
بسم الله الرحمن الرحيم

دانشگاه بیرجند

دانشکده مهندسی

پایان‌نامه دوره کارشناسی ارشد مهندسی برق - الکترونیک

استفاده از روش‌های یادگیری ماشین برای تطبیق دنباله های بیولوژیکی

نگارش:

محمد سروری

استاد راهنما:

دکتر سید حمید ظهیری

استاد مشاور:

دکتر جواد صدری

زمستان ۱۳۹۰

تقدیم به خوبان خوب و یادگاران همیشه جاوید؛

پدرم، مادرم و معلمان دوران ابتدایی تا به امروزم

و تقدیم به فرهیخته استاد مهربان و دلسوزم؛

دکتر جواد صدری

به مصداق «من لم یشکر المخلوق لم یشکر الخالق» بسی شایسته است از اساتید فرهیخته و فرزانه جناب آقایان دکتر سید حمید ظهیری و دکتر جواد صدری که با کرامتی چون خورشید، سرزمین دل را روشنی بخشیدند و گلشن سرای علم و دانش را با راهنمایی های کارساز و سازنده بارور ساختند؛ تقدیر و تشکر نمایم.

ویزکیهم ویعلمهم الکتاب والحکمه.....

معالمقامت زعرش برتر باد همیشه توسن اندیشه ات مظفر باد به نکته های دلاویز و گفته های بلند صحیفه های سخن از تو علم پرور باد

همچنین از پدر و مادر عزیز، دلسوز و مهربانم که آرامش روحی و آسایش فکری مرا فراهم نمودند تا با حمایت های همه جانبه در محیطی مطلوب، مراتب تحصیلی و نیز پایان نامه درسی را به نحو احسن به اتمام برسانم؛ سپاسگزاری می نمایم.

شکر خدا که هر چه طلب کردم از خدا بر منتهای همت خود کامران شدم

چکیده

یکی از مهمترین موضوعات مورد بررسی در زیست‌شناسی و بیواینفورماتیک، مسئله مقایسه و تطبیق دنباله‌های زیستی است. مدل مخفی مارکوف یکی از الگوریتم‌های یادگیری ماشین می‌باشد که در زیست‌شناسی محاسباتی کاربرد فراوانی دارد. در این پایان‌نامه، ساختار مدل مخفی مارکوف و کاربرد آن در هم‌ترازی و خوشه‌بندی دنباله‌های زیستی مورد بررسی قرار گرفته است. از آنجایی‌که مدل مخفی مارکوف یک ابزار قوی برای یافتن شباهت بین داده‌های ترتیبی با طول‌های متغیر است، برای خوشه‌بندی دنباله‌ها، از آن استفاده شده است. در روش ارائه شده، هر دنباله زیستی بوسیله یک مدل مخفی مارکوف که با الگوریتم بهینه‌سازی گروه ذرات بهینه شده، مدل می‌شود. سپس با استفاده از پارامترهای بدست آمده، خوشه‌بندی بر مبنای یک تعریف جدید از ماتریس فاصله، انجام می‌شود. بهینه‌سازی پارامترهای مدل مخفی مارکوف با استفاده از الگوریتم جستجوی گرانشی نیز صورت گرفته است. در نهایت هم‌ترازی دنباله‌ها، با استفاده از روش خوشه‌بندی پیشنهادی و پروفایل مدل مخفی مارکوف، انجام می‌گیرد. آزمایش‌های انجام شده برروی دنباله‌های زیستی نظیر ژن‌ها نشان می‌دهد که الگوریتم پیشنهاد شده می‌تواند در انجام مقایسه بین دنباله‌های زیستی، بصورت بسیار کارامد و موثر، بکار گرفته شود.

کلید واژه‌ها: الگوریتم بهینه‌سازی گروه ذرات، الگوریتم جستجوی گرانشی، دنباله‌های زیستی، خوشه-بندی، ماتریس فاصله، مدل مخفی مارکوف ، هم‌ترازی دنباله.

فهرست مطالب

عنوان **صفحه**

فهرست علایم و نشانه‌ها

عنوان	علامت اختصاری
ماتریس احتمالات گذر	$\boldsymbol{A}$
ماتریس احتمالات نشر	$\boldsymbol{B}$
ماتریس احتمالات شروع	Π
مدل مخفی مارکوف	HMM
پارامتر نشان‌دهنده یک مدل مخفی مارکوف	$\lambda = (A, B, \Pi)$
الگوریتم بهینه‌سازی گروه ذرات	PSO
الگوریتم بهینه‌سازی جستجوی گرانشی	GSA
الگوریتم ژنتیک	GA

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل ۱– مقدمه

۱–۱– پیشگفتار

امروزه، بررسی و مطالعه ژینوم[۱] انسان یکی از موضوعات بسیار مهم و حیاتی در علم زیست‌شناسی می‌باشد. انجام تحقیقات آزمایشگاهی بر روی دنباله‌های زیستی نظیر DNA[۲] و پروتئین[۳]، باعث افزایش ضریب سلامت و ارائه راهکارهای جدید برای مقابله با بیماری‌ها می‌شود. در دهه‌های اخیر با پیدایش و توسعه الگوریتم‌های یادگیری ماشین، علم بیواینفورماتیک[۴] شکل گرفته و در کنار تحقیقات آزمایشگاهی زیست‌شناسان، به تجزیه وتحلیل ژینوم انسان پرداخته است. علم بیواینفورماتیک، کاربرد دانش محاسباتی رایانه و تکنولوژی اطلاعات در زمینه زیست‌شناسی و پزشکی می‌باشد که با الگوریتم‌ها، پایگاه داده‌ها، هوش مصنوعی و محاسبات نرم سروکار دارد. همچنین پردازش تصویر، مدل‌سازی و شبیه‌سازی، پردازش سیگنال، الگوریتم‌های تکاملی و شبکه‌های عصبی در بیواینفورماتیک کاربرد زیادی دارند.

از اهداف علم بیواینفورماتیک، می‌توان به یافتن ژن‌ها، هم‌ترازی دنباله‌ها، پیشگویی ساختار پروتئین‌ها، طراحی داروها، مشخص کردن عامل بیماری‌ها وغیره اشاره کرد. از آن‌جایی‌که حجم داده‌های دنباله‌های زیستی بسیار زیاد است و این حجم داده‌ها به مرور زمان در حال گسترش می‌باشد، بنابراین برای دست یابی به این اهداف، از الگوریتم‌های یادگیری ماشین، استخراج داده و تشخیص الگو استفاده زیادی می-شود.

[1] - Genome
[2] - Dioxide Nonleotide Acid(DNA)
[3] - Protein
[4] - Bioinformatic

۱-۲- تاریخچه

۱-۲-۱- تاریخچه پیدایش علم بیواینفورماتیک

واژه بیواینفورماتیک، قبل از ایجاد تحول بزرگ در زمینه ژنتیک ابداع و بکاربرده شد. در سال ۱۹۷۸ میلادی، *Paulien Hogeweg* و *Ben Hesper* این واژه را برای بیان و ارجاع به عبارت "مطالعه اطلاعات پردازشی در سیستم‌های حیاتی" استفاده کردند[1,2] . در آن زمان این تعریف بعنوان یک زمینه کاری در راستای علوم بیوفیزیک[۱] و بیوشیمی[۲] استفاده می‌شد. با این حال واژه بیواینفورماتیک با معنی امروزی برای اولین بار در سال ۱۹۸۰ برای تشریح کاربردهای علم رایانه و اطلاعات جهت تشریح داده‌های زیستی، مورد استفاده قرار گرفت. بدون شک بزرگترین تحول در زمینه ژنتیک و بیواینفورماتیک، با آغاز پروژه ژینوم انسان[۳] آغاز شد.

۱-۲-۲- پروژه ژینوم انسان

پروژه ژینوم انسان، یکی از پروژه‌های بین المللی علمی می‌باشد که با هدف اولیه ترتیب‌گذاری[۴] نوکلئوتایدهای[۵] دنباله *DNA* انسان، در سال ۱۹۹۰ میلادی شروع شد. سرپرست این پروژه،*Ari Patrinos* -رئیس اداره زیست‌شناسی و تحقیقات محیط زیستی دانشکده انرژی در کشور بریتانیا- بود. یک پیش نویس از این پژوهش در سال ۲۰۰۰ اعلام شد و نسخه کامل‌شده آن در سال ۲۰۰۳ منتشر شد. مهمترین نتایجی که این پروژه بین المللی بدست آورد به شرح زیر است:

- وجود تقریبا ۲۰۰۰۰ تا ۲۵۰۰۰ ژن در *DNA* انسان
- مشخص کردن ترتیب حدود سه بیلیون نوکلئوتاید در *DNA* انسان
- ذخیره سازی این اطلاعات در پایگاه داده‌ها

[1] - Biophysics
[2] - Biochemistry
[3] - Human Genome Project (HGP)
[4] - Sequencing
[5] - Nocleotide

- ارائه و بهبود ابزاری برای تجزیه وتحلیل داده‌ها

از هنگامی که پروژه ژینوم انسان به پایان رسید، تجزیه و تحلیل داده‌های بدست آمده از آن تاکنون ادامه دارد. این تجزیه و تحلیل‌ها شامل بررسی ساختار کروموزوم‌ها، ژن‌ها و پروتئین‌ها و ارائه یک الگوریتم محاسباتی دقیق جهت مدل سازی، خوشه‌بندی و هم‌ترازی[1] آن‌ها می‌باشد.

۱-۳- تجزیه وتحلیل دنباله های زیستی

قبل از شروع پروژه ژینوم انسان در سال ۱۹۹۰، نوکلئوتایدهای *DNA* اولین دنباله زیستی در سال ۱۹۷۷ ترتیب‌گذاری شد. این دنباله زیستی یک باکتری به نام *PhageQ-X174* بود [3]. بعد از آن و به دنبال شروع پروژه ژینوم انسان، دنباله‌های دیگری نیز ترتیب‌گذاری شده و اطلاعات آن در پایگاه‌های داده ذخیره‌سازی شدند. اطلاعات این دنباله‌ها برای مشخص کردن ژن‌هایی که پروتئین‌های خاصی را رمزگذاری کردند، *RNA*[2]ها و ساختار موتیف‌ها[3] و غیره مورد تجزیه و تحلیل قرار گرفت. انجام مقایسه بین ژن‌ها و خوشه‌بندی آن‌ها، باعث شناسایی توابع مشترک و رمز گذاری پروتئین‌های مشترکی شد که ژن‌های موجود در یک طبقه با یکدیگر به اشتراک می‌گذارند. با افزایش حجم زیاد دنباله‌های ترتیب-گذاری شده و داده‌های موجود در آن‌ها، تجزیه و تحلیل آن‌ها بصورت دستی تقریبا غیر ممکن شد. امروزه بوسیله الگوریتم‌های کامپیوتری نظیر بلاست[4]، دنباله‌های زیادی با شمار نوکلئوتایدهایی حدود ۱۹۰ بیلیون در مدت زمان بسیار کوتاهی مورد تجزیه و تحلیل قرار می‌گیرند[4].

۱-۴- هم ترازی و خوشه بندی

در بیواینفورماتیک، منظور از هم‌ترازی دنباله‌ها، ارائه یک روش برای چیدن و هم‌راستا کردن آن‌ها می باشد[5] و [6]. این دنباله‌ها می‌توانند *DNA*، پروتئین و یا *RNA* باشند. هدف از هم‌ترازی، پیدا کردن

[1] - Alignment
[2] - Ribo Nucleic Acid (RNA)
[3] - Motif
[4] - Basic Local Alignment Search Tool (BLAST)

نواحی شبیه به هم در دنباله‌ها می‌باشد که ممکن است یک ساختار تکاملی مشترک و یا یک ناحیه عملیاتی یکسانی داشته باشند. دنباله‌های هم‌تراز شده بصورت یک شکل ماتریس مانند نشان داده می-شوند که سطرهای این ماتریس دنباله‌های هم‌تراز شده و ستون‌های آن، حروف هم‌تراز شده در دنباله‌ها را نشان می‌دهند. همچنین واژه هم‌ترازی دنباله‌ها، برای داده‌های غیرزیستی نظیر داده‌های مالی و یا داده های هواشناسی نیز بکار می‌رود. شکل ۱-۱ قسمتی از هم‌ترازی چند دنباله از پروتئین‌ها را نشان می‌دهد.

```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
```

شکل ۱-۱: بخشی از هم‌ترازی تعدادی از دنباله های پروتئین

منظور از خوشه‌بندی مجموعه‌ای از داده‌ها، تقسیم‌بندی آن‌ها به چندین دسته و گروه می‌باشد. این تقسیم‌بندی باید بصورتی باشد که داده‌هایی که در یک گروه قرار می‌گیرند بیشترین میزان شباهت را به یکدیگر داشته باشند، در عین حالی که بین گروه‌ها بیشترین فاصله ممکن، ایجاد شود. در بیواینفورماتیک، خوشه‌بندی دنباله‌های زیستی نیز از اهمیت ویژه ای برخوردار است. چرا که بعنوان مثال با خوشه‌بندی ژن‌ها، می‌توان ژن‌هایی را که یک یا چند تابع مشخص به اشتراک می‌گذارند، در یک گروه قرار داد[7].

هم‌ترازی و خوشه‌بندی، رابطه معناداری با یکدیگر دارند. به این صورت که هر دو موضوع، به نحوی شباهت و همسانی بین دنباله‌ها را جستجو می‌کنند. از این رو گاهی از اوقات بوسیله هم‌ترازی، عملیات خوشه‌بندی انجام می شود[11-8] و گاهی هم از خوشه‌بندی برای انجام هم‌ترازی بین دنباله‌ها استفاده شده است[12]. از آن جایی که مسئله هم‌ترازی دنباله‌ها، بسیار پیچیده‌تر از خوشه‌بندی آن‌ها می‌باشد، در زمینه انجام هم‌ترازی بوسیله نتایج خوشه‌بندی کارهای بسیار کمتری انجام شده است.

۱–۵– مدل مخفی مارکوف[1]

یکی از مهمترین الگوریتم‌های یادگیری ماشین که در بیواینفورماتیک کاربرد فراوانی دارد، مدل مخفی مارکوف می‌باشد. این مدل برای اولین بار توسط *Baum* و دیگر مولفین در یک سری از ژورنال‌های آماری در اواخر دهه ۶۰ میلادی توصیف شد. اولین کاربرد *HMM* در کاربردهای تشخیص صدا بود که در اواسط دهه ۷۰ به آن پرداخته شد[16-13]. در نیمه دوم دهه ۸۰، استفاده از *HMM* در تجزیه و تحلیل دنباله-های زیستی مخصوصا *DNA* آغاز شد [17]. از آن زمان تاکنون *HMM* در تمامی زمینه‌های بیواینفورماتیک، مورداستفاده فراوانی قرار گرفته است [5].

۱–۶– هدف از انجام پایان نامه

در زمینه تجزیه و تحلیل دنباله‌های زیستی، بحث هم‌ترازی و خوشه‌بندی از دیدگاه زیست‌شناسی و همچنین محاسباتی بسیار حائز اهمیت است. از طرفی دیگر رشد روز افزون داده‌های زیستی و همچنین لزوم یافتن میزان شباهت[2] و یا میزان غیرهمسانی[3] بین داده‌های جدید و داده‌های قدیم ،ارائه الگوریتم-هایی جهت انجام دقیق و سریع عملیات هم‌ترازی و خوشه‌بندی را اجتناب‌ناپذیر می‌کند.

در این پایان‌نامه، ساختار دنباله‌های زیستی و اجرای مسئله هم‌ترازی بوسیله روش‌های متعارف نظیر برنامه‌نویسی پویا مورد بررسی قرار گرفته است. همچنین از آنجایی‌که *HMM* در بسیاری از مسائل مربوط به بیواینفورماتیک وارد شده، ساختار این مدل بطور کامل تشریح شده است. علاوه بر آن، استفاده از *HMM* در حل مسئله هم‌ترازی دنباله‌های زیستی و خوشه‌بندی آنها نیز مورد تجزیه و تحلیل قرار گرفته است.

[1] - Hidden Markov Model (HMM)
[2] - Similarity
[3] - Dissimilarity

۱–۷– نوآوری پایان نامه

نوآوری ارائه شده در پایان‌نامه، مربوط به انجام یک روش جدید خوشه‌بندی دنباله‌های زیستی و استفاده از آن برای انجام هم‌ترازی بین دنباله‌ها می‌باشد. سپس الگوریتم خوشه‌بندی پیشنهادی بر روی دنباله‌های ژن مرتبط با سرطان ریه و الگوریتم هم‌ترازی بر روی دنباله‌های محک *BALiBASE*[۱] آزمایش شده است.

در روش پیشنهادی، ابتدا هر کدام از دنباله‌ها بوسیله یک *HMM* که توسط الگوریتم بهینه‌سازی گروه ذرات[۲] بهینه شده است، مدل می‌شوند. سپس با استفاده از خاصیت مدل مارکوف، فاصله بین هر دو دنباله از یکدیگر بدست می‌آید و یک ماتریس فاصله D مشخص می‌شود. در ادامه با استفاده از ماتریس فاصله D، دنباله‌ها با روش خوشه‌بندی *DPAM*[۳] ، خوشه‌بندی می‌شوند. بهینه‌سازی پارامترهای *HMM* با استفاده از الگوریتم جستجوی گرانشی[۴] نیز انجام شده است. عملیات هم‌ترازی با استفاده از روش خوشه-بندی صورت می‌گیرد. بدین ترتیب که برای انجام هم‌ترازی، یک مدل پروفایل مخفی مارکوف[۵] برای بهترین خوشه بدست آمده از بهترین حالت خوشه‌بندی، مدل‌سازی می‌شود. سپس تمامی دنباله‌ها، بر اساس کیفیت خوشه‌ای که به آن تعلق گرفته‌اند، به‌ترتیب با *Profile HMM*، هم‌تراز می‌شوند. نتایج بدست‌آمده نشان می‌دهد که الگوریتم پیشنهادی دارای عملکرد مناسبی جهت انجام خوشه‌بندی و هم-ترازی می‌باشد.

۱–۸– ساختار پایان نامه

این پایان‌نامه، از شش فصل تشکیل شده است. در فصل دوم، پس از تعریف مفاهیم اولیه در مورد دنباله‌های زیستی،مفهوم هم‌ترازی و نحوه انجام آن توسط الگوریتم های متعارف مورد بررسی قرار می-

---

1- Benchmark ALignment dataBASE
2- Particle Swarm Optimization
3 - Descriptor Partition Around Medoid
4- Gravitational Search Algorithm(GSA)
5- Profile HMM

گیرد. در فصل سوم، ساختار HMM و محاسبات و نحوه آموزش آن مورد تجزیه و تحلیل قرار می گیرد . در فصل چهارم، مسئله هم‌ترازی بوسیله HMM، بررسی شده و نحوه شکل‌گیری یک مدل پروفایل مخفی مارکوف جهت انجام هم‌ترازی چندگانه تشریح می‌شود.

در فصل پنجم، مسئله خوشه‌بندی و ارتباط آن با هم‌ترازی مورد بررسی قرار می‌گیرد و الگوریتم‌های خوشه‌بندی داده‌های ترتیبی عنوان شده و استفاده از *HMM* در خوشه‌بندی آنها بررسی می‌شود. در ادامه این فصل، کارهای انجام شده قبلی در زمینه خوشه‌بندی داده‌ها بوسیله *HMM* مورد بررسی قرار می-گیرد. در قسمت آخر فصل پنجم، روش پیشنهادی جهت انجام خوشه‌بندی و هم‌ترازی مطرح شده و در فصل ششم، نتیجه‌گیری و پیشنهادات ارائه می‌گردد.

فصل ۲– مقدمه ای بر ساختار دنباله های زیستی و هم ترازی بین آنها

۲–۱– مقدمه

در این فصل بطور مختصر، ساختار برخی از دنباله‌های زیستی نظیر *DNA* و پروتئین بررسی می‌شود. در ادامه مفهوم هم‌ترازی و نحوه انجام آن بوسیله الگوریتم‌های متعارف ، بیان خواهد شد.

۲–۲– دنباله های زیستی

DNA –۱–۲–۲

ساختار ملکول *DNA* انسان، از دو رشته نردبانی شکل تشکیل شده که این دو رشته به صورت فنر مانند در هم پیچیده شده‌اند[18]. ماده تشکیل دهنده اضلاع این رشته نردبانی شکل، زنجیره شکر-فسفات می‌باشد که این زنجیره توسط چهار نوع ماده مختلف به هم متصل می‌شوند(شبیه پله های نردبان).این چهار ماده عبارتند از:

$$Guanine(G)\,,\,,\ Cyto\sin e(C)\,,\ Thy\min e(T)\ \ Adenine(A)$$

در هر پله رشته نردبانی شکل *DNA*، دو ماده مشخص می تواند قرار بگیرند. به این نحو که ماده نوع A و نوع T با هم و ماده نوع C و نوع G هم با می‌توانند ظاهر شوند[19]. به هر کدام از این مواد مشخص، یک نوکلئوتاید[1] یا یک جفت پایه[2] گفته می‌شود (شکل ۲-۱).

[1]- Nokleotide
[2]- Base pair