



دانشگاه شهید چمران اهواز

دانشکده مهندسی

گروه مهندسی کامپیوتر

درجه کارشناسی ارشد گرایش هوش مصنوعی

داده کاوی در داده‌های ریز آرایه جهت تشخیص سرطان

نگارش:

فاطمه امین زاده

استاد راهنما:

جناب آقای دکتر علیرضا عصاره

استاد مشاور:

سرکار خانم دکتر بیتا شادکار

۱۳۸۸

به نام خداوند بخشنده مهربان



دانشکده مهندسی
گروه مهندسی کامپیوتر

درجه کارشناسی ارشد گرایش هوش مصنوعی

داده کاوی در داده‌های ریز آرایه جهت تشخیص سرطان

نگارش:

فاطمه امین زاده

استاد راهنما:

جناب آقای دکتر علیرضا عصاره

استاد مشاور:

سرکار خانم دکتر بیتا شادگار

۱۳۸۸

نظر هیئت داوران

استاد راهنما : جناب آقای دکتر عصاره

تاریخ و امضا

استاد داور : جناب آقای

تاریخ و امضا

تعهد نامه

بدینوسیله اینجانب فاطمه امینزاده قصرالدشتی تعهد می‌نمایم که مطالب ارائه شده در این پایان نامه حاصل کار پژوهشی و تحقیق اینجانب می‌باشد و قبلاً برای احراز هیچ مدرک دیگری ارائه نشده است. رجوع به دست آوردهای دیگران که در این پایان نامه از آنها استفاده شده مطابق مقررات ارجاع داده شده است. کلیه حقوق مادی و معنوی این اثر متعلق به دانشکده مهندسی دانشگاه شهید چمران اهواز می‌باشد.

نام و نام خانوادگی دانشجو: فاطمه امینزاده

امضاء:

تاریخ:

صفحه تقديم

صفحه تقدیر و تشکر

چکیده

تکنولوژی ریزآرایه باعث تولید حجم انبوهی داده دسته‌بندی در بسیاری از زمینه‌ها شده است. تحلیل داده‌های ریزآرایه و کلاس‌بندی آن‌ها، نشان می‌دهد این روش در تشخیص بیماری‌ها و سرطان تأثیر بسزایی دارد. با توجه به تحقیقات بسیاری که در مورد کلاس‌بندی داده‌های ریزآرایه صورت گرفته است، اعمال روش‌های معمول یادگیری ماشین دارای معایبی ذاتی برای رسیدن به یک کلاس‌بندی پایدار و دقیق است. بنابراین مطلوب‌تر است که از ترکیب دسته‌بندی‌کننده‌های خبره، به جای تکیه بر نتیجه‌ی یک دسته‌بندی‌کننده منفرد استفاده شود. در این تحقیق، دسته‌بندی داده‌های ریزآرایه‌ای با استفاده از روش‌های موثر انتخاب ژن‌ها و دسته‌بندی‌کننده‌ی ترکیبی RotBoost انجام شده است. این دسته‌بندی‌کننده از ترکیب تکنیک‌های AdaBoost و Rotation Forest به‌وجود می‌آید. با توجه به تحقیقات قبلی انجام شده در زمینه‌ی ریزآرایه‌ها، این نخستین باری است که تکنیک RotBoost بر روی دسته‌بندی‌کننده‌ی داده‌های ریزآرایه اعمال گردیده است. این تحقیق بر روی ۸ مجموعه‌داده‌ی ریزآرایه واقعی، پیاده‌سازی شده و از درخت تصمیم به عنوان دسته‌بندی‌کننده‌ی پایه استفاده شده است. بررسی نتایج به‌دست آمده نشان می‌دهد که در اکثر این ریزآرایه‌ها، روش پیشنهادی از صحت بالاتری نسبت به تکنیک‌هایی نظیر AdaBoost و Rotation Forest برخوردار است.

واژه‌های کلیدی: داده‌های ریزآرایه DNA، دسته‌بندی‌کننده ترکیبی، دسته بندی سرطان، راپر، فیلتر.

صفحه	فهرست مطالب
۴	فهرست اشکال
۶	فهرست جداول
	فصل اول
۷	۱- مفاهیم مقدماتی
۸	۱-۱- بیولوژی سرطان
۹	۲-۱- بیولوژی مولکولی
۱۱	۱-۲-۱- DNA، عامل انتقال اطلاعات وراثتی
۱۲	۲-۲-۱- اسیدهای نوکلئیک
۱۴	۳-۲-۱- همانند سازی
۱۶	۴-۲-۱- اصل مرکزی، مسیری یک طرفه
۱۷	۵-۲-۱- RNA
۱۹	۳-۱- ریزآرایه
۱۹	۱-۳-۱- بیان ژنی و فناوری ریزآرایه
۲۰	۲-۳-۱- انواع ریزآرایه‌ها
۲۳	۱-۳-۱- استخراج داده‌ها از تصاویر ریزآرای‌های
۲۴	۴-۱- ساختار پایان‌نامه
	فصل دوم
۲۵	۲- پیشینه تحقیق
۲۶	۱-۲- انتخاب ژن
۲۸	۲-۲- خوشه‌بندی ژن‌ها
۳۱	۳-۲- دسته‌بندی ژن‌ها
	فصل سوم
۳۶	۳- مبانی و روش تحقیق
۳۶	۱-۳- مقدمه
۳۷	۲-۳- مشخصات مجموعه داده‌ها
۳۸	۳-۳- انتخاب ویژگی
۳۸	۱-۳-۳- گزینش روش انتخاب ویژگی بر اساس هدف:
۳۹	۲-۳-۳- انتخاب ویژگی براساس نوع خروجی مورد انتظار:
۴۰	۳-۳-۳- انتخاب ویژگی براساس استراتژی جستجو

۴۱	۳-۳-۴ گزینش روش انتخاب ویژگی بر اساس خصوصیات مجموعه داده:
۴۳	۳-۴-۴ روش انتخاب ویژگی مناسب این تحقیق
۴۴	۳-۵-۴ اعتبارسنجی داخلی و خارجی
۴۵	۳-۶-۴ بسته نرم افزاری انتخاب ژن
۴۶	۳-۷-۴ الگوریتمهای جست و جو و ارزیابی به کار رفته در این تحقیق
۴۶	۳-۷-۱ روشهای جست و جو
۴۷	۳-۷-۲ مدل فیلتر
۴۷	۳-۷-۲-۱ تکنیک FCBF:
۵۴	۳-۷-۲-۲ تکنیک Relief:
۵۵	۳-۷-۲-۳ تکنیک F-Statistic:
۵۵	۳-۷-۲-۴ تکنیک GSNR:
۵۶	۳-۷-۲-۵ تکنیک mRMR :
۵۶	۳-۷-۲-۶ مدل رایبر
۵۷	۳-۸-۴ کلاس بندی داده های ریز آرایه
۵۷	۳-۸-۱ درخت تصمیم
۵۹	۳-۸-۲ دسته بندی کننده های ترکیبی
۵۹	۳-۸-۲-۱ تکنیک Bagging
۶۰	۳-۸-۲-۲ تکنیک AdaBoost
۶۳	۳-۸-۳ تکنیک Rotation Forest
۶۶	۳-۸-۴ تکنیک RotBoost
۶۸	۳-۸-۴-۱ ماتریس چرخش و نگاشت مجموعه داده ها
۷۱	۳-۹-۴ چگونگی ارزیابی کارایی دسته بندی کننده ها
	فصل چهارم
۷۴	۴-بررسی و تحلیل نتایج
۷۴	۴-۱ مقدمه
۷۴	۴-۲ نتایج مرحله انتخاب ویژگی
۷۸	۴-۳ نتایج مرحله دسته بندی
۷۹	۴-۳-۱ استفاده از روش ارزیابی Bootstrap .632
۷۹	۴-۳-۲ مقایسه ی کارایی دسته بندی کننده ی ترکیبی RotBoost با دسته بندی کننده ی منفرد
۸۱	۴-۳-۳ مقایسه ی کارایی دسته بندی کننده ی ترکیبی RotBoost با دسته بندی کننده های ترکیبی معمول
۸۱	۴-۳-۱ مقایسه ی میانگین و انحراف معیار
۸۳	۴-۳-۲ ماتریس چرخش مبتنی بر PCA و ICA
۸۵	۴-۳-۳ مقایسه ی صحت و تمایز با استفاده از دیاگرام Kappa

۹۹	۴-۳-۳-۴- بررسی تأثیر تعداد ژن‌های انتخاب شده در دسته‌بندی‌کننده RotBoost
۱۰۲	۴-۳-۳-۵- بررسی تأثیر تغییر پارامتر تعداد ویژگی‌های موجود در هر زیرمجموعه از ویژگی‌ها (M)
۱۰۳	۴-۳-۳-۶- مقایسه‌ی کارایی روش پیشنهادی با روش‌های قبلی

فصل پنجم

۱۰۵	۵- نتیجه‌گیری و پیشنهادات آتی
۱۰۵	۵-۱- نتایج
۱۰۷	۵-۲- پیشنهادات آتی
۱۰۹	فهرست مراجع:

فهرست اشکال

فصل اول

- شکل ۱-۱ مراحل پیشرفت سرطان ۱۰
- شکل ۲-۱ نمودار رشد داده‌ها در پایگاه Gen Bank در سالهای اخیر ۱۷
- شکل ۳-۱ مارپیچ DNA درون یک موجود زنده؛ رشته‌ی DNA ۱۹
- شکل ۴-۱ چهار باز آلی نیتروژن دار حلقوی تشکیل دهنده نوکلئوتیدها ۲۰
- شکل ۵-۱ الف نحوه اتصال و جفت شدن بازها ۲۰
- شکل ۶-۱ همانندسازی DNA ۲۲
- شکل ۷-۱ اصل مرکزی بیولوژی مولکولی ۲۴
- شکل ۸-۱ سنتز RNA از روی DNA ۲۵
- شکل ۹-۱ مراحل تهیه ریزآرایه‌های الیگونوکلئوتید ۲۸
- شکل ۱۰-۱ مراحل تهیه ریزآرایه‌های cDNA ۲۹
- شکل ۱۱-۱ ریزآرایه‌های یک کاناله و دو کاناله ۳۰

فصل سوم

- شکل ۱-۳ یک الگوریتم ترکیبی از فیلتر و راپر ۶۱
- شکل ۲-۳ الگوریتم FCBF ۶۹
- شکل ۳-۳ الگوریتم AdaBoost ۷۱
- شکل ۴-۳ الگوریتم Rotation Forest ۷۴
- شکل ۵-۳ الگوریتم RotBoost ۷۷
- شکل ۶-۳ ICA ماتریس داده‌ها را به ضرب دوماتریس S و A تبدیل می‌کند ۸۳

فصل چهارم

- شکل ۱-۴ میانگین صحت دسته‌بندی مجموعه داده‌های ریزآرایه با استفاده از ژن‌های انتخاب شده توسط فیلترهای گوناگون ۸۷
- شکل ۲-۴ میانگین صحت دسته‌بندی مجموعه داده‌های ریزآرایه با استفاده از ژن‌های انتخاب شده توسط فیلترهای گوناگون ۹۰
- شکل ۳-۴ میانگین صحت دسته‌بندی با استفاده از دسته‌بندی کننده ترکیبی RotBoost و منفرد درخت تصمیم ۹۱
- شکل ۴-۴ میانگین صحت دسته‌بندی کننده‌های منفرد و ترکیبی مختلف ۹۳
- شکل ۵-۴ میانگین صحت دست‌بندی کننده‌ی RotBoost با استفاده از تبدیلات PCA و ICA ۹۵
- شکل ۶-۴ دیاگرام ابر Kappa به ازای مجموعه داده‌ی SRBCT ۹۷
- شکل ۷-۴ مرکز دیاگرام Kappa به ازای مجموعه داده‌های مختلف، به ترتیب (a) Breast، (b) CNS، (c) Colon. ۹۸

- Leukemia (d)
- شکل ۴-۸ مرکز دیاگرام Kappa به ازای مجموعه داده‌های مختلف، به ترتیب، Lung(e)، MLL(f)، Ovarian(g) و ۹۹ SRBCT(h).
- شکل ۴-۹ دیاگرام Kappa-error مربوط به Breast و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۰ RotBoost و
- شکل ۴-۱۰ دیاگرام Kappa-error مربوط به CNS و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۱ RotBoost
- شکل ۴-۱۱ دیاگرام Kappa-error مربوط به Colon و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۲ RotBoost و
- شکل ۴-۱۲ دیاگرام Kappa-error مربوط به Leukemia و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۳ RotBoost و Forest
- شکل ۴-۱۳ دیاگرام Kappa-error مربوط به Lung و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۴ RotBoost
- شکل ۴-۱۴ دیاگرام Kappa-error مربوط به MLL و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۵ RotBoost
- شکل ۴-۱۵ دیاگرام Kappa-error مربوط به Ovarian و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۶ RotBoost و Forest
- شکل ۴-۱۶ دیاگرام Kappa-error مربوط به SRBCT و روش‌های ترکیبی Bagging، AdaBoost، Rotation Forest و ۱۰۷ RotBoost و Forest
- شکل ۴-۱۷ نمودارهای میانگین صحت دسته‌بندی کننده RotBoost بر اساس تعداد ژن‌های متفاوت مشخص شده توسط الگوریتم Relief
- شکل ۴-۱۸ نمودارهای میانگین صحت دسته‌بندی کننده RotBoost بر اساس تعداد ژن‌های متفاوت مشخص شده توسط الگوریتم GSNR
- شکل ۴-۱۹ نمودارهای میانگین صحت RotBoost با توجه به تغییر تعداد ژن‌ها با استفاده از فیلتر F-Statistic
- شکل ۴-۲۰ نمودارهای میانگین و انحراف معیار صحت RotBoost با توجه به تغییر پارامتر M

فهرست جداول

فصل اول

جدول ۱-۳- مشخصات هر یک از مجموعه داده‌ها شامل تعداد نمونه‌ها، تعداد کلاس‌ها و تعداد ویژگی‌ها یا ژن‌ها ۴۴

۸۱

فصل چهارم

جدول ۱-۴- میانگین صحت دسته‌بندی مجموعه داده‌های ریزآرایه با استفاده از ژن‌های انتخاب شده توسط فیلترهای گوناگون ۸۲

جدول ۲-۴- تعداد ژن‌های انتخاب شده توسط فیلترهای مختل ۸۴

جدول ۳-۴- میانگین صحت دسته‌بندی مجموعه داده‌های ریزآرایه با استفاده از ژن‌های انتخاب شده توسط راپر ۸۶

جدول ۴-۴- میانگین و انحراف معیار صحت دسته‌بندی کننده‌ی ترکیبی RotBoost و منفرد درخت تصمیم ۸۸

جدول ۵-۴- میانگین و انحراف معیار صحت دسته‌بندی کننده‌های منفرد و ترکیبی مختلف ۹۰

جدول ۶-۴- میانگین و انحراف معیار RotBoost با استفاده از تبدیلات PCA و ICA ۱۱۰

جدول ۷-۴- تغییرات میانگین و انحراف معیار صحت روش RotBoost با پارامتر M ۱۱۱

فصل اول

مقدمه

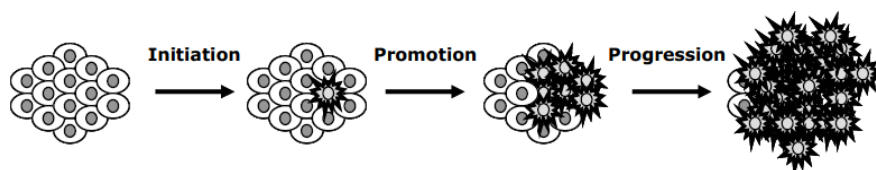
۱- مفاهیم مقدماتی

سالیان متمادی است که بشر می‌کوشد و علت وجودی آن‌ها را دریابد. بیماری‌های جسمی و روانی از جمله مسائل مهم انسانی هستند، که توجه بسیاری از متخصصین را در هر عصر و زمانی به خود معطوف

نموده‌اند. از کوشش برای ارتباط دادن بسیاری از بیماری‌ها به نیروهای ماوراءالطبیعه در سالیان دور تا کشف پنی سیلین و در سال‌های اخیر ظهور تکنولوژی مولکولی، جملگی تلاش‌هایی است که بشر در این راستا انجام داده است.

۱-۱- بیولوژی سرطان

سرطان چهارمین بیماری شایع و یکی از علل بسیار مهم مرگ‌ومیر در سراسر جهان است که هر ساله منجر به بیش از ۶ میلیون مورد مرگ می‌شود [۱]. سرطان بار مالی سنگینی به سیستم‌های پزشکی و ضایعات بسیار مهم‌تری به بیماران و خانواده‌های آن‌ها وارد می‌کند. سرطان توسط یک سری عوامل که باعث تغییر شکل سلول‌های نرمال به غیرنرمال و در نتیجه رشد و تقسیم غیرقابل کنترل آن‌ها می‌گردد، ایجاد می‌شود. همان‌گونه که شکل ۱-۱ نشان می‌دهد، این فرآیند شامل سه مرحله آغاز^۱، پیشرفت^۲ و پیشرفت تصاعدی^۳ است.



شکل ۱-۱ مراحل پیشرفت سرطان

در مرحله‌ی اول، عواملی مانند ویروس، مشکلات محیطی (آلودگی هوا، مصرف دخانیات، تشعشعات و غیره) باعث جهش^۴ DNA می‌شود. مرحله‌ی پیشرفت شامل پروسه‌ی رشد و ازدیاد سلول‌های جهش‌یافته‌ای است که منجر به تومورهای کوچک می‌گردد. در مرحله‌ی پیشرفت تصاعدی این تومورها دچار رشد بیشتر و در نتیجه تخریب بافت اصلی می‌شوند [۱].

علی‌رغم پیشرفت‌های تکنولوژی طی یک دهه‌ی گذشته در زمینه‌ی سرطان موفقیت‌های محدودی برای مقابله با این بیماری ویران‌گر حاصل شده است. بنابراین تشخیص دقیق، دسته‌بندی و پیش‌بینی به‌موقع آن، از اهمیت به‌سزایی

¹ Initiation

² Promotion

³ Progression

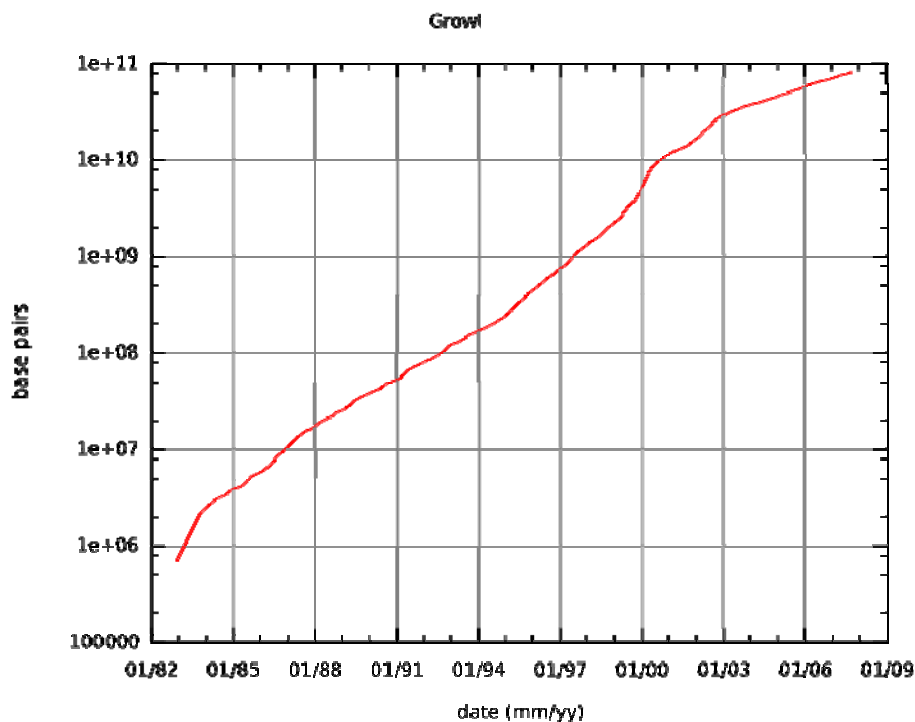
⁴ Deoxyribo Nucleic Acid

در مطالعات و تحقیقات برخوردار است. لذا به کارگیری مناسب تکنولوژی جدید به منظور پیش‌بینی و دسته‌بندی سرطان از اهمیت روزافزونی برخوردار شده است.

۱-۲- بیولوژی مولکولی

بیولوژی مولکولی به سرعت به سوی نظام‌های مبتنی بر داده در حرکت است و در نتیجه بر نقش ابزارها و الگوریتم‌های محاسباتی نیز روزبه‌روز افزوده شده تا علاوه بر بالا رفتن سرعت، با کاهش تعداد آزمایش‌ها برای یافتن پاسخ مسأله، در هزینه‌های آزمایشگاهی صرفه‌جویی شود. به تدریج روش‌های هوشمند با توجه به فضای جست‌وجوی بسیار بزرگ در برخی مسائل، به کار گرفته شده و امروزه شاهد استفاده از این الگوریتم‌ها در اغلب مسائل از جمله تشخیص و درمان بیماری‌ها به‌ویژه سرطان می‌باشیم که این کاربردها، مدام در حال افزایش است [۴].

بیولوژی مولکولی محاسباتی از شاخه‌های نسبتاً جدیدی است که با وجود نو بودن، توجه مشتاقان زیادی را به خود معطوف نموده است. اگرچه درک مکانیزم‌های حیات، از عرصه‌هایی است که همه‌روزه مرزهای آن فراتر می‌رود، اما با توجه به تاثیر عمیق محاسبات در تمامی شئون زندگی امروز، کنترل فرآیندهای حیات در سطح مولکولی، هنوز پتانسیل بالایی جهت پیشرفت دارد. همان‌گونه که اشاره شد، بیولوژی مولکولی به‌سوی نظام‌های مبتنی بر داده در حرکت است و در نتیجه نقش ابزارها و الگوریتم‌های محاسباتی هر روزه پررنگ‌تر می‌شود. بیش از ده سال پیش، دانشمندان زیستی شروع به جمع‌آوری داده‌های بیولوژیک در سطح وسیعی نمودند و هم‌اکنون پروژه‌های توالی‌یابی ژنی و پروتئینی، میزان قابل توجهی از این داده‌ها را تشکیل می‌دهد. به‌عنوان مثال تا انتهای سال ۲۰۰۴، بیش از ۴۴ میلیارد نوکلئوتید در ۴۰ میلیون رکورد از توالی‌های ژنی در GenBank لیست شده است و این میزان تقریباً هر ۱۴ ماه دو برابر می‌شود [۲]. شکل ۱-۲ نشان‌دهنده‌ی این موضوع است.



شکل ۱-۲ نمودار رشد داده‌ها در پایگاه Gen Bank در سال‌های اخیر رشد بسیار بالا بوده و به‌طور متوسط هر ۱۴ ماه دو برابر شده است.

تحلیل دقیق این داده‌های حجیم و در حال افزایش، می‌تواند به درک صحیح فعالیت‌های زیستی در سطح مولکولی بیانجامد. سیستم‌های کامپیوتری با فراهم نمودن امکان ذخیره‌سازی و به اشتراک گذاشتن داده‌ها و سود جستن از سرعت بالا در محاسبات، نقش کلیدی را در این تحقیقات ایفا می‌کند. اگر چه با پیشرفت‌های اخیر در تکنولوژی، سرعت و ظرفیت این سیستم‌ها به‌طور عمده‌ای افزایش یافته است، لیکن تحلیل و ارائه این حجم از داده‌های زیستی و توانایی پیش‌بینی در مورد داده‌های آینده، ابزارهای قدرتمند و پیچیده‌تری را طلب می‌کند. در اینجا الگوریتم‌های پیشنهادی هوش مصنوعی به مدد روش‌های سنتی آمده و راه‌حلی مؤثر برای مسائلی ارائه نموده که حل برخی از آن‌ها پیش از این (چه به‌دلیل کمبود دانش و چه به‌دلیل هزینه‌های بالای زمانی و حافظه‌ای در محاسبات) غیر عملی می‌نمود [۴].

۱-۲-۱- DNA، عامل انتقال اطلاعات وراثتی

قدم اول در بیولوژی مولکولی، درک صحیح فرآیند توارث است. بشر همیشه بر این باور بوده که در هنگام تولید-مثل، خصوصیات ظاهری از والد یا والدین به مولود منتقل می‌شود. حتی بسیاری از پیشینیان عقیده داشتند که فرزند، علاوه بر خصوصیات جسمی، خصوصیات روحی را نیز به ارث می‌برد [۲].

سلول کوچک‌ترین واحد زنده یک موجود زنده است و کلیه موجودات، طی فرآیند میتوز^۱، از تقسیم سلولی یک سلول به وجود می‌آیند. بدن متشکل از سلول‌های متفاوت با کارکردهای مختلف است. سلول، از غشاء سیتوپلاسمی، اندامک‌ها و هسته تشکیل شده است. کروموزوم‌ها درون هسته‌ی سلول قرار دارند که حاوی اطلاعات ژنتیکی (ژنوم یا مجموعه ژن‌ها) و متشکل از دو DNA می‌باشند. ژن‌ها تعیین‌کننده خصوصیات فیزیکی موجودات هستند به طوری که تغییرات کوچک در ژن‌ها باعث ایجاد تغییرات کوچک در ویژگی‌های فیزیکی می‌شود و محل قرار گرفتن آن‌ها در رشته DNA، loci یا مکان هندسی^۲ نامیده می‌شود. به مجموعه کلیه اطلاعات ژنتیکی موجود در ژنوتیپ یک موجود، ژنوم گفته می‌شود [۲].

در قرون معاصر، Mendel اولین شخصی بود که پدیده‌ی فرآیند توارث را به شیوه جدید مورد بررسی قرار داد. وی در سال ۱۸۶۵ یک مدل ریاضی جهت ارث‌بری ارائه داد و در آن، واحد اساسی توارث را ژن نام نهاد. کار مندل بعدها تقریباً به فراموشی سپرده شد تا اینکه در اوایل قرن بیستم با توجه به رشد سریع ریاضیات، دوباره بررسی‌هایی بر روی آن انجام شد. با تمام این‌ها، چپستی ژن تا سال ۱۹۴۴ که دانشمندان پی‌بردند ژن از ماده‌ای به نام DNA تشکیل شده است، هنوز برهمه پوشیده بود. در سال ۱۹۵۳ واتسون و کریک، ساختار مارپیچ دورشته‌ای معروف DNA را پیشنهاد دادند. این ساختار از آن جهت مورد توجه قرار گرفت که توانسته بود یک مدل فیزیکی برای بیان چگونگی تقسیم یک مولکول DNA به دو مولکول مشابه ارائه دهد. امروزه مشخص شده است که مولکول DNA، پلیمری از مولکول‌های کوچک‌تری به نام نوکلئوتیدها^۳ است که اطلاعات مربوط به توارث، ناشی از ترتیب قرار گرفتن آن‌ها در کنار یکدیگر می‌باشد (شکل ۱-۳). مارپیچ DNA درون یک سلول یک موجود زنده؛ این رشته پس از تا خوردن فراوان به صورت کروموزوم درون هسته سلول قرار می‌گیرد [۵].

¹ Mitoses

² Locus

³ Nucleotide