# دانشگاه شهید بهشتی

## دانشکده‌ی علوم ریاضی

پایان‌نامه‌ی کارشناسی ارشد

ریاضی محض

عنوان:

# الگوریتم *Libra*:

# روشی جدید برای یافتن موتیف بر مبنای جبر خطّی

۱۳۸۷/۱۰/ - ۶

استاد راهنما: دکتر چنگیز اصلاح‌چی

استاد مشاور: دکتر مهدی صادقی

نگارش:

**علیرضا شیخ‌عطّار**

زمستان ۱۳۸۶

تاریخ .....
شماره .....
پیوست .....

دانشگاه شهید بهشتی

«بسمه تعالی»+

# «صورتجلسه دفاع از پایان نامه دانشجویان دوره کارشناسی ارشد»

ان ۱۹۸۳۹۶۳۱۱۳ اوین
ز: ۲۹۹۰۱

بازگشت به مجوز دفاع شمارهٔ **۲۰۰/۳۹۳۲/ت/د** مورخ **۸۶/۱۱/۸** جلسه هیأت داوران ارزیابی پایان نامه: **آقای علیرضا شیخ عطار** شماره شناسنامه: **۲۶۸۹** صادره از: **تهران** متولد: **۱۳۶۰** دانشجوی دوره کارشناسی ارشد: **ریاضی محض**

با عنوان:

**الگوریتم Libra : روش جدید برای یافتن موتیف بر مبنای جبر خطی**

به راهنمایی:

**آقای دکتر چنگیز اصلاح چی**

طبق دعوت قبلی در تاریخ **۸۶/۱۱/۱۰** تشکیل گردید و بر اساس رأی هیأت داوری و با عنایت به ماده ۲۰ آئین نامه کارشناسی ارشد مورخ ۷۵/۱۰/۲۵ پایان نامه مزبور با نمره ۱۹,۷۵ (نوزده و هفتاد و پنج) و درجه عالی مورد تصویب قرار گرفت.

| | مرتبه علمی | نام دانشگاه | امضاء |
|---|---|---|---|
| ۱- **استاد راهنما:** آقای دکتر چنگیز اصلاح چی | دانشیار | شهید بهشتی | |
| ۲-**مشاور:** آقای دکتر مهدی صادقی | استادیار | پژوهشگاه مهندسی ژنتیک | |
| ۳- **داور:** آقای دکتر سهرابعلی یوسفی | استادیار | شهید بهشتی | |
| ۴- **داور:** آقای دکتر حمید پزشک | دانشیار | تهران | |
| ۵- **مدیر گروه:** آقای دکتر علیرضا سالمکار | استادیار | شهید بهشتی | |

# الگوریتم *Libra*:

## روشی جدید برای یافتن موتیف بر مبنای جبرخطّی


### چکیده

امروزه یک مسئله‌ی کلّی در بیوانفورماتیک، یافتن الگوهای تقریباً مشابه (موتیف) روی توالی‌های *DNA* و پروتئین می‌باشد که توجّه بسیاری از زیست‌شناسان و دانشمندان علوم‌کامپیوتر و ریاضی را به خود جلب کرده‌است. به طور خاص، یافتن مکان‌های اتّصال یک فاکتور نسخه برداری روی *DNA* از مهم‌ترین مسائل پیداکردن موتیف‌ها به شمار می‌آید. الگوریتم‌های بسیاری برای یافتن موتیف‌ها ارائه شده است امّا به سختی می‌توان از بین آنها بهترین را انتخاب کرد. این در حالی است که به جز تعریف بیولوژیک موتیف، تعریف روشنی در محاسبات کامپیوتری برای آن ارائه نشده است که به توافق همه رسیده باشد. علاوه بر آن تقریباً همه‌ی الگوریتم‌های موجود در مقابل داده‌های گوناگون حسّاس هستند و دیده شده است که الگوریتمی برای *DNA* های گونه‌های مختلف به طور موفقیت‌آمیز کار می‌کند امّا برای گونه‌های پیچیده‌تری مانند انسان موفقیت‌آمیز نیست. بالاخره اینکه الگوریتم‌های متفاوت اغلب پارامترهای متفاوتی به عنوان ورودی در نظر می‌گیرند که قابل تبدیل به هم نیستند. یعنی به سختی می‌توان شرایط یکسانی را برای مقایسه الگوریتم‌ها فراهم‌کرد. با توجّه به این دلایل مقایسه عملکرد الگوریتم‌های موجود کاری بسیار مشکل است. وجود چنین نقایصی انگیزه‌ای شد تا الگوریتمی ارائه دهیم که اوّلاً هر گونه از داده‌ها را بدون حسّاسیت بپذیرد و ثانیاً با کمترین پارامترهای لازم بتواند به جوابی راضی‌کننده در زمانی معقول دست یابد. ما در این پایان‌نامه ابتدا با دسته‌بندی الگوریتم‌های موجود آن‌ها را بررسی می‌کنیم. در گام بعد الگوریتمی به نام *Libra* ارائه می‌دهیم که بر مبنای جبرخطّی ابتدا فضای جستجو را به زیرقطعه‌هایی از توالی‌های *DNA* کاهش داده و سپس با نگرشی جدید از شباهت در موتیف، قطعه‌های مشابه را پیدا می‌کنیم. سپس از روی آنها مدلی برای موتیف می‌سازیم و آنگاه با مدل‌هایی که از موتیف بدست می‌آوریم به جواب‌های راضی‌کننده‌ای در زمانی معقول دست می‌یابیم. در پایان این الگوریتم را با یکی از معروفترین الگوریتم‌های موجود به نام *MEME* مقایسه خواهیم کرد.



**واژه‌های کلیدی:** ۱) فضای برداری، ۲) گراف، ۳)همردیفی دوگانه و چندگانه، ۴)پیچیدگی محاسباتی.

# ۱

# مقدّمه

## ۱.۱ نگاهی بر زیست‌شناسی مولکولی

### ۱.۱.۱ پیش‌درآمد

مادّه‌ی وراثتی در همه‌ی موجودات زنده *DNA*[۱] است[۱]. زمانی که یک سلّول در بدن موجود زنده در حال تقسیم شدن است، برگردان‌هایی از *DNA* ساخته می‌شود و هر سلّول نوزاد یک برگردان از *DNA* را دریافت می‌کند. توالی[۲] *DNA* همه‌ی اطّلاعات وراثتی یک موجود زنده که ژن[۳] نام دارند را در بر دارد که به عنوان الگوی ساخت *RNA*[۴] به کار می‌روند. سپس *RNA* می‌تواند همانند یک قالب برای تولید پروتئین مورد استفاده قرارگیرد. بروز صفات سلّول و انجام فعّالیت‌های حیاتی، به عهده‌ی پروتئین‌ها می‌باشد.

---

۱ *Deoxy – ribonucleic Acid*
۲ *Sequence*
۳ *Gene*
۴ *Ribonucleic acid*

## ۱.۱.۲ *DNA*

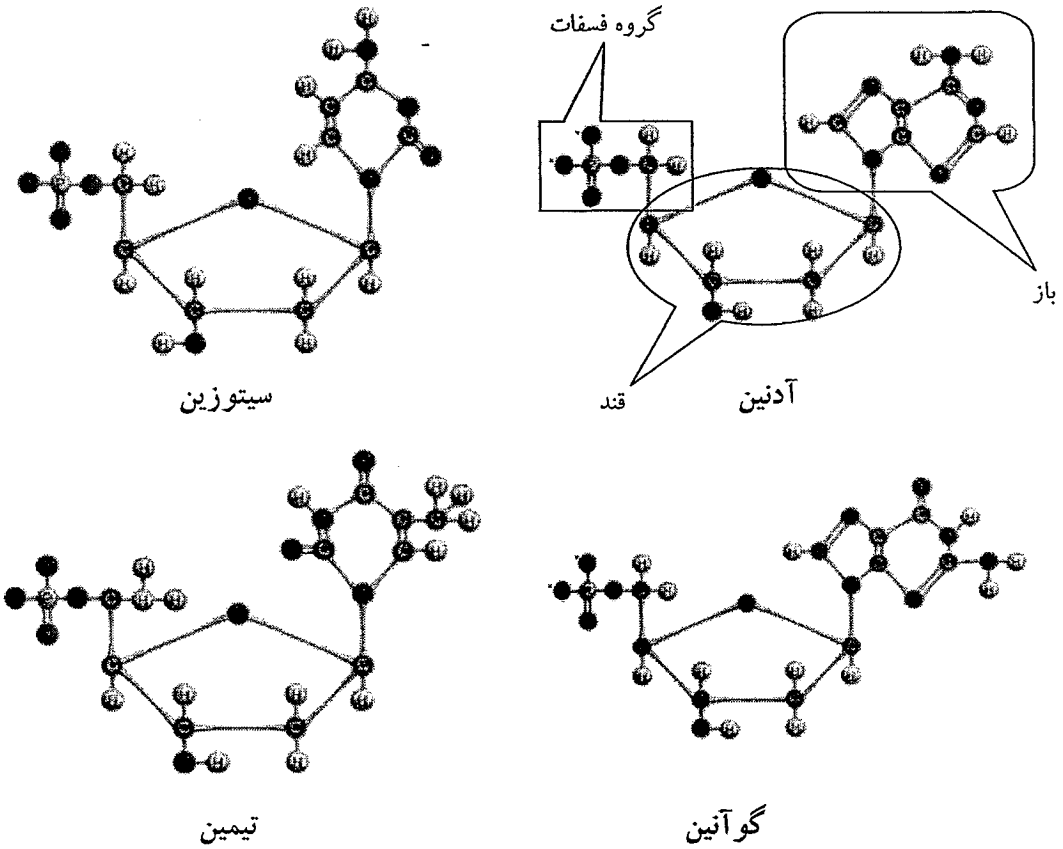
*DNA* زنجیره‌ای از نوکلئوتیدهای[۵] به هم متّصل می‌باشد. هر نوکلئوتید از سه قسمت تشکیل شده است(شکل ۱.۲.۱.۱):

۱) یک عدد مولکول قند

۲) یک عدد گروه فسفات

۳) یکی از چهار بازآلی یعنی آدِنین[۶] $(A)$، سیتوزین[۷] $(C)$، گوانین[۸] $(G)$ و یا تیمین[۹] $(T)$.

نوکلئوتیدهای متفاوت دارای بازهای آلی متفاوتی هستند. لذا صرف نظر از چگونگی اتّصال نوکلئوتیدها به‌یکدیگر، توالی نوکلئوتیدها را با توالی بازهای آلی‌شان مشخّص می‌کنیم. این چهار بازآلی در واقع حروف رمز برای توالی *DNA* هستند.



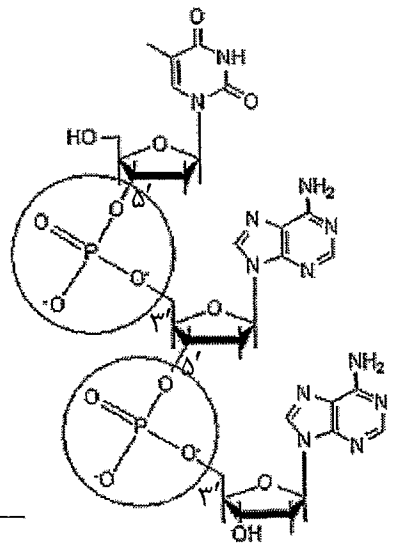

شکل ۱.۲.۱.۱

---

[۵] *Nucleotide*
[۶] *Adenine*
[۷] *Cytosine*
[۸] *Guanine*
[۹] *Thymine*

قندی که در هر نوکلئوتید وجود دارد دارای دو کربن به نام کربن شماره‌ی سه و کربن شماره‌ی پنج می‌باشد که به ترتیب به صورت کربن $۳'$ و کربن $۵'$ نشان می‌دهیم. گروه فسفات که به کربن $۵'$ هر نوکلئوتید متّصل است به کربن $۳'$ نوکلئوتید بعدی می‌پیوندد(شکل ۲.۲.۱.۱). به این صورت پیوندی به نام فسفو-دی‌استر[۱۰] نوکلئوتیدها را مانند یک زنجیر به هم متّصل می‌کند. ساختار $DNA$ به شکل دو رشته‌ی مارپیچ در هم تنیده می‌باشد(شکل ۳.۲.۱.۱). اگر این مارپیچ از هم باز شود، حاصل دو رشته‌ی موازی از نوکلئوتیدهایی خواهدبود که جفت جفت روبروی هم قرارگرفته‌اند(شکل ۴.۲.۱.۱). پیوندهای هیدروژنی موجود بین جفت بازآلی روبروی هم، این دو رشته را به هم چسبانده‌است. هر جفت بازآلی روبروی هم، شامل یک باز پورین[۱۱] ($A$ یا $G$) و یک باز پریمیدین[۱۲] ($T$ یا $C$) می‌باشد که طبق قانون، $G$ فقط با $C$ جفت می‌شود و $A$ فقط با $T$. لذا با دانستن توالی یکی از دو رشته‌ی $DNA$، رشته‌ی مقابل آن به طور یکتا بدست می‌آید. در واقع دو رشته مکمّل یکدیگرند. این رابطه بین دو رشته‌ی $DNA$ را رابطه‌ی مکمّلی می‌گوئیم. برای همین اغلب اوقات توالی $DNA$ را با یک رشته بیان می‌کنند.

از آنجایی که پیوند فسفو-دی‌استر، کربن‌های $۵'$ را به $۳'$ هم متّصل می‌کند، زنجیره‌ی نوکلئوتیدها دارای جهت بوده و این جهت را به صورت $۵' \rightarrow ۳'$ نشان می‌دهند. هنگام تشکیل مارپیچ، رشته‌ها به صورت موازی متقابل قرار می‌گیرند. یعنی اگر روی یکی از رشته‌ها در جهت $۵' \rightarrow ۳'$ حرکت کنیم، رشته‌ی دیگر را در جهت $۳' \rightarrow ۵'$ خواهیم‌دید. به جهت $۵' \rightarrow ۳'$ ($۳' \rightarrow ۵'$) بالادست[۱۳] (پائین‌دست[۱۴]) گویند.

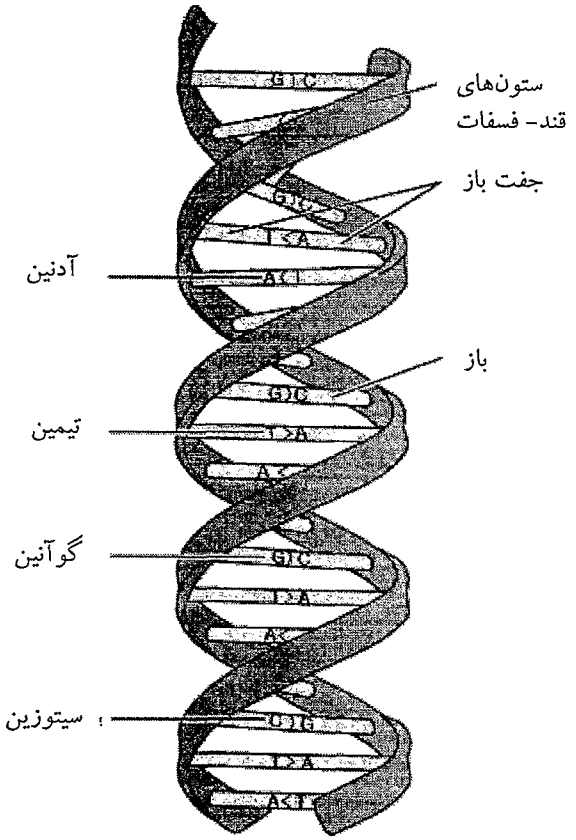



شکل ۲.۲.۱.۱

پیوند فسفو دی-استری: از پیوند کربن $۵'$ به کربن $۳'$، دو نوکلئوتید متوالی به هم متّصل می‌گردند.

[۱۰] *Phospho diester*
[۱۱] *Purine*
[۱۲] *Pyrimidine*
[۱۳] *Upstream*
[۱۴] *Downstream*

تعداد توالی‌های *DNA* در یک سلّول موجودات زنده‌ی گوناگون، متفاوت است. در هر سلّول انسان، ۴۶ توالی *DNA* وجود دارد که هر *DNA* در یک کروموزوم[۱۵] بسته‌بندی شده‌است. این ۴۶ کروموزوم به صورت ۲۳ جفت در سلّول انسان یافت می‌شود. هر ژن ناحیه‌ای بر روی *DNA* است که شامل اطّلاعاتی از یک خصوصیت وراثتی می‌باشد. همه‌ی اطّلاعات وراثتی ذخیره‌شده در کروموزوم‌های یک موجود زنده را ، ژنومِ آن موجود می‌نامند. در ژنوم انسان نزدیک به ۱۰۰۰۰۰ ژن وجود دارد که طول ژن‌های شناسایی شده بین ۱۰۰ تا ۲۳۰۰۰۰۰ نوکلئوتید است [۲].



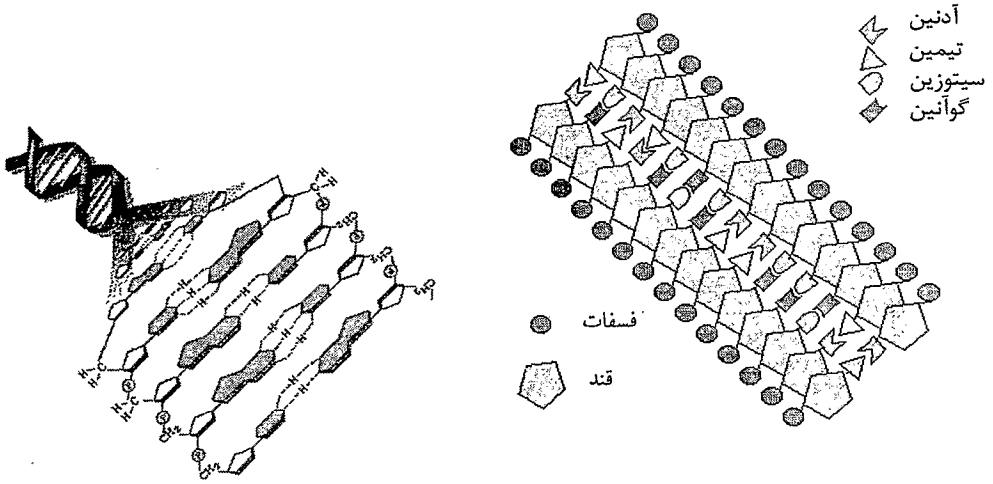

شکل ۳.۲.۱.۱

در سال ۱۹۵۳، تقریباً ۲۰ سال بعد از کشف ساختمان مولکولی *DNA*، ساختمان سه بعدی *DNA* بوسیله‌ی جیمز واتسون و فرانسیس کریک کشف شد. واتسون و کریک با استفاده از مطالعات تفرّق اشعه‌ی *X* رشته‌های *DNA* که

---

[۱۵] *Chromosome*

بوسیله‌ی فرانکلین و ویلکینز تهیّه شده‌بود و همچنین ساختمان مدل‌ها و استنباط‌های شخصی، مدل فضایی خود را ارائه دادند. در سال ۱۹۶۲ واتسون، کریک و ویلکینز به خاطر اهمیت کشف ساختمان *DNA* مشترکاً جایزه‌ی نوبل را دریافت کردند. مدل پیشنهادی آنان را در این شکل می‌بینیم.




شکل ۴.۲.۱.۱

سمت چپ: یک توالی *DNA* باز شده.
سمت راست: ساختار مکمّلی توالی *DNA*.

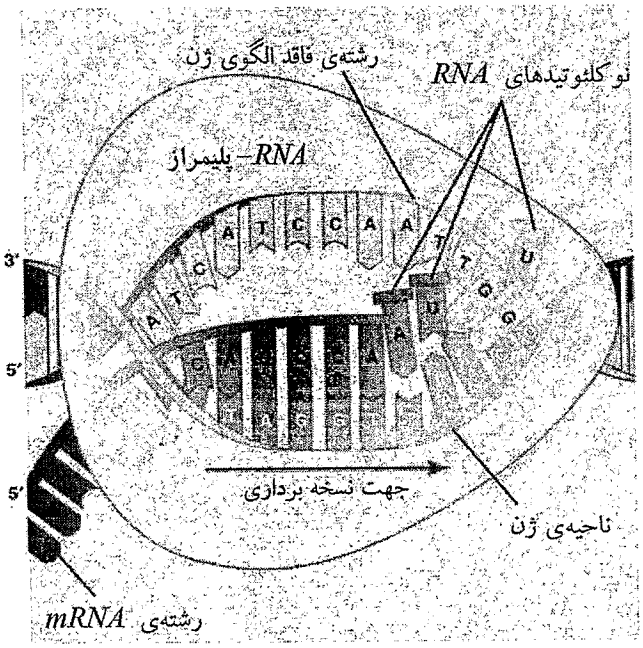
## ۳.۱.۱ *mRNA*، نسخه‌برداری و ترجمه

بروز خصوصیات وراثتی ذخیره‌شده در *DNA*، در گرو ترجمه‌ی یک ژن به رشته‌ای از پروتئین است. این کار به صورت زیر انجام می‌پذیرد:

$$DNA \xrightarrow{\text{نسخه‌برداری}} mRNA \xrightarrow{\text{ترجمه}} Protein$$

هر ژن روی یکی از دو رشته‌ی *DNA* قرار دارد. ابتدا این دو رشته در مکانی که ژن وجود دارد از هم جدا می‌شود و از روی توالی ژن، مکمّلِ آن ساخته می‌شود (شکل ۱.۳.۱.۱). به این تک‌رشته‌ی

مکمّلی $mRNA$[۱۶] ($RNA$ پیک) می‌گویند. ساختمان $mRNA$ شبیه به ساختمان $DNA$ است. $mRNA$ بازهای $A$، $G$ و $C$ را دارد امّا به جای باز $T$، بازآلی دیگری به نام اوراسیل[۱۷] ($U$) را به کار گرفته‌است. پس طبق قانون مکمّلی هر توالی ژن یک و تنها یک رشته‌ی $mRNA$ را می‌سازد که در آن به جای باز $T$، باز $U$ دیده می‌شود. ساخته‌شدن $mRNA$ از $DNA$ را نسخه‌برداری می‌نامند که توسّط آنزیمی به نام $RNA$–پلیمراز[۱۸] انجام می‌پذیرد.




در این مرحله، نوکلئوتیدها به ترتیب وارد $RNA$–پلیمراز می‌شوند و پس از تطبیق با باز مکمّل موجود در رشته‌ی $DNA$ به نوکلئوتید قبلی متّصل می‌شوند. در خاتمه، $RNA$–پلیمراز از رشته‌ی $DNA$ جدا می‌شود و تک رشته‌ای $mRNA$ که الگوی ساخت پروتئین می‌باشد، بدست می‌آید.

شکل ۱.۳.۱.۱

روی $DNA$ زیررشته‌هایی موجود است که حاوی اطّلاعاتی از جمله تنظیم بیان ژن، نقطه‌ی شروع ژن و تعداد نسخه برداری‌های لازم از روی آن می‌باشند[۳]. این قطعات را عناصر تنظیم کننده گویند. اگر در جهت $5' \rightarrow 3'$ روی $DNA$ حرکت کنیم، بالادست نقطه‌ی آغاز ژن یکی از عناصر تنظیم کننده به نام $Promoter$ وجود دارد(شکل ۲.۳.۱.۱). ناحیه‌ی $Promoter$ آنزیم $RNA$–پلیمراز را از نقطه‌ی آغاز نسخه‌برداری مطّلع می‌سازد.

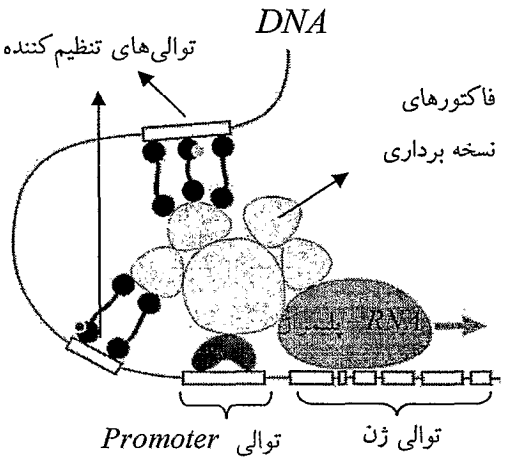
پس آن‌که فاکتورهای نسخه‌برداری زیررشته‌ای از $Promoter$ را برای اتّصال تشخیص می‌دهند، به $DNA$ می‌چسبند و نسخه‌برداری از ژن را کنترل می‌کنند. در واقع این فاکتورها $RNA$–پلیمراز را

---

[۱۶] *messenger RiboNucleic Acid*
[۱۷] *Uracil*
[۱۸] *RNA – Polymerase*

از موقعیت مکانی ژن مطّلع می‌سازند. مکانی از *Promoter* که فاکتورهای نسخه‌برداری به آن می‌چسبند معمولاً به طول پنج تا ۳۰ نوکلئوتید است که به آن مکان اتّصال فاکتور نسخه‌برداری[۱۹] گویند. هرچند که هر فاکتور نسخه‌برداری مکان مخصوصی برای اتّصال دارد امّا بعضی از این مکان‌ها به بعضی دیگر شبیه است [۳]. به این مکان‌های اتّصال شبیه به هم یک موتیف[۲۰] گویند(شکل ۳.۳.۱.۱). پس انگیزه‌ی یافتن موتیف‌ها روشن است: یافتن یک موتیف منجر به یافتن مکان‌های اتّصال و تجلّی یک ژن می‌شود و می‌توان بیان آن ژن را تحت کنترل خود درآورد.

بالادست ناحیه‌ی ژن توالی‌هایی به نام توالی‌های تنظیم‌کننده وجود دارد که حاوی اطّلاعاتی از جمله تنظیم بیان ژن، نقطه‌ی شروع ژن و تعداد نسخه برداری‌های لازم می‌باشند. یکی از این تنظیم‌کننده‌ها توالی *Promoter* نام دارد فاکتور نسخه برداری روی آن متّصل می‌گردد و نسخه برداری از ژن آغاز می‌شود.




شکل ۲.۳.۱.۱

[۱۹] *Transcription Factor Binding Site*
[۲۰] *Motif*


GTGCGGTAATAGCCACTCCCGAGGCCGTATACGCTCAATTGCGTAAGCGCAAGGTGAACGTCCTCGAC : *PRM* ۱
ACTTCAGCTCCAATTGCGTCATTCGTGCCCCGATGGTGCATCGCATGAAAGAATGCCCTCTCGCT : *PRM* ۲
GCTCTGATCACCACAGAACTAGCCGAGCCAATTAAGTCTCATTTCGTCACACTACACGAGGCTCAGG : *PRM* ۳
CAATTGCGATTGCGCCATCCACTCGCGCTGCATCGATCCGCCGCTTGAAACCGATGGGCGGGAATC : *PRM* ۴
GTGTCAAGCGGGTCCGATTGAGTTAGGTGTTTGCGCCACTCCCATGACATATGCACGTAAA : *PRM* ۵
GTTGTGCTTCTTCCCCATCATTGCCCAACTGCACAGGCCGAAGACAGGTGAGTTGGGCCTTGTGCTC : *PRM* ۶
GGTCCGGCTCAACAGTGTTAGATTACATCACTATGCTCTATCCAAACAGCGTTTAATCTTGTAAGTA : *PRM* ۷

شکل ۱.۱.۳.۳

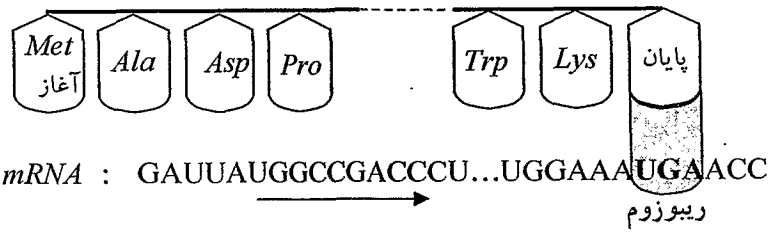
در شکل بالا هفت توالی *Promoter* می‌بینید که فاکتور نسخه برداری مشترکی به آن‌ها متّصل شده‌است. مکان‌های اتّصال این فاکتور نسخه برداری که به صورت پررنگ مشّخص شده، شبیه به هم می‌باشد. این مکان‌ها موتیف نام دارد.

به طور مجزّا هر سه نوکلئوتید متوالی را روی *mRNA* یک کدون[۲۱] می‌نامند. نقطه‌ی آغاز ترجمه‌ی یک *mRNA* به پروتئین، اوّلین کدون *AUG* است(شکل ۱.۱.۳.۴). یک ریبوزوم به این مکان می‌چسبد و شروع به خواندن کدون‌ها می‌کند[۴]. این ریبوزوم نسبت به هر کدون یک آمینواسید مخصوص را به زنجیر پروتئین در حال ساخت می‌چسباند و همین کار را برای کدون بعدی انجام می‌دهد. این کار وقتی خاتمه می‌پذیرد که به یکی از کدون‌های *UAA*، *UAG* و یا *UGA* برسد. در ابتدا به نظر می‌رسد ۴×۴×۴ یعنی ۶۴ آمینواسید وجود داشته باشد. امّا در تحقیقات آزمایشگاهی جدول ۱.۱.۳.۱ بدست آمده‌است که در آن تعدادی از کدون‌های متفاوت آمینواسید یکسانی را مشخّص می‌کنند. طبق این جدول ۲۰ نوع آمینواسید متفاوت موجود است. در نهایت ریبوزوم با استفاده از *mRNA* پروتئین را می‌سازد و ژن تجلّی می‌یابد.

| | U | C | A | G |
|---|---|---|---|---|
| U | UUU Phe | UCU Ser | UAU Tyr | UGU Cys |
| | UUC Phe | UCC Ser | UAC Tyr | UGC Cys |
| | UUA Leu | UCA Ser | UAA *Stop* | UGA *Stop* |
| | UUG Leu | UCG Ser | UAG *Stop* | UGG Trp |
| C | CUU Leu | CCU Pro | CAU His | CGU Arg |
| | CUC Leu | CCC Pro | CAC His | CGC Arg |
| | CUA Leu | CCA Pro | CAA Gln | CGA Arg |
| | CUG Leu | CCG Pro | CAG Gln | CGG Arg |
| A | AUU Ile | ACU Thr | AAU Asn | AGU Ser |
| | AUC Ile | ACC Thr | AAC Asn | AGC Ser |
| | AUA Ile | ACA Thr | AAA Lys | AGA Arg |
| | AUG Met | ACG Thr | AAG Lys | AGG Arg |
| G | GUU Val | GCU Ala | GAU Asp | GGU Gly |
| | GUC Val | GCC Ala | GAC Asp | GGC Gly |
| | GUA Val | GCA Ala | GAA Glu | GGA Gly |
| | GUG Val | GCG Ala | GAG Glu | GGG Gly |

جدول ۱.۱.۳.۱

[۲۱] *Codon*




شکل ۱.۱.۳.۴

ریبوزوم با شناسایی کدونِ *AUG* شروع به خواندن *mRNA* می‌کند و به هر سه نوکلئوتید متوالی یک اسیدآمینه را طبق جدول روبرو به پروتئین در حال ساخت می‌چسباند تا در خاتمه به یکی از کدون‌های *UAA*، *UAG* و یا *UGA* برسد.

## ۲.۱ راهبردهای اخیر

همان‌طور که در بخش ۳.۱.۱ توضیح داده‌شد، فاکتورهای نسخه‌برداری، بیان ژن را با اتّصال به ناحیه‌ی خاصّی از *Promoter* کنترل می‌کنند. مشخّص‌کردن مکان‌های اتّصال روی توالی‌های *DNA*، هنوز یک مسئله‌ی سخت در زیست‌شناسی مولکولی باقی مانده‌است. مهم‌ترین عامل سختی این مسئله آن است که فاکتورهای نسخه‌برداری به مکان‌های بسیار متنوّعی متّصل می‌شوند که پیش‌بینی توالی این مکان‌ها، از قبل بسیار مشکل می‌باشد. هرچند که مکان‌های اتّصال یک فاکتور نسخه‌برداری الگوی مشابهی دارند امّا این الگوها به طور دقیق مشخّص نیست. در نتیجه پیدا کردن آن کاری بسیار مشکل است.

یک راه برای فائق آمدن بر این مشکل استفاده از آزمایش‌های زیست‌شناسی است که اغلب پُرهزینه و زمان‌بَر[۲۲] است. دسترسی اخیر به توالی‌های کامل ژنوم، تلاش دانشمندان را به کشف طرز کار تنظیم ژن توسّط فاکتورهای نسخه‌برداری با استفاده از تحلیل محاسباتی برانگیخت. الگوریتم‌ها و ابزارهای جستجو برای عناصر تنظیم کننده به دو دسته‌ی عمده تقسیم می‌شوند:

[۲۲] *Time Consuming*

۱) روش‌هایی که مکان‌های اتّصال شناخته شده را جستجو می‌کنند. پیدا کردن مکان‌های جدید اتّصال در تعدادی توالی *Promoter* منجر به یافتن طرز کار ژن(های) جدید می‌شود. لذا این الگوریتم‌ها قادر به یافتن الگوی جدیدی برای مکان‌های اتّصال و در نتیجه ژن(های) جدید نیستند.

۲) روش‌هایی که به دنبال الگوهای جدیدی برای مکان اتّصال فاکتورهای نسخه‌برداری روی توالی‌های *Promoter* می‌گردند. این الگوریتم‌ها بیش از الگوریتم‌های دسته‌ی اوّل با سختی مسئله روبرو هستند. در فصل سه الگوریتم‌های موجود را دقیق‌تر بررسی می‌کنیم.

## ۳.۱ خط مشی ما

الگوریتم ما از لحاظ تقسیم بندی بخش ۲.۱ در دسته‌ی دوّم می‌گنجد. یعنی الگوریتم *Libra* به دنبال الگوهای جدیدی برای مکان اتّصال فاکتورهای نسخه‌برداری در مجموعه‌ای از توالی‌های *Promoter* می‌گردد که این توالی‌ها تنظیم‌کننده‌ی ژن مشترکی دارند. طول توالی‌های *Promoter* در بعضی موارد به چندهزار نوکلئوتید هم می‌رسد. لذا طول زیاد توالی‌ها نیز مسئله را سخت می‌کند (در فصل دو سختی مسئله‌ی یافتن موتیف‌ها را به طور دقیق‌تر بررسی خواهیم‌کرد). در نتیجه کم کردن فضای جستجو کمک شایانی برای رسیدن به جواب در زمان معقول‌تری خواهدبود. ما به چهار نوکلئوتید $A$، $C$، $G$ و $T$ به ترتیب چهار پایه‌ی مرتّب فضای برداری $\mathbb{R}^4$ نسبت می‌دهیم. پس به جای رشته‌ای از نوکلئوتیدها، دنباله‌ای از بردارها داریم. یک تابع پیش‌بینی تعریف می‌کنیم و با استفاده از هر چهار نوکلئوتید ، نوکلئوتید پنجم را با یک بردار تخمین می‌زنیم. بر اساس فاصله‌ی اقلیدسی، فاصله‌ی هر نوکلئوتید را تا بردار تخمین زده شده رتبه‌بندی می‌کنیم و رتبه‌ی نوکلئوتید پنجم را محاسبه می‌کنیم. این رتبه یکی از اعداد یک، دو، سه و یا چهار خواهدبود. سپس رتبه‌ی هر نوکلئوتید را با میانگین رتبه‌ی نوکلئوتید بعدی و قبلی آن تعویض می‌کنیم و این کار را ۱۵ بار انجام خواهیم‌داد تا نمودار هموارتری بدست‌آوریم. چون مکان‌های اتّصال فاکتورها توالی‌های شبیه به هم هستند انتظار داریم در نموداری که بدست‌آورده‌ایم به طور مشابهی پیش‌بینی شده‌باشد. یعنی نمودار در آن مکان‌ها شبیه به هم باشند. لذا با استفاده از نمودار توالی *DNA* را به قطعاتی تکّه‌تکّه می‌کنیم و تکّه‌های شبیه به هم را در یک مجموعه قرارخواهیم داد. در این صورت چهار مجموعه به ترتیب برای قطعاتی که در آن‌جا

نمودار نزولی، صعودی، مقعّر به بالا و مقعّر به پائین می‌باشد را بدست می‌آوریم. تا اینجا فضای جستجو را به تکّه‌های بدست‌آمده محدود کرده‌ایم. همردیفی تکّه‌های بدست‌آمده را بدست می‌آوریم و با استفاده از قطعاتی که نسبت به تعریف ما شبیه به هم هستند، الگویی برای موتیف بدست می‌آوریم. در نهایت مکان‌هایی را روی توالی *DNA* جستجو می‌کنیم که به الگو شبیه هستند. این مکان‌ها پیشنهادی برای موتیف خواهد بود. در فصل چهار الگوریتم *Libra* را به دقّت شرح خواهیم داد.

# ۲

# مسئله‌ی یافتن موتیف

## ۱.۲ پیش‌درآمد

زیست‌شناسان در آزمایشگاه توالی‌های مختلف *Promoter* را با فاکتورهای نسخه‌برداری متفاوتی مورد آزمایش قرار می‌دهند. سپس توالی‌های *Promoter* را بر حسب اینکه فاکتور یکسانی به آن‌ها متّصل شده‌باشد، دسته‌بندی می‌کنند. این توالی‌ها در واقع ژن یکسانی را شناسایی می‌کنند. به این مجموعه از *Promoter*ها توالی‌های هم‌خانواده[۱] می‌گوئیم. مشاهده نشان داده‌است که مکان اتّصال[۲] در توالی‌های هم‌خانواده به هم شبیه هستند. یعنی در هر جایگاه[۳] از توالی‌های هم‌خانواده، فراوانی یکی از چهار نوکلئوتیدِ $A$، $C$، $G$ و $T$ بیش از بقیه است. البتّه در بعضی جایگاه‌ها ممکن است بیش از یک نوکلئوتید از زیادترین فراوانی برخوردار باشند. می‌دانیم که مشاهدات تجربی هیچ تعریف جامع و مانعی برای شباهت ارائه نمی‌کند. با این حال به این مکان‌های شبیه به هم موتیف گویند. در واقع مشاهده شده‌است که مکان اتّصال فاکتورهای نسخه‌برداری نیز موتیف هستند.

---

[۱] *orthologous*
[۲] *Binding Site*
[۳] *Position*

## ۲.۲ صورت مسئله‌ی یافتن موتیف

فرض کنید $S_1, S_2, \ldots, S_n$ ، $n$ توالی $DNA$ هم‌خانواده باشند. می‌دانیم این توالی‌ها دارای جهت هستند(بخش ۲.۱.۱). همه‌ی $n$ توالی مذکور را در یک جهت پشت سر هم قرار می‌دهیم و رشته‌ی حاصل را $S$ می‌نامیم. در واقع رشته‌ی $S$ دنباله‌ای از همه‌ی نوکلئوتیدهای $S_1, S_2, \ldots, S_n$ می‌باشد. فرض کنید $s_i$ و $s_j$ به ترتیب دو زیررشته‌ی[۴] هم‌طول از توالی‌های $S_i$ و $S_j$ باشند(شکل ۱.۲.۲). فاصله‌ی همینگ دو زیر رشته‌ی $s_i$ و $s_j$ به صورت زیر تعریف می‌شود:

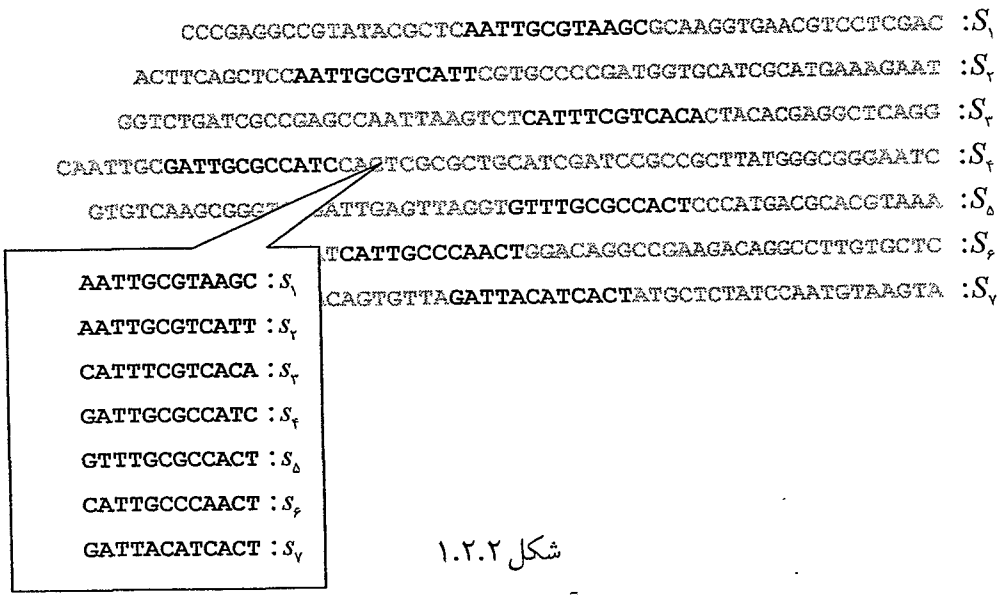
**فاصله‌ی همینگ[۵] دو زیر رشته‌ی $s_i$ و $s_j$ :** اگر $s_i(k)$ $k$امین نوکلئوتید زیررشته‌ی $s_i$ باشد به تعداد اعضای مجموعه‌ی $\{k \mid s_i(k) \neq s_j(k)\ ,\ k = 1,\ldots,l = |s_i|\}$ که در آن $|s_i|$ طول زیررشته‌ی $s_i$ می‌باشد، فاصله‌ی همینگ دو زیر رشته‌ی $s_i$ و $s_j$ گویند و آن را با $d_H(s_i, s_j)$ نشان می‌دهیم. خلاصه‌تر آنکه به تعداد جایگاه‌هایی که در آن‌ها دو زیررشته‌ی $s_i$ و $s_j$ نوکلئوتیدهای غیریکسان دارند فاصله‌ی همینگ دو زیررشته‌ی $s_i$ و $s_j$ گویند.

دو زیررشته‌ی $s_i$ و $s_j$ به طول $l$ را **شبیه به‌هم** گوئیم، اگر $d_H(s_i, s_j) \leq d$ باشد. مقدار $d$ به دلخواه تغییر می‌کند. به طور دقیق‌تر دو زیررشته‌ی $s_i$ و $s_j$ را **شبیه به‌هم با پارامتر** $(l,d)$ گوئیم. ساده‌ترین تعریف موتیف به صورت زیر است:

**موتیف [۵]:** فرض کنید رشته‌ی $S$ به دو زیرمجموعه‌ی جدا ازهم به نام $B$ و $M$ افراز شده است. اگر مجموعه‌ی $M$ شامل زیرقطعه‌هایی دوبه‌دو **شبیه به‌هم با پارامتر** $(l,d)$ باشد که شبیه آنها در مجموعه‌ی $B$ به **ندرت** یافت شود، مجموعه $M$ را **موتیف** می‌نامند. مجموعه‌ی $B$ پس‌زمینه[۶] نام دارد( شکل ۱.۲.۲).

---

[۴] *Subsequence*
[۵] *Hamming Distance*
[۶] *Background*




شکل ۱.۲.۲

مجموعه توالی‌های $S=\{S_1,...,S_7\}$ و مجموعه زیررشته‌های $s=\{s_1,...,s_7\}$ را می‌بینید که دو به دو به هم شبیه هستند. مجموعه‌ی $s$ یک موتیف است. نوکلئوتیدهای پس‌زمینه کم‌رنگ و عناصر موتیف پررنگ مشخّص شده‌اند.

کلّی‌ترین صورت مسئله‌ی یافتن موتیف، مطابق با ساده‌ترین تعریف موتیف می‌باشد:

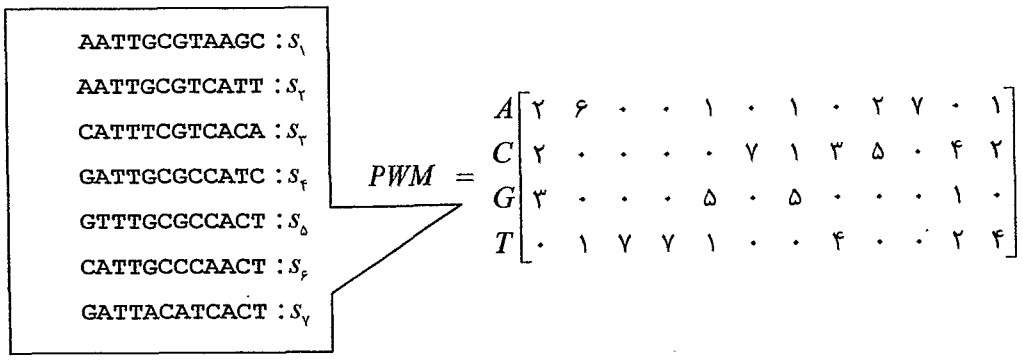
**مسئله‌ی یافتن موتیف ( در حالت کلّی) :** برای $n$ توالی هم‌خانواده‌ی $S_1, S_2,...,S_n$ که در یک جهت پشت سرهم قرارگرفته‌اند، دو مجموعه‌ی **پس‌زمینه** $(B)$ و **موتیف** $(M)$ را بیابید.

الگوریتم‌ها برای نمایش‌دادن یک موتیف آن را به صورت‌های گوناگونی توصیف می‌کنند. هر یک از این توصیفات را یک مدل برای موتیف گوئیم. تعاریف گوناگون موتیف و صورت مسئله‌های متفاوت برای آن زائیده‌ی همین مدل‌های گوناگون است. پس هر الگوریتمی به طور خاص به دنبال بهترین پیشنهاد برای موتیف طبق تعریف خود است. در فصل سه الگوریتم‌های متفاوت برای یافتن موتیف را بررسی خواهیم کرد.

## ۳.۲ مدل‌هایی برای توصیف موتیف

فرض کنید $M=\{s_1,s_2,...,s_k\}$ یک موتیف شامل $k$ زیررشته‌ی هم‌اندازه به طول $l$ باشد. برای راحتی کار، چهار نوکلئوتید $A$، $C$، $G$ و $T$ را به صورت چهارتایی مرتّب $N=(A,C,G,T)$ در نظر می‌گیریم. فرض کنید $k$ زیررشته‌ی هم‌طول $s_1,s_2,...,s_k$ را در ماتریس $C=\left[c_{ij}\right]_{k\times l}$ ذخیره کرده‌ایم به طوریکه $c_{ij}$ برابر با نوکلئوتید $j$ام در زیررشته‌ی $i$ام است. در این صورت موتیف $M$ را می‌توان به سه روش توصیف کرد [۶]:

۱) **مدل ماتریس وزن جایگاه $(PWM)$[۷]:** در این مدل موتیفِ $M$ به صورت یک ماتریس $4\times l$ به نام ماتریس وزن جایگاه توصیف می‌شود. اگر $PWM=\left[f_{ij}\right]_{4\times l}$ باشد آنگاه $f_{ij}$ برابر با تعداد نوکلئوتید $N(i)$ واقع در ستون $j$ام ماتریس $C$ است(شکل ۲.۳.۱). اگر $j$اُمین نوکلئوتید $s_k$ را با $s_k(j)$ نشان دهیم، آنگاه $s_k$ را شبیه به مدل $PWM$ گوئیم اگر $\Sigma_{s_k}^{PWM}=\sum_j PWM\left(s_k(j),j\right)$ به اندازه‌ی کافی بزرگ باشد. مقدار $\Sigma_{s_k}^{PWM}$ را امتیاز شباهت $s_k$ به مدل $PWM$ می‌نامیم.




شکل ۱.۳.۲

ماتریس $PWM$ فراوانی هر نوکلئوتید را در هر ستون نشان می‌دهد.

---

[۷] *Position Weight Matrix*