده می کنیم. در شکل ۱ استراتژی تولید ماتریس ورودی ما خلاصه شده است.

دستگاه ما شبیه به چنگ و همکاران (۲۰۱۱) و دونگ و همکاران. (۲۰۱۲) است، به جز اینکه ما عمدتا در مناطق اطراف TSS تمرکز می کنیم به جای این که از نواحی ژن بدن یا محل پایان رونویسی (TTS) استفاده نماییم. این بر اساس مشاهدات چنگ و همکارانش است (۲۰۱۱) و نشان می دهد که سیگنال های نزدیک به TSS پر اطلاعات ترین سیگنال ها هستند، بنابراین، نیاز دستیابی به بین ها از مناطق انتهایی ژن ها حذف م شود. علاوه بر این، با توجه به مقیاس پذیری CNN ها، قادر به استفاده از مناطق بزرگتر TSS نسبت به مطالعات قبلی بودیم تا جلوه های سیگنال های دفاعی را بیشتر جذب کنیم و همچنین مناطق بیشتری را

پوشش دهیم. بنابراین، این امکان فراهم می شود تا مدل تعاملات طولانی مدت در تغییرات هیستون مدل سازی شوند.

جدول ۲- پنج علامت تغیرات هیستون اصلی ، همانطور که توسط کونداج و همکاران(۲۰۱۵) تعریف شده است، همراه با دسته های کاربردی آنها

Histone mark	Associated with	Functional category
H3K4me3	Promoter regions	Promoter mark
H3K4me1	Enhancer regions	Distal mark
H3K36me3	Transcribed regions	Structural mark
H3K9me3	Heterochromatin regions	Repressor mark
H3K27me3	Polycomb repression	Repressor mark

۳,۲ معماری پایان به پایان بر اساس شبکه عصبی کانولوشن

شبکه های عصبی کانولوشن (CNNs) برای اولین بار توسط لیکون و همکاران (۱۹۹۸) مورد استفاده قرار گرفتند و از آن زمان به طور گسترده ای برای بسیاری از برنامه ها استفاده می شوند. در این مقاله، ما یک CNN را برای انجام وظیفه طبقه بندی بیان ژن با استفاده از چارچوب Torch7 اجرا کردیم (کلوبرت و همکاران، ۲۰۱۱). مدل DeepChrome ما که در شکل ۲ خلاصه شده، از پنج مرحله تشکیل شده است. فرض می کنیم مجموعه آموزش ما شامل N_{samp} نمونه های ژن از فرم جفت برچسب خورده $(\mathbf{X}^{(n)}, y^{(n)})$ باشد، جایی که $\mathbf{X}^{(n)}$ ماتریس های اندازه $N_f\ (=5) \times b\ (=100)$ and $y^{(n)} \in \{-1, +1\}$ برای $n \in \{1, \ldots, N_{samp}\}$ هستند.

۱.کانولوشن: از کانولوشن موقتی با فیلترهای N_{out} استفاده می کنیم، طول هر کدام از آنها k است. این یک عمل پنجره لغزشی را در تمام موقعیت های بین انجام می دهد که یک نقشه خروجی از اندازه $N_{out} \times (b - k + 1)$ تولید می کند. هر پنجره عملیات لغزشی، N_{out} فیلتر خطی متنوعی را در k کد ورودی متوالی از موقعیت $p = 1$ به $(b - k + 1)$ اعمال می کند. در شکل ۲، یک مستطیل قرمز عملیات پنجره لغزشی را با $k=4$ و $p=1$ نشان می دهد. با توجه به نمونه ورودی x به اندازه $N_f \times b$، نقشه ویژگی Z کانولوشن به صورت زیر محاسبه می شود:

$$Z = f_{conv}(X)$$

$$Z_{p,i} = B_i + \sum_{j=1}^{N_f} \sum_{r=1}^{k} W_{ij,r} X_{p+r-1,j} \qquad (1)$$

این برای پنجره لغزشی همسایه p ام و فیلتر پنهان I ام ساخته شده است، که در آن $p \in \{1, \ldots, (b-k+1)\}$ و $i \in \{1, \ldots, N_{\text{out}}\}$. W با اندازه $N_{\text{out}} \times N_f \times k$ و B با اندازه $N_{\text{out}} \times 1$، پارامترهای قابل تمرین از لایه کانولوشون است، همچنین N_{out} تعداد فیلترها را نشان می دهد.

۲. تصحیح: در این مرحله، یک تابع غیر خطی با نام واحد خطی تصحیح شده (ReLU) اعمال می کنیم. ReLU یک عمل عاقلانه است که تمام مقادیر منفی را به صفر متصل می کند:

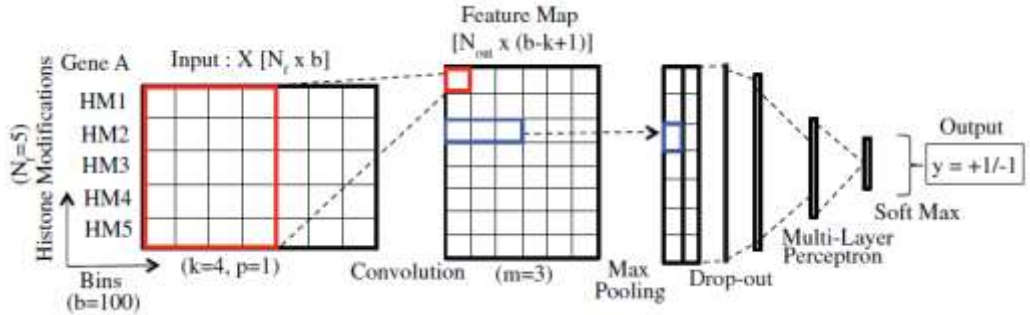
$$f_{\text{relu}}(z) = \text{relu}(z) = \max(0, z) \qquad (2)$$

۳. جمع آوری: در قدم بعدی، به منظور یادگیری ویژگی های تداخل ناپیوسته، ما از حداکثر پهنای موقتی در خروجی دو مرحله اول استفاده می کنیم. Maxpooling به سادگی حداکثر مقادیر در محدوده خاصی را انتخاب می کند، که مقادیر کوچکتر از یک منطقه نزدیک به مبدا بزرگ- TSS برای یک ژن خاص را تشکیل می دهد. Maxpooling روی ورودی Z اندازه $N_{\text{out}} \times P$ اعمال می شود، جایی که $P = (b - k + 1)$. در اندازه جمع کردن m، خروجی V با اندازه $N_{\text{out}} \times \lfloor \frac{P}{m} \rfloor$ بدست می آوریم:

$$\begin{aligned} \mathrm{V} &= f_{\text{maxpool}}(\mathrm{Z}) \\ \mathrm{V}_{i,p} &= \max_{j=1}^{m} \mathrm{Z}_{i,m(p-1)+j} \end{aligned} \qquad (3)$$

جایی که $p \in \{1, \ldots, \lfloor \frac{P}{m} \rfloor\}$ و $i \in \{1, \ldots, N_{\text{out}}\}$ هستند. در شکل ۲، مستطیل آبی نتیجه کار عملیات Maxpooling را در نقشه ویژگی نشان می دهد که در آن m=3 است.

۴. Dropout: خروجی سپس از یک لایه Dropout عبور می کند (سیرواستاوا و همکاران.، 2014)، که به طور تصادفی ورودی های لایه بعدی در طول آموزش را با انتخاب احتمال ۰,۵ صفر می کند. این کار شبکه را تنظیم می کند و مانع از فیت شدن بیش از حد می شود. این شباهت به تکنیک های گروهی مانند بگینگ یا متوسط سازی مدل است که در بیوانفورماتیک بسیار محبوب هستند.



شکل ۲. مدل شبکه عصبی کانولوشن DeepChrome (CNN). ماتریس ورودی x، شامل ۱۰۰ بین با سیگنال هایی از پنج اصلاح هیستون، این مدل دارای مراحل مختلف CNN هستند. این مراحل عبارتند از: کانولوشن، جمع کردن پس از

Dropout ، و MLP با جایگزینی خطوط خطی و غیر خطی. تابع Softmax در پایان، خروجی مدل را برای پیش بینی طبقه بندی می کنند.

۵. لایه های شبکه عصبی رو به جلو کلاسیک: سپس، نمایش منطقه آموخته شده وارد کلاسیفار MLP می شود تا تابع طبقه بندی تطبیق یافته با برچسب های بیان ژن یاد گرفته شوند. این شبکه استاندارد و کاملا متصل به شبکه MLP دارای چندین خط متناوب و غیر خطی است. هر لایه یاد میگیرد که ورودی را به یک فضای ویژگی پنهان وارد کند، و آخرین لایه خروجی از طریق یک تابع نقشه برداری softmax از یک فضای مخفی به فضای برچسب کلاس خروجی (۱-/۱+) یاد داده می شود. شکل ۲ MLP با ۲ لایه پنهان و یک تابع softmax در انتها را نشان می دهد. این مرحله به صورت $f_{mlp}(.)$ نمایش داده می شود.

فرم خروجی کل شبکه می تواند به صورت زیر نوشته شود:

$$f(\mathbf{X}^{(n)}) = f_{mlp}(f_{\text{maxpool}}(f_{\text{relu}}(f_{\text{conv}}(\mathbf{X}^{(n)})))) \qquad (4)$$

تمام مراحل فوق تکنیک های موثری هستند که به طور گسترده در زمینه یادگیری عمیق مورد استفاده قرار می گیرند. همه پارامترهایی که علامت Θ را به خود میگیرند، در طول تمرین به منظور به حداقل رساندن تابع تلفات آموخته شده که تفاوت بین برچسب های واقعی y و پیش بینی بهتر نمرات را از $f(.)$ یاد میگیرند. (هنگام آموزش این مدل عمیق، پارامترها، در ابتدا، به صورت تصادفی مقداردهی می شوند و نمونه های ورودی از طریق شبکه وارد می شوند. خروجی این شبکه پیش بینی مربوط به یک نمونه است. تفاوت بین پیش بینی خروجی f (X) و برچسب واقعی آن y از طریق مرحله " بازگشت به عقب " به شبکه برگردانده می شوند.). تابع تلفات L، در کل مجموعه آموزشی با اندازه n، به صورت زیر تعریف شده است:

$$L = \sum_{n=1}^{N_{\text{samp}}} \text{loss}(f(\mathbf{X}^{(n)}), y^{(n)}) \qquad (5)$$

از گرادیان نزولی تصادفی (SGD) (بوتو، 2004) برای آموزش مدل به واسطه «بازگشت به عقب» استفاده می کنیم. در مجموعه ای از نمونه های آموزشی، به جای محاسبه گرادیان واقعی هدف با استفاده از تمام نمونه های آموزش، SGD گرادیان هر نمونه را محاسبه می کند و بر اساس هر نمونه تمرینی آنها را به روز رسانی می نماید. برای تابع هدف ما، تلفات L(.) [معادله (۵)] توسط مرحله نزول گرادیان که برای به روز رسانی پارامترهای شبکه Θ به شرح زیر به حداقل می رسد:

$$\Theta \leftarrow \Theta - \eta \frac{\partial L}{\partial \Theta} \qquad (6)$$

که در آن η نرخ یادگیری می باشد (برابر با ۰,۰۰۱ تنظیم شده است).

۳,۳ تجسم اثر ترکیبی از طریق بهینه سازی

DeepChrome علاوه بر اینکه قابلیت پیش بینی دقیق وظایف بیان ژن را دارا است، نقش مهم دیگری نیز دارد که به ما امکان می دهد تا روابط ترکیبی بین تغییرات هیستون مختلف را کشف و تجسم کنیم، در واقع این که به چنین پیش بینی هایی منجر می شود. به تازگی، شبکه های عصبی عمیق به علت ویژگی های خودکار آموخته شده پوشش چندین لایه به عنوان "جعبه سیاه" شناخته می شدند. از آنجایی که بیان ژن به تعاملات ترکیبی بین تغییرات هیستون بستگی دارد، با اینحال این مهم است که بدانیم چگونه ویژگی های شبکه استخراج شده و پیش بینی های آن انجام می شوند. به عبارت دیگر، میخواهیم الگوهای ترکیبی تغییرات هیستون را درک کنیم، الگوهایی که منجر به پیش بینی نتایج ژن زیاد و کم توسط شبکه می شوند. تلاش می کنیم این کار با استخراج یک نقشه از ویژگی های الگو بسیار تاثیر گذار در پیش بینی بیان ژن که به طور مستقیم از شبکه آموزش می بیند، انجام می شود. این رویکرد، رویکرد شبکه محور نامیده می شو (یوسینسکی و همکاران، ۲۰۱۵)، ویژگی های خاص کلاس را از مدل آموزش دیده پیدا می کند و مستقل از نمونه های آزمایش خاص است.

تکنیک ما برای تولید این تجسم از کارهای سیمونیان و همکاران(۲۰۱۳) و یوسینسکی و همکاران (۲۰۱۵) الهام گرفته شده، که در پی این تا درک کند چگونه یک شبکه عصبی کانولوشنی یک کلاس تصویر خاص را در وظیفه تشخیص ابعاد تفسیر می کند. در عوض ما به دنبال پیدا کردن این هستیم که چگونه شبکه کلاس بیان ژن (زیاد و کم) را تفسیر می کند. با توجه به یک مدل CNN آموزش دیده و برچسب مورد علاقه (+۱ یا -۱) در مورد ما، یک روش بهینه سازی عددی در مدل برای ایجاد نقشه الگوی ویژگی که بهترین نشان دهنده کلاس است، اجرا می شود. این بهینه سازی شامل چهار مرحله عمده است:

۱. به صورت تصادفی ورودی X_c (of size $N_f(=5) \times b(=100)$) را اولویت بندی کنید.

۲. با بهینه سازی معادله زیر (۷) بهترین مقادیر Xc را پیدا کنید. برای به گونه ای برای Xc جستجو کردیم که تابع تلفات با توجه به برچسب های مورد نظر + ۱ (بیان ژن = بالا) یا -۱ (بیان ژن = کم) باشد. با استفاده از معادله (۴)، f(Xc) برچسب پیش بینی شده را با استفاده از مدل DeepChrome آموزش یافته در ورودی Xc ایجاد می کند. ما دوست داریم الگوی ویژگی مطلوب Xc را به نحوی پیدا کنیم که برچسب پیشنهادی f (Xc) نزدیک به برچسب کلاس مورد نظر c باشد:

$$\arg\min_{X_c} L_{\text{visual}} = \arg\min_{X_c}\{L(f(\mathrm{X}_c), y = c) + \lambda\|\mathrm{X}_c\|_2^2\} \qquad (7)$$

که در آن c = +1 یا ۱- است، L(.) تابع تلفات تعریف شده در معادله (۵) می باشد. تنظیم مقادیر L_2، $\|\mathrm{X}_c\|_2^2$، برای ایجاد مقادیر سیگنال در Xc و λ پارامتر تنظیم کننده است. Xc که به طور موضعی بهینه شده را می توان توسط روش بازگشت پشتی پیدا کرد. این مرحله شبیه به روش آموزش CNN است، که در

آن به منظور به حداقل رساندن تابع تلفات از پارامترهای شبکه Θ استفاده می شود. با این حال، در این مورد، بهینه سازی با توجه به مقادیر ورودی (Xc) و شبکه انجام می شود. پارامترها با توجه به مقادیر به دست آمده از آموزش طبقه بندی ثابت می شوند. Xc به روش زیر بهینه شده است:

$$X_c^{t+1} \leftarrow X_c^t - \alpha \frac{\partial L_{visual}}{\partial X_c} \qquad (8)$$

جایی که α پارامتر سرعت یادگیری و t نشان دهنده مرحله تکرار بهینه سازی است.

۳. سپس در مرحله بعد، تمام مقادیر خروجی منفی را بر ۰ تنظیم میکنیم و $X_c \in [0,1]$ را نرمال می کنیم:

$$X_{c(norm)} = \frac{X_c}{\max(X_c)} \qquad (9)$$

۴. در نهایت، ما یک آستانه ۰,۲۵ را برای تعریف بین "اکتیو" تنظیم می کنیم. بین ها در Xc (استاندارد) با مقادیر بیش از ۰,۲۵ در نظر گرفته می شوند زیرا نشان می دهد که چنین سیگنال هایی از تغیرات هیستون برای پیش بینی کلاس های خاص مهم هستند. تعداد فرکانس این بین ها اکتیو را با توجه به علامت اصلاح خاص هیستون شمارش می کنیم. تعداد فرکانس های بالا (بیشتر از میانگین تعداد فرکانس در تمام نشانه های تغیرات هیستون) بین اکتیو نشان دهنده تاثیر مهم این سیگنال های تغیرات هیستون بر سطح بیان ژن هدف می باشد.

این تکنیک تجسم نشان دهنده مفهوم کلاسی است که توسط مدل DeepChrome آموخته شده و به یک ژن خاص اختصاص ندارد. نقشه الگوی ویژگی بهینه شده $Xc_{(norm)}$ برای یک برچسب بیان خاص ژن ۱+ (بالا) یا ۱- (کم) نشان داده شده است. در شکل ۵، DeepChrome $Xc_{(norm)}$ را به عنوان نقشه های حرارت نمایش می دهد. از طریق این نقشه ها، خروجی های بصری برای درک اثرات ترکیبی تغییرات هیستون در تنظیمات ژن دریافت می کنیم.

۴ ست آپ آزمایش

۴,۱ مجموعه داده

ما سطوح بیان ژن و داده های سیگنال را برای پنج سیگنال اصلاح هیستون برای ۵۶ نوع سلول های مختلف از پایگاه داده REMC دانلود کردیم. REMC یک منبع عمومی از اطلاعات اپی ژنومیک انسانی است که از صدها نوع سلولی تولید می شوند. علامت اصلاح هسته هیستون، همانطور که توسط کونداج و همکاران تعریف شده (۲۰۱۵)، در جدول ۲ ذکر شده اند و مشخص شد نقش مهمی در تنظیم ژن دارند. ما در مورد این تغییرات هیستون، تمرکز کردیم؛ زیرا آنها از طریق تکنولوژی متوالی برای تمام انواع ۵۶ سلول، یکپارچه شده

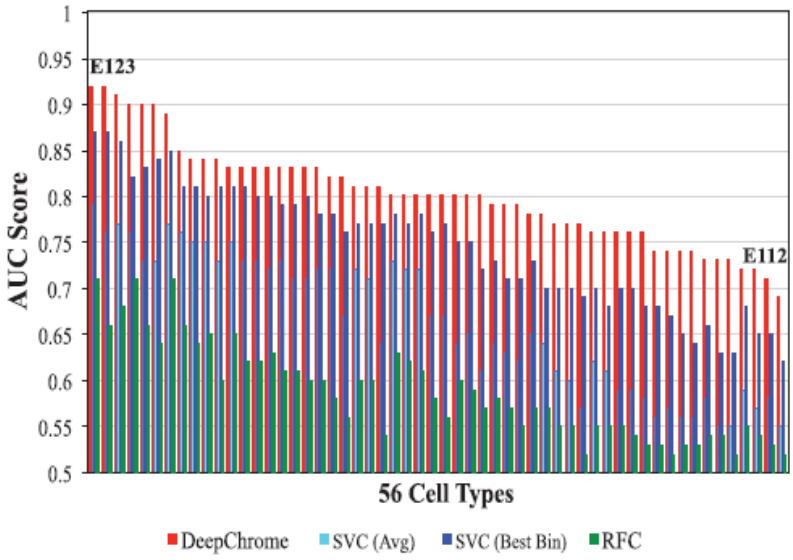
اند. داده های بیان ژن برای تمام ژن های حاوی نکته در ژنوم انسان محاسبه شده و برای تمامی ۵۶ سلول در پایگاه داده REMC نرمال شده است. همانطور که قبلا ذکر شد، مشکل هدف به عنوان یک تسک دسته بندی باینری فرموله شده است. بنابراین، هر ژن نمونه با برچسب ۱+/۱- به ترتیب نشان می دهد که بیان ژن زیاد و کم است. مقادیر بیان ژن با استفاده از میانه ی بیان ژن در میان تمام ژن ها برای یک سلول نوع خاصی تفسیر شدند.

۴,۲ خطوط

ما DeepChrome دو مطالعه خطوط چنگ و همکاران (۲۰۱۱)، که از ماشین های بردار پشتیبانی (SVM) استفاده کردند و دونگ و همکاران (۲۰۱۲) که از کلاسیفایر جنگل تصادفی استفاده نمودند را با هم مقایسه می کنیم. استراتژی های اجرای آنها به شرح زیر است:

• SVM (چنگ و همکاران، ۲۰۱۱): نویسندگان ۱۶۰ بین از مناطق حاوی ژن TSS و TTS را انتخاب کردند. هر موقعیت بنیادی از یک مدل طبقه بندی جداگانه SVM استفاده می کند که در مجموع ۱۶۰ مدل متفاوت را به همراه می آورد. این باعث بدست آوردن آگاهی از موقعیت های مهم بنیاد برای طبقه بندی بیان ژن به عنوان زیاد و کم می شود. پس از این استراتژی مدل خاص، نتایج عملکرد بهترین بین (SVM Best Bin) همراه با عملکرد متوسط در تمام بین ها (SVM Avg) را در بخش ۵,۱ و در شکل ۳ ارائه می کنیم.

• کلاسیفایر جنگل تصادفی (دونگ و همکاران، 2012): در این مطالعه، بین ها از مناطق متصل به TSS، TTS و ژن بدن انتخاب شدند.



شکل ۳- ارزیابی عملکرد در مجموعه تست. (بهترین رنگ در نظر گرفته شده) نمودار میله ای مقادیر AUC DeepChrome را نسبت به با مدل های خطوط جدید برای ۵۶ نوع سلول نشان می دهد (به عنوان مثال ۵۶ تسک طبقه بندی متفاوت). نتایج به دست آمده از بهترین نوع سلول (E123) تا بدترین نوع سلول (E112) برای مجموعه تست (۶۶۰۰ ژن) ارائه شده است. DeepChrome (میانگین AUC - 0.80) به طور مداوم بهتر از SVM (میانگین AUC: SVM

بهترین بین - ۰,۷۵ و 0.66 - SVM Avg) و کلاسیفایر جنگل تصادفی (میانگین 0.59 - AUC) برای تسک دسته بندی کردن باینری بیان ژن عمل نموده است. بر اساس مدل مبتنی بر SVM یک مدل جداگانه برای هر بین (مدل خاص بین) وجود دارد، بنابراین نتایج هر دو میانگین نمرات AUC در تمام بین ها (SVM Avg) و بهترین نمره AUC در بین همه بین ها (SVM Best Bin) ارائه می شود.

در این مطالعه مقادیر بین با بالاترین همبستگی برای بیان ژن به عنوان "بهترین بین" انتخاب شده اند. یک ماتریس با تمام ژن ها و بهترین بین ها برای هر سیگنال تغیرات هیستون به عنوان ورودی مدل برای پیش بینی برچسب های ژنی (+ ۱ / - ۱) به عنوان خروجی استفاده شد. از آنجا که این خطوط با انتخاب بهترین حالت و بهترین استراتژی بین عمل می کند، در آزمایش ما از بهترین بین عملکرد جنگل تصادفی به عنوان پایه در شکل ۳ استفاده شده است.

ما این پایه ها را با استفاده از بسته Scikit یادگیری مبتنی بر پایتون اجرا کردیم (پدروگسا و همکاران، ۲۰۱۱).

۴,۳ تنظیم هایپر پارامتر

برای هر نوع سلول، مجموعه نمونه ما از مجموع ۱۹۸۰۲ژن به ۳ قسمت مساوی تقسیم شده بود: آموزش (۶۶۰۱ ژن)، اعتبارسنجی (۶۶۰۱ ژن) و مجموعه (۶۶۰۰ ژن). ما DeepChrome را با استفاده از هایپر پارامترهای زیر آموزش دادیم: اندازه فیلتر $(k = \{10, 5\})$، تعداد فیلترهای کنولوشن $(N_{\text{out}} = \{20, 50, 100\})$ و حجم جمع آوری maxpooling $(m = \{2, 5\})$. جدول ۳ نتایج حاصل از اعتبارسنجی را برای تنظیم ترکیب های مختلفی از اندازه k و حجم جمع آوری m نشان می دهد. k نمایانگر بین های موضعی محسوب می شود. m نشان دهنده مناطق انتخاب شده در مدل CNN ما است. حداکثر، حداقل و میانگین نمرات AUC در ۵۶ نوع سلولی را به دست می آوریم. عملکرد مدل ها با استفاده از این مقادیر هایپر پارامتر تفاوت چندانی با هم دیگر نداشتند (P-value = 0.92). ما همچنین یک مدل عمیق را با ۲ لایه کانولوشن آموزش دادیم و هیچ افزایش قابل توجهی در عملکرد مشاهده نشد (P-value 0.939).

• ما مقادیر $k = 10, N_{\text{out}} = 50$ و M=5 را برای آموزش مدل نهایی CNN بر اساس بالاترین مقدار حداکثر و حداقلAUC در جدول ۳ انتخاب کردیم. تعداد واحدهای مخفی انتخاب شده برای دو لایه MLP به ترتیب ۶۲۵ و ۱۲۵ بود. ما مدل را برای ۱۰۰ دوره آموزش دادیم و متوجه شدیم که ابتدا آزمایش همگرا شدند (حدود ۱۵-۲۰ دوره).

• برای پیاده سازی SVM، یک هسته RBF انتخاب شد و مدل بر روی متغیرهای هایپرپارامتر $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$ و $\gamma \in \{0.01, 0.1, 1, 2\}$ آموزش داده شد. پارامتر C رابطه

بین نمونه های آزمایشی طبقه بندی نشده و سادگی تصمیم گیری را متعادل می کند، در حالی که پارامتر γ ، میزان تاثیر یک نمونه ی آموزشی را تعریف می کند.

• برای پیاده سازی کلاسیفایر جنگل تصادفی، تعدادی از درخت های تصمیم گیری را به $n_d \in \{10, 20, \ldots, 200\}$ در هر مدل آموزشی تغییر دادیم.

جدول ۳ نتایج مجموعه اعتبار سنجی (۶۶۰۱ ژن) در هنگام تنظیم ترکیب های مختلف از اندازه کرنل k و اندازه جمع آوری m

Kernel size, pool size (k, m)	AUC scores (validation set)		
	Max	Min	Mean
(5,2)	0.94	0.65	0.77
(5,5)	0.94	0.65	0.77
(10,2)	0.94	0.65	0.76
(10,5)	0.94	0.66	0.77

تمام مدل های فوق در مجموعه آموزشی آموزش دید شده اند و پارامترهای آزمون بر اساس نتایج آنها بر روی مجموعه اعتبار سنجی انتخاب گردیدند. سپس مدل های انتخاب شده را بر روی مجموعه داده های آزمون اعمال کردیم. نمرات AUC (معیار عملکرد) در بخش ۵,۱ گزارش شده اند [منحنی نمرات سطح زیر نمودار (AUC) بدست آمده از منحنی عملیاتی گیرنده (ROC) به عنوان احتمالی تفسیر می شوند که در آن یک رویداد که به صورت تصادفی با سوء ظن بیشتری از حالت غیر رویدادی که تصادفی انتخاب می گردد (از لحاظ اندازه گیری پیوستگی آن). نمره AUC بین ۰ و ۱، جایی که مقادیر نزدیک به ۱، پیش بینی های موفق تری را نشان می دهد.].

۵ نتایج

۵,۱ ارزیابی عملکرد

نمودار میله ای در شکل ۳، عملکرد DeepChrome و سه سطح پایه را در مجموعه داده های آزمون برای طبقه بندی بیان ژن در ۵۶ نوع مختلف سلول (یا وظایف) مقایسه می کند. DeepChrome (میانگین AUC = 0.80) از تمامی مقادیر سلولی که در امتداد محور x نشان داده می شود، پایه ریزتری دارد. همانطور که قبلا مورد بحث بود، چنگ و همکاران (۲۰۱۱) یک مدل SVM متفاوت برای هر موقعیت باینری اجرا می کند. بنابراین، ما هر دو میانگین نمره AUC برای همه بینها (SVM Avg) و همچنین بهترین نمره AUC را در میان تمام بینها (SVM Best Bin) گزارش می کنیم. SVM Best Bin (میانگین AUC = 0.75) نتایج بهتر از SVM Avg (میانگین AUC = 0.66) را نشان می دهد. با این حال، نمرات AUC آن هنوز

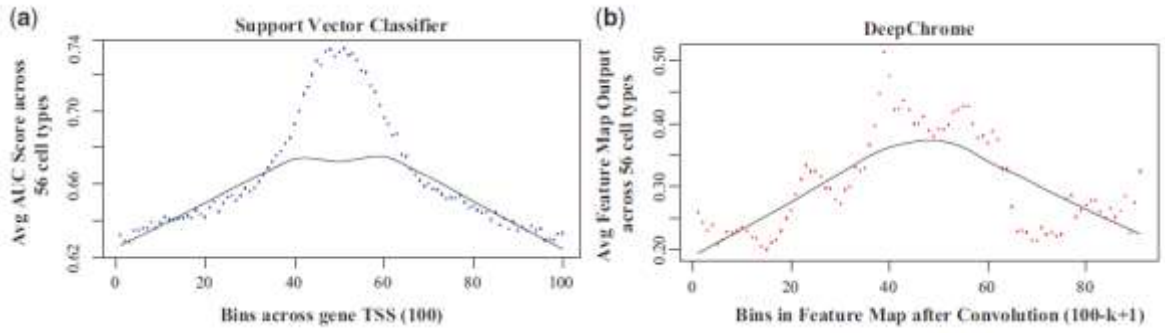
پایین تر از DeepChrome است. جنگل تصادفی بدترین عملکرد را به ارمغان آورده است (میانگین AUC = 0.59). علاوه بر این، مشاهده می کنیم که عملکرد هر سه مدل در انواع مختلف سلول ها متفاوت است و روند مشابهی را دنبال می کنند. برای برخی از انواع سلول، مانند E123، پیش بینی نتایج AUC بالاتری را در میان تمام مدل ها نسبت به سایر سلول ها به دست آورده است.

۵,۲ تأیید تأثیر موقعیت های بین در پیش بینی

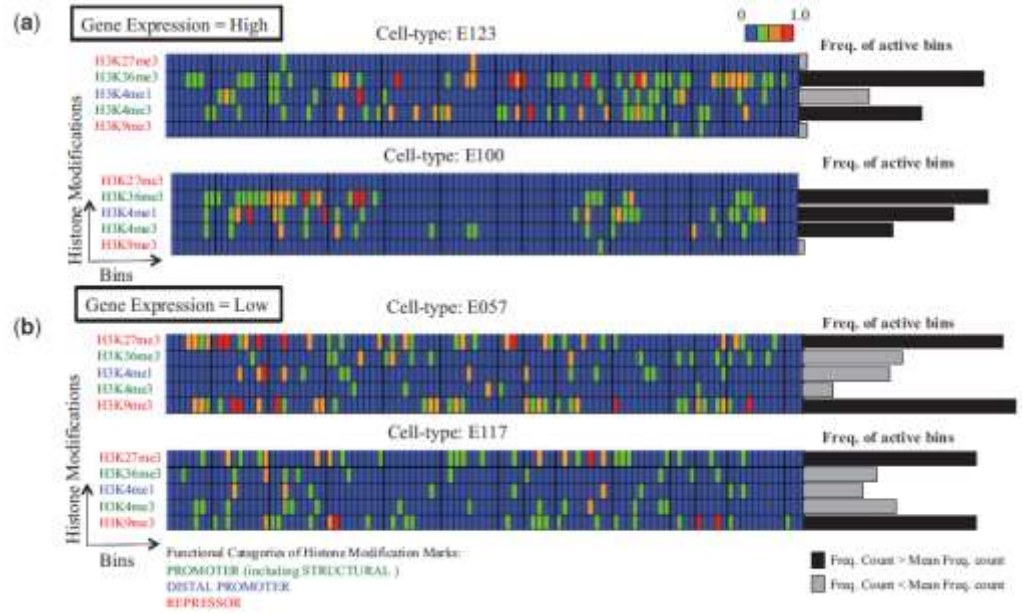
چنگ و همکاران (۲۰۱۱) پیش بینی های مربوط به هر سطر (به دلیل استراتژی بین مشخص) را دریافت کرد و گزارش دادند که به طور متوسط بهترین نمرات AUC از بین های نزدیک به TSS بدست آمد. شکل ۴ (الف) نشان می دهد که اجرای ما بر این خط پایه SVM این مشاهدات را تایید کرده است. از آنجایی که شبکه کانولوشن ما در کل منطقه ی جانبی (یعنی تمام بینها را حداقل یک بار) پیش بینی می کند، نمی توانیم AUC را برای هر یک از بین های فرد ارزیابی کنیم. با این حال، می توانیم تقریبا مشخص کنیم کدام سلول ها برای پیش بینی ژن خاصی تاثیر گذار هستند. برای انجام این کار، به قوی ترین اکتیو سازی در میان خروجی مرحله کانولوشن نگاه می کنیم (نقشه ویژگی، همانطور که در شکل ۲ نشان داده شده است). از آنجا که ستون در نقشه ویژگی مربوط به بین در منطقه ورودی است، می توانیم به سادگی به مقادیر نقشه ویژگی نگاه کنیم و تعیین کنیم که کدام موقعیت های بین برای این پیش بینی تاثیر گذار هستند. به منظور اعتبارسنجی مدل، تمام نمونه های آزمون را از طریق یک شبکه عمیق آموزش دیده اجرا کردیم و میانگین تمام نقشه های ویژگی در تمامی ۵۶ مدل را به دست آوردیم. شکل ۴ (b) نشان می دهد که بین ها نزدیک مرکز، نزدیک به TSS، با مقادیر بالاتر توسط لایه های کانولوشن اختصاص داده می شوند. این نشان می دهد که DeepChrome روند مشابهی را حفظ می کند که توسط چنگ و همکاران (۲۰۱۱) گزارش شد. این روند نشان می دهد که سیگنال های تغیرات هیستون بین ها که به TSS نزدیک تر هستند، در پیش بینی های ژن ها تاثیر بیشتری دارند.

۵,۳ تجسم تعاملات ترکیبی در میان تغییرات هیستون

برای تفسیر تعامل ترکیبی میان تغییرات هیستون، یک روش تجسم در بخش ۳,۳ ارائه دادیم.



شکل ۴ - تأیید تأثیر موقعیت ها برای طبقه بندی بیان ژن. چنگ و همکاران (۲۰۱۱) گزارش دادند که موقعیت های بنیادی نزدیک تر به محل شروع رونویسی (TSS) هر ژن در پیش بینی بیان ژن مهم هستند. این امر با اجرای ما از این مدل پایه خاص در (a) تایید شده است. برای هر موقعیت مکانی، میانگین نمره AUC را در تمام انواع سلول نشان می دهد. در (b)، خروجی های فیلتر را از لایه کانولوشن مدل DeepChrome ترسیم می کنیم. برای هر سطر، مقدار آن در تمام فیلترها و انواع سلول ها میانگین گرفته شده است. خطوط توپر بهترین منحنی را برای داده های ارائه شده در شکل نشان می دهند. روند (a) و (b) مشابه است.



شکل ۵. تجسم DeepChrome. (بهترین نما از نظر رنگ) چهار نمونه از نقشه های ساخته شده توسط چهار تکنیک بهینه سازی از مدل های آموزش دیده ارائه شدند. نمرات در این نقشه های ویژگی $\in [0, 1]$ هستند و آستانه ۰,۲۵ برای نشان دادن بین های "اکتیو" (یا مهم) انتخاب شده است. نمودار میله ای نشان دهنده تعداد دفعات اکتیو برای هر تغیرات هیستون است. بیشترین تعداد دفعات (بیشتر از میانگین تعداد فرکانس در تمام نشانه های هیستون) تاثیر بیشتر علامت تغیرات هیستون در پیش بینی برچسب های بیان ژن را نشان می دهد. علائم چندگانه با تعداد فرکانس بالا در نظر گرفته می شود که در بیان ژن به صورت زیاد و کم تاثیر می گذارد. (a) همانطور که انتظار می رود، ارتباط میان ترویج کننده و علامت های اصلاح ساختاری هیستون (H3K4me3 و H3K36me3) در هنگام بیان ژن زیاد است. (b) به طور مشابه، رویکردی مخالف (H3K9me3 و H3K27me3) از روابط ترکیبینشان می دهد که بیان ژن کم است. این الگوی نقشه نه تنها از مشاهدات قبلی چنگ و همکاران (۲۰۱۱) و دونگ و همکاران. (۲۰۱۲) پشتیبانی می کند، بلکه دیدگاه های جدیدی ارائه می دهد که توسط مطالعات بیولوژیکی جدید حمایت می شوند. به عنوان مثال، مطالعه اخیر توسط بروس و همکاران (۲۰۱۴) گزارش کرده است که شواهد همزیستی از تغییرات H3K27me3 و H3K9me3 در سکوت ژن گزارش شده است.

شکل ۵، چهار نتیجه از تجسم DeepChrome را در چهار نوع سلول با نمرات AUC بالا ارائه می دهد. هر نتیجه تجسم یک نقشه گرمایی است که الگوی ترکیبی هیستون را نشان می دهد و بیان می دارد که بهترین بیان ژن بالا (برچسب = +۱) یا کم (برچسب = −۱) است. توجه داشته باشید که این تفاوت با توجه به بخش ۵,۲ است که در آن ما اهمیت موقعیت بین را به طور کلی تأیید کردیم، نه تعاملات ترکیبی را برای کلاسی خاص. مقادیر حرارتی در محدوده [۰،۱] قرار دارند، که نشان دهنده اهمیت خاصی برای پیش بینی طبقه

مورد علاقه می باشند. آستانه ۰,۲۵ برای فیلتر کردن "اکتیو" یا مهمترین بینها انتخاب شد که برای طبقه بندی خاص تاثیر گذار هستند. ما تعداد دفعات اکتیو کردن بین ها برای هر گونه تغییر هیستون را به دست آورده ایم. علامت هیستون با تعداد فرکانس بالا (> میانگین تعداد فرکانس در تمام نشانه های هیستون) در نظر گرفته می شود که به شدت تحت تاثیر بیان ژن زیاد به کم است. همانطور که انتظار می رود، هنگامی که بیان ژن زیاد است، رابطه بین پروتئین ها و علامت های اصلاح ساختاری هیستون (H3K4me3 و H3K36me3) را برای ۴۷ مورد از ۵۶ (۸۴٪) نوع سلولی مشاهده می کنیم. به همین ترتیب، بیان علامت های سرکوبگر (H3K9me3 و H3K27me3) نشان می دهد زمانی که بیان ژن پایین است رابطه ترکیبی برای ۵۰ از ۵۶ (۸۹٪) سلول ها وجود دارد. به عبارت دیگر، مدل ما به طور خودکار یاد می گیرد که به منظور طبقه بندی کردن بیان ژن زیاد و کم، به ترتیب، باید تعداد زیادی از علامت های ترویج کننده یا مانع شونده را داشته باشد.

نتایج تجسم خودمان را با یافته های قبلی تایید کردیم. هر دو مقاله مرجع، چنگ و همکاران (۲۰۱۱) و دونگ و همکاران (۲۰۱۲) نشان داده شد که همبستگی ترکیبی بین H3K4me3 (علامت ترویج کننده) و H3K36me3 (علامت مانع شونده) وجود دارد. این الگوی را می توان در شکل ۵ برای موارد بیان ژن بالا دید. به همین ترتیب، دونگ و همکاران (۲۰۱۲) همبستگی ترکیبی بین علامت ترویج کننده (H3K4me3) و علامت مانع شونده را (H3K4me1) گزارش نمودند، که همچنین توسط تجسم DeepChrome برای ۳۵ از ۵۶ نوع سلول (۶۲,۵٪) تایید شده است. علاوه بر این، مطالعات تجربی نشان داده است که این ترویج کننده ها نقش مهمی در اکتیو سازی ژن ها دارند و این روند در تجسم ما زمانی دیده می شود که برچسب اختصاصی +۱ باشد.

یکی دیگر از الگوهای ترکیبی که در اکثر نمونه های سلولی (۸۹٪، یعنی ۵۰ از ۵۶ نوع سلول) مشاهده شد، H3K9me3و H3K27me3 برای نمونه های بیان ژن کم است (-1 = برچسب) ما این یافته ها را در چندین مطالعات زیست شناختی اخیر مانند بروس و همکاران (۲۰۱۴) مشاهده کردیم. این مطالعه نشان می دهد که این دو علامت مخالف همسو می شوند و در خاموش شدن ژن همکاری می کنند. تقریبا بدون دانش تخصصی، توانستیم این ترکیب را از طریق DeepChrome پیدا کنیم و تجسم کنیم. به نظر ما، هیچ یک از مطالعات محاسباتی قبلی این اثر ترکیبی بین H3K9me3 و H3K27me3 را گزارش نکرده است. به طور خلاصه، روش DeepChrome توانایی یادگیری عمیق در روابط ترکیبی بین تغییرات هیستون و تنظیم ژن را فراهم می کند.

۶ بحث

ما DeepChrome را در یک چارچوب یادگیری عمیق ارائه کرده ایم که نه تنها سطح بیان ژن را به دقت با استفاده از تغییرات هیستون به عنوان ورودی طبقه بندی می کند، بلکه همچنین روابط ترکیبی بین این علائم

تغییراتی که ژن ها را تنظیم می کند را نیز یاد می گیرد. ما یک مدل مبتنی بر شبکه عصبی کانولوشن اجرا می کنیم که از کار یادگیری عمیق در برنامه های تشخیص تصویر الهام گرفته شده و عملکرد آن را در ۵۶ نوع سلولی از آخرین مجموعه داده های REMC ارزیابی می کند. از نظر ما، برای این اولین بار است که آموزش یادگیری عمیق در وظیفه طبقه بندی بیان ژن با استفاده از سیگنال های تغیرات هیستون صورت گرفته است.

DeepChrome با استفاده از SVM و جنگل تصادفی برای تسک هدف بیشتر از ۵۶ نوع سلولی (یا وظایف) بهتر از مدل های پیشرفته تر عمل می کند. علاوه بر این، یک استراتژی بهینه سازی برای استخراج روابط ترکیبی بین تغییرات هیستون از مدل های آموزش یافته پیشنهاد می کنیم. یافته های ما نه تنها مشاهدات قبلی را تایید می کنند، بلکه بینش جدید برای مکانیسم های تنظیم ژن های که در مطالعات تجربی اخیر مشاهده شده، ارائه می نمایند. یادآوری می کنیم که این بینش در حال حاضر محدود به منابع جستجو شده است. بنابراین، الگوهای هیستون بهینه شده از مدل های DeepChrome را برای تمامی انواع ۵۶ سلول برای بیان ژن آنلاین را به صورت کم و زیاد طبقه بندی می کنیم (www.deepchrome.org). ما امیدواریم که زیست شناسان بتوانند از این نتایج برای تهیه فرضیه های مهم در مورد متابولیسم تغییر هیستون که موجب اکتیو شدن یا خاموش شدن ژن می شود، استفاده کنند.

برای کارهای آینده، ما می خواهیم عملکرد DeepChrome را حین اضافه کردن اصلاحات هیستون در نظر بگیریم تا اثرات ترکیبی خود را نیز ببینیم. همچنین پیش بینی های سلولی را انجام می دهیم، که در آن یک مدل بر روی داده های سلولی آموزش داده می شود و پیش بینی ها بر روی انواع دیگر سلول ها انجام می گیرد. مطالعات قبلی گزارش داده اند که همبستگی بین تغییرات هیستون در انواع سلول ها همیشگی است. با این حال، کاهش عملکرد (دم راست در شکل ۳) برای برخی از انواع سلول ها در نتایج ما نشان می دهد که پتانسیل کشفیات بیشتری در این موضوعات وجود دارد. یکی دیگر از رویکردهای معتبر، شناخت اثر روابط بین تغییرات هیستون برای تنظیم ژن های فردی است. این می تواند به زیست شناسان در طراحی "داروهای اپی ژنتیک" کمک کند که می تواند علامت های تغیرات هیستون و بیان یک هدف خاص ژن را کنترل کند.

به طور خلاصه، DeepChrome چند راه جدید برای مطالعه و بررسی تنظیمات ژنتیکی از طریق عوامل اپی ژنتیک باز می کند. این به دلیل توانایی یادگیری عمیق برای رسیدگی به حجم زیادی از داده های موجود است و همچنین به طور خودکار ویژگی های مهم و تعاملات پیچیده را استخراج می کند، به ما امکان می دهد بینش های مهم را ایجاد کنیم. تکنیک هایی مانند DeepChrome توانایی ما را یک گام به بررسی درست تر مکانیزم های تنظیم ژن نزدیک می کند، که به نوبه خود می تواند منجر به درک بیماری های ژنتیکی شود.