

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه بیرجند
دانشکده مهندسی

پایان نامه دوره کارشناسی ارشد مهندسی برق - الکترونیک

استفاده از روش های 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 فریخته استاد مهربان و دلسوزم؛

دکتر خواصدری

به مصداق «من لم یسکر المخلوق لم یسکر الخالق» بسی شایسته است از اساتید فریخته و فرزانه جناب آقایان دکتر سید حمید ظهیری و دکتر حواد صدری که با

گرامتی چون خورشید، سرزمین دل را روشنی بخشیدند و گلشن سراسی علم و دانش را با راهبانی های کار ساز و سازنده بارور ساختند؛ تقدیر و شکر نمایم.

و نیز کیم و یعلیمم الکتاب و الحکمہ.....

مقامت ز عرش بر تریباد همیشه توسن اندیشه ات مفر باد
به نکته های دلاویز و گفته های بلند، صحیفه های سخن از تو علم پرور باد

بچنین از پدر و مادر عزیز، دلسوز و مهربانم که آراش روحی و آسایش فکری مرا فراهم نمودند تا با حمایت های همه جانبه در محیطی مطلوب، مراتب تحصیلی و

نیربایان نامه درسی را به نحو احسن به اتمام برسانم؛ پاسگزاری می نمایم.

شکر خدا که هر چه طلب کردم از خدا بر تنهای بهمت 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 آنجایی که 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 بر روی دنباله‌های زیستی نظیر ژن‌ها نشان می‌دهد که الگوریتم پیشنهاد شده می‌تواند در انجام مقایسه بین دنباله‌های زیستی، بصورت بسیار کارآمد 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

صفحه	عنوان
ط	فهرست علایم و نشانه‌ها
ی	فهرست جدول‌ها
ک	فهرست شکل‌ها
۱	فصل ۱- مقدمه
۱-۱	۱-۱- پیشگفتار
۲-۱	۲-۱- تاریخچه
۲-۱-۱	۲-۱-۱- تاریخچه پیدایش علم بیوانفورماتیک
۲-۲-۱	۲-۲-۱- پروژه ژینوم انسان
۳-۱	۳-۱- تجزیه و تحلیل دنباله‌های زیستی
۴-۱	۴-۱- هم‌ترازی و خوشه‌بندی
۵-۱	۵-۱- مدل مخفی مارکوف
۶-۱	۶-۱- هدف از انجام پایان‌نامه
۷-۱	۷-۱- نوآوری پایان‌نامه
۸-۱	۸-۱- ساختار پایان‌نامه
۸	فصل ۲- مقدمه ای بر ساختار دنباله‌های زیستی و هم‌ترازی بین آنها
۱-۲	۱-۲- مقدمه
۲-۲	۲-۲- دنباله‌های زیستی
۱-۲-۲	۱-۲-۲- DNA
۲-۲-۲	۲-۲-۲- پروتئین
۱-۲-۲-۲	۱-۲-۲-۲- رونوشت برداری
۲-۲-۲-۲	۲-۲-۲-۲- ترجمه کردن
۳-۲	۳-۲- هم‌ترازی
۱-۳-۲	۱-۳-۲- هم‌ترازی دوگانه
۲-۳-۲	۲-۳-۲- هم‌ترازی چندگانه
۴-۲	۴-۲- هم‌ترازی چندگانه دنباله‌های زیستی
۱-۴-۲	۱-۴-۲- مدل امتیازدهی به هم‌ترازی
۲-۴-۲	۲-۴-۲- جریمه جای خالی

۱۷	مدل جریمه جای خالی خطی.....	۲-۴-۱-۱
۱۷	مدل جریمه جای خالی سببی.....	۲-۴-۲-۱
۱۸	الگوریتم‌های هم‌ترازی.....	۲-۵-۱
۱۸	برنامه‌نویسی پویا.....	۲-۵-۱-۱
۱۸	هم‌ترازی سراسری.....	۲-۵-۱-۱-۱
۲۲	هم‌ترازی محلی.....	۲-۵-۱-۲
۲۵	فصل ۳- مدل مخفی مارکوف.....	
۲۵	مقدمه.....	۳-۱-۱
۲۵	مدل مخفی مارکوف.....	۳-۲-۱
۲۶	مدل‌های مارکوف مرتبه اول.....	۳-۳-۱
۲۸	مدل‌های مخفی مارکوف مرتبه اول.....	۳-۴-۱
۲۹	محاسبات مدل مخفی مارکوف.....	۳-۴-۱-۱
۳۰	ارزیابی.....	۳-۴-۲
۳۳	الگوریتم پسرو.....	۳-۴-۲-۱
۳۷	رمزگشایی.....	۳-۴-۳
۴۰	یادگیری.....	۳-۴-۴
۴۰	الگوریتم پیش‌رو- پس‌رو.....	۳-۴-۴-۱
۴۲	استفاده از الگوریتم ژنتیک برای آموزش ساختار مدل مخفی مارکوف.....	۳-۵-۱
۴۵	عملگرهای ژنتیک برای <i>GA-HMM</i>	۳-۵-۱-۱
۴۷	الگوریتم <i>Baum-Welch</i> انتخابی.....	۳-۵-۲
۴۹	مقدار برازندگی.....	۳-۵-۳
۵۳	مدل‌سازی ناحیه کدگذاری شده <i>C.jejuni</i>	۳-۵-۴
۵۶	فصل ۴- استفاده از مدل مخفی مارکوف در هم‌ترازی.....	
۵۶	مقدمه.....	۴-۱-۱
۵۶	ساختار <i>HMM</i> در هم‌ترازی دنباله‌های زیستی.....	۴-۲-۱
۵۷	استفاده از <i>HMM</i> جهت انجام هم‌ترازی دوگانه.....	۴-۳-۱
۶۱	استفاده از مدل مخفی مارکوف جهت هم‌ترازی چندگانه.....	۴-۴-۱
۶۲	پروفایل‌های مدل مخفی مارکوف.....	۴-۴-۱-۱
۶۳	ساختار <i>Profile HMM</i>	۴-۴-۱-۱-۱
۶۵	امتیازدهی به یک هم‌ترازی چندگانه.....	۴-۵-۱
۶۵	امتیازدهی به روش <i>SP</i>	۴-۵-۱-۱
۶۶	پایگاه داده پی‌فام.....	۴-۶-۱
۶۶	نرم افزار <i>HMMer</i>	۴-۷-۱
۶۷	کارهای دیگر انجام شده برای حل مسئله هم‌ترازی.....	۴-۸-۱

فصل ۵ - خوشه‌بندی و هم‌ترازی.....	۶۹
۱-۵ - مقدمه.....	۶۹
۲-۵ - خوشه‌بندی داده‌های ترتیبی.....	۶۹
۳-۵ - خوشه‌بندی مبتنی بر شباهت.....	۷۲
۴-۵ - کارهای انجام شده قبلی در رابطه با خوشه‌بندی داده‌های ترتیبی بوسیله <i>HMM</i>	۷۲
۵-۵ - کارهای انجام شده در رابطه با انجام هم‌ترازی با استفاده از نتایج بدست آمده از خوشه‌بندی.....	۷۴
۶-۵ - خوشه‌بندی دنباله‌های زیستی با استفاده از مدل مخفی مارکوف بهینه‌شده با الگوریتم بهینه-سازی گروه ذرات.....	۷۵
۱-۶-۵ - الگوریتم بهینه‌سازی گروه ذرات.....	۷۵
۲-۶-۵ - روش پیشنهاد شده جهت انجام خوشه‌بندی و هم‌ترازی.....	۷۶
۳-۶-۵ - استفاده از <i>HMM</i> برای مدل‌سازی دنباله‌ها.....	۷۷
۴-۶-۵ - نحوه اعمال <i>PSO</i> جهت بهینه‌سازی <i>HMM</i>	۷۹
۱-۴-۶-۵ - تابع برازندگی.....	۸۰
۵-۶-۵ - ماتریس فاصله <i>D</i>	۸۱
۶-۶-۵ - الگوریتم خوشه‌بندی <i>DPAM</i>	۸۲
۷-۵ - نتایج آزمایش‌ها.....	۸۳
۸-۵ - بهینه‌سازی پارامترهای <i>HMM</i> توسط الگوریتم جستجوی گرانشی (<i>GSA</i>).....	۸۶
۹-۵ - اعمال یک هم‌ترازی چندگانه بر اساس نتایج بدست‌آمده از خوشه‌بندی.....	۹۰
فصل ۶ - نتیجه‌گیری و پیشنهادها.....	۹۳
۱-۶ - نتیجه‌گیری.....	۹۳
۲-۶ - پیشنهادها.....	۹۴
منابع و ماخذ.....	۹۵
واژه‌نامه فارسی به انگلیسی.....	۱۰۳
واژه‌نامه انگلیسی به فارسی.....	۱۰۴

فهرست علائم و نشانه‌ها

عنوان	علامت اختصاری
ماتریس احتمالات گذر	A
ماتریس احتمالات نشر	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

صفحه

عنوان

جدول ۱-۳: پارامترهای GA	۵۱
جدول ۱-۵: تعداد حالات خوشه‌بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس فاصله <i>D</i> و <i>HMM</i> های بهینه‌شده با <i>PSO</i>	۸۵
جدول ۲-۵: تعداد حالات خوشه بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس شباهت <i>Dij</i> ارائه شده در [99] و <i>HMM</i> های بهینه‌شده با <i>PSO</i>	۸۵
جدول ۳-۵: تعداد حالات خوشه‌بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس فاصله ارائه شده در [78] و <i>HMM</i> های بهینه‌شده با <i>PSO</i>	۸۶
جدول ۴-۵: تعداد حالات خوشه‌بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس فاصله <i>D</i> و <i>HMM</i> های بهینه‌شده با <i>GSA</i>	۸۹
جدول ۵-۵: تعداد حالات خوشه‌بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس شباهت <i>Dij</i> ارائه شده در [99] و <i>HMM</i> های بهینه‌شده با <i>GSA</i>	۸۹
جدول ۶-۵: تعداد حالات خوشه‌بندی و مقادیر <i>DBL</i> بدست‌آمده برای هر کدام با خوشه‌بندی بر اساس ماتریس فاصله ارائه شده در [78] و <i>HMM</i> های بهینه‌شده با <i>GSA</i>	۸۹
جدول ۷-۵: مقادیر <i>SP</i> بدست آمده با استفاده از پایگاه داده <i>BALiBASE</i>	۹۱

فهرست شکل‌ها

صفحه	عنوان
۴	شکل ۱-۱: بخشی از هم ترازی تعدادی از دنباله‌های پروتئین
۹	شکل ۱-۲: ساختار <i>DNA</i> (شکل سمت راست مدل واقعی و شکل سمت چپ مدل باز شده آن می باشد).
۱۱	شکل ۲-۲: مراحل ساخت پروتئین
۱۵	شکل ۳-۲: ماتریس جانشین <i>BLOSUM 50</i>
۱۶	شکل ۴-۲: یک ماتریس جانشین برای هم ترازی کروموزوم ها
۱۹	شکل ۵-۲: سمت چپ: هم ترازی نوع ۱، وسط: هم ترازی نوع ۲، سمت راست: هم ترازی نوع ۳
۲۰	شکل ۶-۲: انتخاب $F(i,j)$ از سه سلول همسایه
۲۱	شکل ۷-۲: ماتریس کامل برنامه‌نویسی پویا برای دو دنباله نمونه.
۲۴	شکل ۸-۲: ماتریس هم‌ترازی محلی بین دو دنباله
۲۶	شکل ۱-۳: یک مدل مارکوف مرتبه اول
۲۸	شکل ۲-۳: نمونه ای از یک مدل مخفی مارکوف با سه حالت مخفی
۳۳	شکل ۳-۳: محاسبه احتمالات بوسیله الگوریتم پیشرو می تواند مانند یک شبکه داربستی تصور شود.
۳۵	شکل ۴-۳: مثال مربوط به بررسی مدل ارزیابی. در این جا یک مدل مخفی مارکوف دارای چهار حالت مخفی و پنج حالت قابل مشاهده نشان داده شده است.
۳۶	شکل ۵-۳: یک <i>HMM</i> چپ به راست برای استفاده در کاربردهای تشخیص صدا
۳۸	شکل ۶-۳: محاسبات مربوط به الگوریتم رمزگشایی
۳۹	شکل ۷-۳: بررسی مسئله رمزگشایی برای مثال ۴
۴۴	شکل ۸-۳: الگوریتم <i>GA-HMM</i>
۴۶	شکل ۹-۳: چهار نوع عملگر جهش. (a)الحاق حالت(الحاق کردن یک حالت در دومین مکان)، (b)حذف حالت (حذف سومین حالت)، (c) حذف گذر و (d) الحاق گذر.
۴۷	شکل ۱۰-۳: فرآیند یک تقاطع. در ضمن تقاطع گذرهای خروجی جابجا می‌شوند.
۴۸	شکل ۱۱-۳: منفی لگاریتم احتمال بر حسب تعداد تکرار <i>Baum-Welch</i> برای یک <i>HMM</i>
۵۰	شکل ۱۲-۳: یک مدل برای پیشگویی ناحیه پیشین <i>C.jejuni</i> که در [47] ارائه شده است.
۵۰	شکل ۱۳-۳: مدل دستی ساخته شده برای ناحیه پیشین <i>C.jejuni</i> که در [47] استفاده شده است.
۵۲	شکل ۱۴-۳: نتایج یادگیری <i>GA-HMM</i> : (a) بهترین مقدار برازندگی را در هر تکرار نشان می دهد و (b) میانگین تعداد حالات <i>HMM</i> را در هر تکرار نشان می‌دهد.

- شکل ۳-۱۵: بعد از آموزش دنباله‌های *C.jejuni* ، *GA-HMM* یک مدل را برای سیگنال متناوب پیدا می‌کند. ۵۳
- شکل ۳-۱۶: ساختار *HMM* برای ناحیه کد گذاری شده *C.jejuni* ۵۴
- شکل ۳-۱۷: ساختار یک *HMM* جهت مدل سازی یک کودون ۵۴
- شکل ۳-۱۸: نتایج شبیه‌سازی نواحی کدگذاری شده *C.jejuni* ۵۵
- شکل ۴-۱: ساده ترین ساختار *HMM* ۵۶
- شکل ۴-۲: ساختار *HMM* با حذف حالات دلخواه ۵۷
- شکل ۴-۳: ساختار *HMM* با حذف حالات دلخواه توسط حالات بی‌صدا ۵۷
- شکل ۴-۴: ساختار یک *HMM* جهت انجام هم‌ترازی دوگانه ۵۸
- شکل ۴-۵: ساختار کامل یک *HMM* جهت انجام هم‌ترازی دوگانه ۵۹
- شکل ۴-۶: مدل سازی یک هم‌ترازی بوسیله *Pair HMM* ۶۰
- شکل ۴-۷: یک *Pair HMM* جهت انجام هم‌ترازی محلی ۶۱
- شکل ۴-۸: هم‌ترازی چند گانه ۷ دنباله از پروتئین‌ها ۶۲
- شکل ۴-۹: ساختار یک *Profile HMM* ۶۴
- شکل ۴-۱۰: یک *Profile HMM* برای انجام هم‌ترازی محلی ۶۴
- شکل ۵-۱: بلوک دیاگرام روش پیشنهادی جهت خوشه‌بندی ۷۷
- شکل ۵-۲: ساختار یک *HMM* برای مدل سازی ژن. احتمالات گذر با خطوط نقطه چین و احتمالات نشر با خطوط تیره نشان داده شده اند. ۷۸
- شکل ۵-۳: درصد مقادیر *DBL* بدست آمده در برابر تعداد خوشه‌ها برای خوشه‌بندی بوسیله *HMM* های بهینه‌شده با *PSO* ۸۶
- شکل ۵-۴: درصد مقادیر *DBL* بدست آمده در برابر تعداد خوشه‌ها برای خوشه‌بندی بوسیله *HMM* های بهینه‌شده با *GSA* ۹۰
- شکل ۵-۵: مقادیر جدول ۵-۷ بر حسب درصد ۹۱

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 و تحلیل 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 و غیره اشاره کرد. از آنجایی که حجم داده‌های دنباله‌های زیستی بسیار زیاد است و این حجم داده‌ها به مرور زمان در حال گسترش می‌باشد، بنابراین برای دست‌یابی به این اهداف، از الگوریتم‌های یادگیری ماشین، استخراج داده و تشخیص الگو استفاده زیادی می‌شود.

^۱ - Genome

^۲ - Dioxide Nonleotide Acid(DNA)

^۳ - Protein

^۴ - 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 نوکلئوتایدی‌های^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 پیش‌نویس از این پژوهش در سال ۲۰۰۰ اعلام شد و نسخه کامل شده آن در سال ۲۰۰۳ منتشر شد. مهمترین نتایجی که این پروژه بین المللی بدست آورد به شرح زیر است:

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

¹ - Biophysics

² - Biochemistry

³ - Human Genome Project (HGP)

⁴ - Sequencing

⁵ - Nucleotide

● ارائه و بهبود ابزاری برای تجزیه و تحلیل 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^۱ آن‌ها می‌باشد.

۱-۳- تجزیه و تحلیل دنباله‌های 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،^۲ *RNA*ها و ساختار موتیف‌ها^۳ 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 تقریباً غیر ممکن شد. امروزه بوسیله الگوریتم‌های کامپیوتری نظیر بلاست^۴، دنباله‌های زیادی با شمار نوکلئوتاید‌هایی حدود ۱۹۰ بیلیون در مدت زمان بسیار کوتاهی مورد تجزیه و تحلیل قرار می‌گیرند [4].

۱-۴- هم‌ترازی و خوشه بندی

در بیوانفورماتیک، منظور از هم‌ترازی دنباله‌ها، ارائه یک روش برای چیدن و هم‌راستا کردن آن‌ها می‌باشد [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 می‌شوند 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 داده‌های 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 دنباله‌های 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 بیوانفورماتیک، خوشه‌بندی دنباله‌های زیستی نیز از اهمیت ویژه‌ای 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 می‌شود [8-11] 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^۱

یکی از مهمترین الگوریتم‌های یادگیری ماشین که در بیوانفورماتیک 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 [13-16]. در نیمه دوم دهه ۸۰، استفاده از *HMM* در تجزیه و تحلیل دنباله‌های زیستی مخصوصاً *DNA* آغاز شد [17]. از آن زمان تاکنون *HMM* در تمامی زمینه‌های بیوانفورماتیک، 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^۲ و یا میزان غیرهمسانی^۳ بین داده‌های جدید و داده‌های قدیم، ارائه الگوریتم‌هایی 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 آنجایی که *HMM* در بسیاری از مسائل مربوط به بیوانفورماتیک 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.

^۱ - Hidden Markov Model (HMM)

^۲ - Similarity

^۳ - Dissimilarity

۷-۱- نوآوری پایان نامه

نوآوری ارائه شده در پایان نامه، 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^۱ BALiBASE 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 خوشه‌بندی 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، به ترتیب با *Profile HMM*، هم‌تراز می‌شوند. نتایج بدست‌آمده نشان می‌دهد که الگوریتم پیشنهادی دارای عملکرد مناسبی جهت انجام خوشه‌بندی و هم‌ترازی می‌باشد.

۸-۱- 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 الگوریتم‌های متعارف مورد بررسی قرار می‌-

^۱- Benchmark ALignment dataBASE

^۲- Particle Swarm Optimization

^۳- Descriptor Partition Around Medoid

^۴- Gravitational Search Algorithm(GSA)

^۵- 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

۲-۱-۲ مقدمه

در این فصل بطور مختصر، ساختار برخی از دنباله های زیستی نظیر *DNA* و پروتئین بررسی می شود. در ادامه مفهوم هم ترازی و نحوه انجام آن بوسیله الگوریتم های متعارف، بیان خواهد شد.

۲-۲-۲ دنباله های زیستی

DNA - ۱-۲-۲

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

Guanine(G) , *Cytosine(C)* , *Thymine(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 هم می توانند ظاهر شوند [19]. به هر کدام از این مواد مشخص، یک نوکلئوتاید^۱ یا یک جفت پایه^۲ گفته می شود (شکل ۲-۱).

^۱ - Nucleotide

^۲ - Base pair