



دانشگاه تهران
مرکز تحقیقات بیوشیمی و بیوفیزیک

مسائل محاسباتی در شناسائی هاپلوتیپ‌ها

نگارش
سید علی کتان‌فروش

اساتید راهنما: دکتر مهدی صادقی و دکتر حمید پزشکی

استاد مشاور: دکتر چنگیز اصلاحچی

رساله برای دریافت درجه
دکترای در رشته بیوانفورماتیک

مهر ۱۳۸۸

مسائل محاسباتی در شناسائی هاپلوتیپ‌ها

چکیده

توالی نوکلئوتیدها در ژنوم انسان بجز برخی از جایگاه‌های معین بر روی آن، در بین همه افراد یکسان است. با این حال همین تفاوت‌های اندک، عامل اصلی بروز صفات متفاوت در بین افراد جمعیت است. رایج‌ترین نوع این تفاوت‌ها، چندریختی تک نوکلئوتیدی یا اسنپ (Single Nucleotide Polymorphism, SNP) است. در هر جایگاه اسنپ، بیش از یک نوع نوکلئوتید بر روی ژنوم افراد مختلف جمعیت مشاهده می‌شود. یک هاپلوتیپ، نشان‌دهنده‌ی توالی نوکلئوتیدهای ژنوم یک فرد بر روی تعدادی از اسنپ‌ها است. ژنوم هر فرد ترکیبی از دو هاپلوتیپ به ارث رسیده از والدین است که این ترکیب را اصطلاحاً ژنوتیپ فرد می‌نامند. به نظر می‌رسد تنوع هاپلوتیپ‌ها طی نسل‌های متوالی، در نواحی معینی از ژنوم بدون تغییر باقی می‌ماند. این نواحی، ژنوم را به مجموعه‌ای از بلوک‌های هاپلوتیپی افراز می‌کنند.

در رساله‌ی پیش رو، پس از مرور روش‌های توالی‌یابی ژنوم و تعیین ژنوتیپ‌ها، دو مسئله‌ی محاسباتی در شناسائی هاپلوتیپ‌ها مورد بحث قرار می‌گیرند؛ مسئله‌ی استنباط هاپلوتیپ‌ها از داده‌های ژنوتیپ، تحت مدل بیشترین پارسیمونی و مسئله‌ی تعیین بلوک‌های هاپلوتیپ. در مسئله‌ی اول، ترکیبی از یک الگوریتم ژنتیک و رده‌ای از روال‌های سودجویانه برای حل مسئله‌ی بیشترین پارسیمونی بکار گرفته می‌شوند. در مسئله‌ی دوم، برای تعیین ساختار بلوکی ژنوم به وسیله‌ی نمونه‌ای از هاپلوتیپ‌های جمعیت، ابتدا یک شاخص برای تعیین همبستگی بین اسنپ‌ها تعریف می‌شود سپس، بلوک‌های هاپلوتیپ با حل یک مسئله‌ی بهینه‌سازی مقید بدست

می‌آیند. بر این اساس، بلوک‌های هاپلوتیپ به قسمی تعریف می‌شوند که بیشترین تعداد جفت اسنیپ‌های “همبسته” در ژنوم را در برگیرند و تعداد جفت اسنیپ‌های “مستقل” درون بلوک‌ها از کسر معینی تجاوز نکند. در این روش، از آزمون دقیق فیشر برای تعیین سطح معناداری همبستگی بین اسنیپ‌ها و برنامه‌ریزی پویا برای بدست آوردن افراز بهینه استفاده می‌شود.

در این رساله چند طرح جدید برای ارزیابی جنبه‌های مختلف ساختارهای بلوکی معرفی می‌شوند که از طریق آنها، شباهت بین ساختارهای بلوکی، ثبات بلوک‌های هاپلوتیپ و کارایی افرازهای بلوکی در شناسائی نقاط پراحتمال نوترکیبی و نیز شناسائی جایگاه ژنی مرتبط با بیماری مورد بررسی قرار می‌گیرند. برخلاف نتایج نه چندان رضایت‌بخش بدست آمده از الگوریتم پیشنهادی برای استنباط هاپلوتیپ‌ها، روش پیشنهادی برای افراز بلوکی هاپلوتیپ‌ها در بیشتر جنبه‌های مورد بررسی، برتری مطلوبی نسبت به دیگر روش‌های رایج نشان می‌دهد.

واژه‌های کلیدی: ژنوتیپ، هاپلوتیپ، اسنیپ *SNP*، ژنتیک جمعیت، بازتوالی‌یابی ژنوم، آزمون همبستگی

فیشر، برنامه‌ریزی خطی، برنامه‌ریزی پویا، ساختار بلوکی کروموزوم، نرخ نوترکیبی، مدل‌های بیماری، شناسائی

جایگاه بیماری، مطالعه‌ی *case-control*.

پیشگفتار

در قرن جدید، دانش زیست‌شناسی بیشترین سهم در تحقیقات را متوجه خود می‌کند. زیست‌شناسی در مقایسه با دیگر علوم پایه مثل فیزیک و شیمی، قرن‌ها از داشتن مدل‌های نظری به قدر کافی دقیق و در عین حال جامع بی‌بهره بود و از این رو بیشتر به توده‌ای از واژگان می‌مانست که تنها برای نامگذاری پدیده‌های مرتبط با حیات ابداع می‌شدند در حالیکه در تبیین روابط علّی بین آنها نوسیدکننده می‌نمود. تجربیات مندل در زمینه‌ی وراثت و نظریه‌ی تکاملی داروین تا پیش از قرن بیستم زمینه‌ی اولیه‌ی بنیان‌های تئوری پیدایش و بقای جانداران را فراهم کردند. کشف ساختار مارپیچ دو رشته‌ای DNA توسط واتسون و کریک در نیمه‌ی قرن گذشته تقریباً تمام آنچه را زیست‌شناسی به عنوان اصول موضوعه نیاز داشت فراهم کرد و مفهوم ذهنی ژن به عنوان عامل وراثت را در شکل ماده‌ای شیمیایی با ساختاری کاملاً پیچیده عینیت بخشید. اکنون، در کنار این پارادایم جدید، گردآیه‌ای از دیگر علوم پایه، ریاضیات و آمار، و علوم مهندسی قرار گرفته‌اند تا درک ما از دنیای حیات را با سرعت بیشتری گسترش بخشند. از این میان، همزمان با دوران آغازین پروژه‌ی ژنوم و نیاز به ابزارهای کارآمد برای سازماندهی و جستجوی اطلاعات مرتبط با توالی‌های زیستی، رشته‌ی جدیدی تحت عنوان بیوانفورماتیک ظهور کرد. اکنون، بیوانفورماتیک با معرفی و توسعه‌ی ابزارهایی از علوم کامپیوتر برای حل مسائل زیست‌شناسی و به طور خاص زیست‌شناسی سلولی و مولکولی و ژنتیک جایگاه ویژه‌ای میان زمینه‌های تحقیقاتی در دنیای علم دارد.

رساله‌ی حاضر حاوی بخشهایی از مجموعه فعالیت‌های پژوهشی نگارنده در طول دوره‌ی دکترای تخصصی بیوانفورماتیک در دانشگاه تهران از ابتدای سال تحصیلی ۱۳۸۳ است که در ارتباط با موضوع «مسائل محاسباتی در شناسائی هاپلوتیپ‌ها» به عنوان پایان‌نامه‌ی دکترای به انجام رسیده‌اند. در این رساله، دو مسئله‌ی متفاوت مورد بحث و بررسی قرار می‌گیرند: مسئله‌ی استنباط هاپلوتیپ‌ها از داده‌های ژنوتیپ توسط یک الگوریتم ژنتیک و مسئله‌ی تعریف یک ساختار بلوکی بر روی ژنوم افراد یک زیرجمعیت انسانی. هر یک از این مسائل با خاستگاه‌های متفاوتی به عنوان موضوع تحقیق مطرح می‌شدند و از اینرو پیشرفت یکسانی نداشته‌اند. ایده‌ی بکارگیری الگوریتم ژنتیک در مسئله‌ی استنباط هاپلوتیپ‌ها به عنوان اولین موضوع تحقیق، از کارآیی امیدوارکننده‌ای در مقایسه با دیگر روش‌های رایج برخوردار نبود. با این حال مطالعه‌ی بخش‌های مرتبط با شیوه‌ی پیاده‌سازی الگوریتم ژنتیک برای حل مسئله‌ی استنباط هاپلوتیپ‌ها و مرور دیگر رویکردهای حل این مسئله در این رساله می‌تواند اطلاعات مفیدی از جزئیات و حدود کارآیی چنین روش‌هایی در اختیار کسانی که به این مبحث علاقمندند قرار دهد. مسئله‌ی دیگر یعنی تعریف ساختار بلوکی ژنوم در زیرجمعیت‌های انسانی

حجم اصلی این رساله را در بر می گیرد. کارآیی روش پیشنهاد شده برای این مسئله در این رساله به تفصیل با بکارگیری در دو مسئله دیگر مورد بررسی قرار می گیرد: مسئله ای از ژنتیک آماری درباره ی تعیین نقاط پراحتمال نوترکیبی در ژنوم و دیگری مسئله ی شناسائی جایگاه ژنی مرتبط با بیماری در نمونه های case و control.

به عنوان موضوعی رایج در بیوانفورماتیک، در این رساله نیز خواننده با طیف پراکنده ای از ابزارها از آمار و علوم کامپیوتر تا ژنتیک و تکامل که برای حل مسائل طرح شده بکار گرفته شده اند مواجه می شود. هرچند بیشتر این رهیافتها، مثل برنامه ریزی خطی برای حل مسئله ی افراز بلوکی هاپلوتیپها، از پایه مورد بحث قرار نگرفته اند اما برای خواننده ی آشنا به چنین زمینه هایی می تواند قابل توجه باشد. مطالب در این رساله، مطابق با چارچوب استاندارد در علوم زیستی، به ترتیب در فصل های مقدمه - مواد و روشها - نتایج و بحث و نتیجه گیری قرار گرفته اند. بدین ترتیب لازم است، خواننده برای مطالعه ی هر یک از سه مسئله ی مورد بحث، بخشهای مجزا در هر یک از فصل های رساله را به طور جداگانه تعقیب کند.

قدردانی

این رساله حاصل پشتیبانی‌های فکری و روحی افراد بی‌شماری است که نگارنده بر خود لازم می‌داند در اینجا بخشی از آنها را یاد آورد شود. اصولاً، بخش عمده‌ای از پژوهشگران بیوانفورماتیک در ایران، آشنایی با این رشته را مرهون تلاش‌های جناب دکتر بهرام گلیائی مدیر گروه بیوانفورماتیک دانشگاه تهران و جمعی از اساتید بین‌رشته‌ای می‌دانند که هسته‌ی اولیه‌ی این گروه را تشکیل می‌دادند. به ویژه، تلاش‌های جناب دکتر مهدی صادقی در معرفی مسائل گوناگون بیوانفورماتیک و ایجاد ارتباط علمی بین اساتید رشته‌های مختلف، شایسته‌ی قدردانی است. همچنین است راهنمایی‌های جناب دکتر حمید پزشک در مسائل آماری که جزء همیشگی مسائل بیوانفورماتیک هستند. به علاوه لازم است از اساتید محترم جناب دکتر اصلاحچی (دانشگاه شهید بهشتی) و سرکار خانم دکتر الهی به خاطر راهنمایی‌هایشان که به بهبود راه حل‌ها و پرداخت نتایج کمک کرد تشکر کنم. راهنمایی‌های جناب دکتر آرمین مددکار در مورد شیوه‌ی ارائه‌ی مطالب و همکاری بسیار صمیمانه‌ی آقای سید امیر مرعشی در نگارش مقاله نیز درخور قدردانی فراوانند. در اینجا لازم است به طور جداگانه از پژوهشگاه دانش‌های بنیادی و قطب زیست-ریاضی دانشگاه تهران که بخش‌هایی از پژوهش مرتبط با این رساله را مورد حمایت مالی قرار دادند تشکر شود. بسیاری دیگر از اساتید و دانشجویان مرکز تحقیقات بیوشیمی و بیوفیزیک و دانشکده‌ی علوم دانشگاه تهران، پژوهشگاه دانشهای بنیادی و نیز مرکز تحقیقات ژنتیک و زیست‌شناسی مولکولی ماکس پلانک (برلین) هستند که نگارنده را در طی دوره‌ی پژوهشی از دیدگاه‌ها و ایده‌های پرفایده‌یشان در ارتباط با موضوع این تحقیق بهره‌مند ساخته‌اند. در انتها باید از تشویق‌های امیدبخش همسرم برای ادامه این دوره و تکمیل رساله تشکر کنم.

فهرست مطالب

یک	پیشگفتار
چهار	فهرست مطالب
شش	لیست جدول‌ها
هفت	لیست شکل‌ها
۱	۱ مقدمه
۱	۱۰۱ آشنایی با واژگان ژنتیک
۶	۲۰۱ فنآوری‌های تعیین ژنوتیپ
۱۶	۳۰۱ استنباط هاپلوتیپ‌ها با استفاده از داده‌های ژنوتیپ
۳۳	۴۰۱ بلوک‌های هاپلوتیپ
۴۹	۵۰۱ شناسایی جایگاه ژنی خصیصه
۵۷	۲ مواد و روشها
۵۷	۱۰۲ یک الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها
۷۳	۲۰۲ تولید نمونه‌های تصادفی هاپلوتیپ تحت مدل فیلوژنی کامل
۷۹	۳۰۲ یک شاخص برای تعیین وجود همبستگی بین اسنیپ‌ها
۹۰	۴۰۲ روش GPMAP برای افراز بلوکی هاپلوتیپ‌ها
۹۵	۵۰۲ مقایسه‌ی الگوریتم‌های افراز بلوکی هاپلوتیپ‌ها
۹۸	۱۰۵۰۲ نمونه‌گیری از داده‌های HapMap
۹۹	۲۰۵۰۲ تنوع هاپلوتیپ‌ها
۱۰۰	۳۰۵۰۲ محاسبه‌ی تگ‌اسنیپ‌ها
۱۰۳	۴۰۵۰۲ کمیتی برای اندازه‌گیری شباهت بین ساختارهای بلوکی

۵۰۵۰۲	شیوه‌ای برای ارزیابی ثبات الگوریتم‌های افراز بلوکی	۱۰۴
۶۰۵۰۲	سنجش توان شناسائی نقاط پراحتمال نوترکیبی	۱۰۵
۷۰۵۰۲	سنجش توان شناسائی جایگاه ژنی یک خصیصه	۱۰۶

۳ نتایج و بحث ۱۱۸

۱۰۳	کارآیی الگوریتم ژنتیک در استنباط هاپلوتیپ‌ها	۱۱۸
۲۰۳	نمونه‌های از افرازهای بلوکی در ناحیه‌های ENCODE	۱۲۶
۳۰۳	تنوع هاپلوتیپ‌ها در بلوک‌ها	۱۳۰
۴۰۳	تعداد و پوشش htSNP‌ها در بلوک‌ها	۱۳۲
۵۰۳	شباهت بلوک‌های هاپلوتیپی در بین روشهای متفاوت	۱۳۶
۶۰۳	مقایسه‌ی ثبات مدل‌های متفاوت در تعریف بلوک‌های هاپلوتیپی	۱۳۸
۷۰۳	توان شناسائی نقاط پراحتمال نوترکیبی	۱۳۹
۸۰۳	توان شناسائی جایگاه ژنی یک خصیصه	۱۴۱
۱۰۸۰۳	تاثیر نحوه‌ی انتخاب نشانگذارها بر توان	۱۴۳

۴ نتیجه‌گیری ۱۴۷

۱۰۴	الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها	۱۴۷
۲۰۴	الگوریتم GPMAP برای تعیین بلوک‌های هاپلوتیپی	۱۴۸
۳۰۴	تحقیقات آتی	۱۵۱

پیوست الف ۱۵۴

مراجع ۱۵۸

لیست جدول‌ها

۱۰۱	برخی مدل‌های رایج در توارث بیماری‌ها بر حسب ریسک نسبی ژنوتیپ‌ها	۵۳
۱۰۲	روش‌های افراز بلوکی هاپلوتیپ‌های مورد ارزیابی در این رساله	۹۷
۲۰۲	اطلاعات کلی ژنوتیپ‌های HapMap در نواحی ENCODE	۹۹
۱۰۳	بهترین تنظیمات برای پارامترهای الگوریتم ژنتیک GAhap	۱۲۰
۲۰۳	خطای استنباط و تعداد هاپلوتیپ‌های متمایز در نتایج بدست آمده از اجرای الگوریتم‌های مختلف بر روی ژنوتیپ‌های مجموعه‌ی هورن	۱۲۶
۳۰۳	طول زمان اجرا در روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها	۱۲۷
۴۰۳	مشخصات بلوک‌های هاپلوتیپی در نواحی ENCODE به ازای روش‌های مختلف	۱۳۱
۵۰۳	تعداد htSNPsها برای هر یک از نواحی ENCODE به ازای افرازهای بلوکی مختلف	۱۳۳
۶۰۳	شباهت افرازها بین روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها	۱۳۷
۷۰۳	مقایسه‌ی ثبات در تعریف بلوک‌های هاپلوتیپ در بین روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها	۱۳۸
۸۰۳	خطای روش‌های افراز بلوکی هاپلوتیپ‌ها در تشخیص نقاط پراحتمال نوترکیبی	۱۴۰
۹۰۳	خطای نوع اول در شناسائی جایگاه ژنی بیماری بین روش‌های مختلف	۱۴۲
۱۰۰۳	p -مقدار آستانه‌ای آزمون همبستگی مربع کای برای بدست آوردن خطای نوع اول ثابت	
	بین روش‌های مختلف در مطالعه‌ی همبستگی	۱۴۳

لیست شکل‌ها

۱۰۱	ساختمان DNA	۲
۲۰۱	آلل‌های متفاوت پنج اسنپ در قطعه‌ی یکسانی از پنج نمونه‌ی متفاوت از یک کروموزوم	۵
۳۰۱	فناوری نسل جدید توالی‌یابی رشته‌های نوکلئوتیدی	۸
۴۰۱	سازوکار شناسائی توالی نوکلئوتیدی در تراشه‌های SNP-microarray	۱۰
۵۰۱	فرایند آماده‌سازی و خواندن توالی‌های نوکلئوتیدی در فناوری Illumina/Solexa	۱۳
۶۰۱	نگاشت قطعات کوتاه توالی‌یابی شده بر روی ژنوم مرجع و شناسائی جایگاه‌های اسنپ	۱۴
۷۰۱	نمونه‌هایی از ژنوتیپ‌ها و هاپلوتیپ‌های تشکیل‌دهنده‌ی هر یک از آنها	۱۸
۸۰۱	درخت فیلوژنی کامل	۲۳
۹۰۱	بازسازی هاپلوتیپ‌های فردی به وسیله‌ی برهمگذاری قطعات توالی‌یابی شده	۳۲
۱۰۰۱	بلوک‌های هاپلوتیپ در ناحیه‌ی 5q31	۳۴
۱۱۰۱	آزمون چهار گامی برای تعیین بلوک‌های هاپلوتیپ	۴۲
۱۲۰۱	انتخاب هاپلوتیپ-تگ اسنپ‌ها	۴۸
۱۳۰۱	شناسائی هاپلوتیپ مرتبط با بیماری در بین نمونه‌های case و control	۵۱
۱۰۲	نمایش جواب مسئله‌ی تعیین فاز توسط یک رشته‌ی بیتی	۶۳
۲۰۲	”کراس‌اور“ جواب‌ها در الگوریتم ژنتیک ساده برای حل مسئله‌ی تفکیک ژنوتیپ‌ها	۶۵
۳۰۲	”جهش“ جواب‌ها در الگوریتم ژنتیک ساده برای حل مسئله‌ی تفکیک ژنوتیپ‌ها	۶۶
۴۰۲	”کراس‌اور“ جایگشت‌ها در الگوریتم ژنتیک بهبودیافته	۶۹
۵۰۲	تولید هاپلوتیپ‌های تصادفی تحت مدل فیلوژنی کامل توسط الگوریتم RandPerfectHap	۷۶
۶۰۲	مشاهده‌ی نمونه‌هایی با فراوانی‌های حاشیه‌ای یکسان در جمعیت	۸۲
۷۰۲	نحوه‌ی محاسبه‌ی p -مقدار در آزمون دقیق فیشر	۸۶
۸۰۲	نمودار تغییرات آماره‌های همبستگی و p -مقدار بر حسب فراوانی هاپلوتیپ 11	۸۷
۹۰۲	نقشه‌ی LD بین جفت اسنپ‌ها در قسمتی از ناحیه‌ی 4q26 (ENr113)	۱۰۱
۱۰۰۲	جدول توافقی برای مطالعه‌ی همبستگی بین یک اسنپ و خصیصه	۱۰۷

- ۱۱۰۲ جدول توافقی برای مطالعه‌ی همبستگی بین یک بلوک از هاپلوتیپ‌ها و خصیصه ۱۰۹
- ۱۰۳ روند همگرایی به جواب در الگوریتم GAhap به ازای مقادیر مختلف cr ۱۲۲
- ۲۰۳ نمودارهای دقت و کارایی الگوریتم GAhap بر حسب تعداد اسنپ‌های نمونه ۱۲۴
- ۳۰۳ نمودارهای دقت و کارایی الگوریتم GAhap بر حسب تعداد ژنوتیپ‌های نمونه ۱۲۵
- ۴۰۳ افرازهای بلوکی مختلف در ناحیه‌ی (ENr113)4q26 ۱۲۸
- ۵۰۳ افرازهای بلوکی مختلف در ناحیه‌ی (ENm010)7p15.2 ۱۲۹
- ۶۰۳ افرازهای بلوکی مختلف در ناحیه‌ی (ENr131)2q37.1 ۱۳۰
- ۷۰۳ پوشش htSNP‌ها به ازای افرازهای بلوکی متفاوت برای نواحی ENCODE ۱۳۴
- ۸۰۳ دقت بازسازی هاپلوتیپ‌ها توسط htSNP‌ها ۱۳۶
- ۹۰۳ ثبات روش‌های مختلف افراز در تعریف ساختار بلوکی ناحیه‌ی (ENr232)9q34.11 ۱۳۹
- ۱۰۰۳ کارایی روش‌های افراز بلوکی در شناسایی نقاط پراحتمال نوترکیبی ۱۴۲
- ۱۱۰۳ کارایی روش‌های افراز بلوکی هاپلوتیپ‌ها در شناسایی جایگاه ژنی مرتبط با خصیصه، با انتخاب یکنواخت نشانگذارها ۱۴۴
- ۱۲۰۳ کارایی روش‌های افراز بلوکی هاپلوتیپ‌ها در شناسایی جایگاه ژنی مرتبط با خصیصه، با انتخاب اولویت داده شده‌ی نشانگذارها ۱۴۵

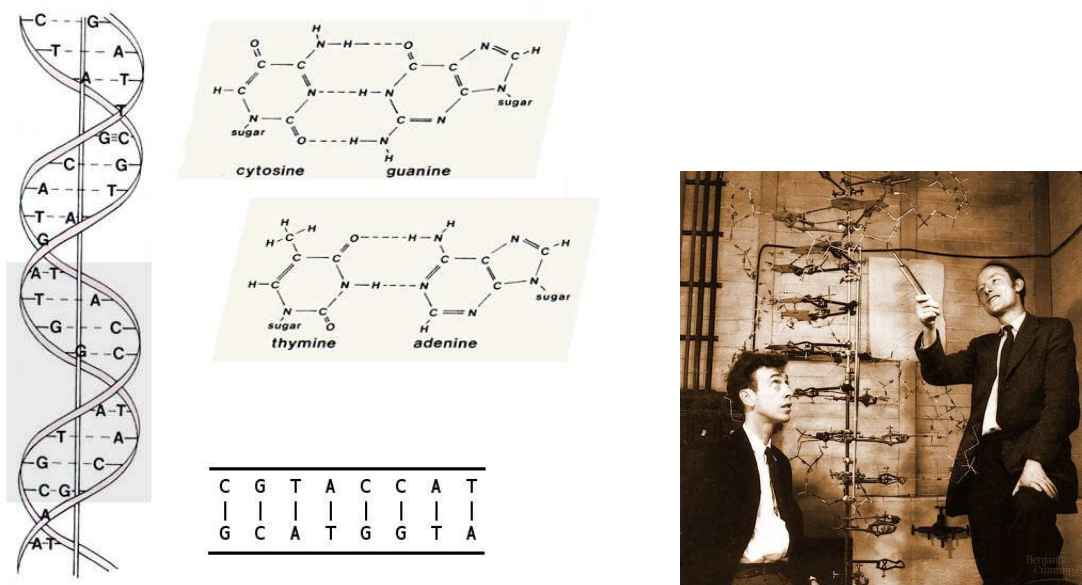
فصل ۱

مقدمه

۱۰۱ آشنایی با واژگان ژنتیک

برای هر موجود زنده، اطلاعات به مفهوم کلی کلمه، از خصیصه‌های ظاهری یک فرد در جمعیتی از جانداران گرفته تا جزئیات دستورالعمل تقسیم سلولی، در واحدهایی که به طور کلی آنها را ژن می‌نامیم نگهداری می‌شوند. ژنها طی تولیدمثل جاندار یا تقسیم سلولی، تکثیر می‌شوند و نسخه‌ی یکسانی از آنها به نسل بعد منتقل می‌شود تا بدین ترتیب موجود جدید تمام اطلاعات لازم برای حیات را در اختیار داشته باشد. تا پیش از انقلابی که با کشف واتسون و کریک در زیست‌شناسی آغاز شد معلوم شده بود که ژنها با ماده‌ای که درون هسته‌ی سلول است در ارتباطند. این ماده با نام شیمیایی اسید دزوکسی ریبونوکلیک یا به اختصار DNA شناخته می‌شد. با کشف ساختار DNA معلوم شد که DNA مولکول پلیمری بسیار طولی از زیرواحدهای ساده‌تری به نام نوکلئوتید است. آنها نشان دادند تنها چهار نوع نوکلئوتید در ساختمان DNA نقش دارند که اکنون آنها را به اختصار با حروف A، C، G و T می‌شناسیم. در واقع ترکیبهای مختلف کنار هم قرار گرفتن این چهار نوع نوکلئوتید در امتداد زنجیر DNA به طور آرمانی توانایی رمزکردن بیشمار ژن را فراهم می‌کند. زنجیر DNA جدا از ساختار نردبانی تاب‌خورده‌اش توالیی از چند صد میلیون نوکلئوتید است که می‌توان آنرا مانند یک رشته از الفبای چهار حرفی {A,C,G,T} در نظر گرفت. البته طبیعت برای حفظ پایداری بیشتر و

مضمون نگهداشتن این اطلاعات از صدمات، اطلاعات را به صورت تکراری، به شکل یک نردبان دورشته‌ای شامل جفت نوکلئوتیدهای روبروی هم نگه می‌دارد. از آنجا که نوکلئوتیدها دارای خاصیت بازی نیز هستند، اندازه‌ی DNA را در واحد جفت‌باز (base pair) می‌شمارند. به عنوان مثال اندازه‌ی رشته‌ی DNA در یک سلول باکتری E.Coli، نزدیک به ۵ میلیون جفت‌باز است.



شکل ۱۰۱: ساختمان DNA

راست) واتسون و کریک در کنار مدل ساختمان DNA. چپ) ساختار اتمی اسیدهای نوکلئیک و موقعیت آنها در قطعه‌ای از زنجیر DNA.

هر ژن جایگاه خاصی از این رشته‌ی عظیم را به خود اختصاص می‌دهد. با اینکه قابل تصور است اگر هر ژن رشته‌ی DNA جداگانه‌ای می‌داشت ولی تکامل، طبیعت را به این سو هدایت کرده است که تمامی ژنها بر روی یک رشته‌ی بزرگ DNA در کنار یکدیگر قرار گیرند. البته این وضعیت در موجودات عالی کمی تغییر می‌کند. بدین ترتیب که ژنها بر روی چندین قطعه‌ی بزرگ DNA در هسته‌ی سلول قرار می‌گیرند. هر یک از این قطعات بزرگ DNA یک کروموزوم را تشکیل می‌دهند. از نظر ساختار فضایی، هر کروموزوم ساختمان در هم پیچیده و فشرده‌شده‌ی یک زنجیر بسیار بلند DNA است که در زیر میکروسکپ به اشکال خاصی قابل مشاهده است. از دیدگاه سلولی، تعداد و اندازه کروموزومها خصیصه‌ای مربوط به گونه است، یعنی در تمام سلولهای افراد یک گونه‌ی خاص (تقریباً) یکسان است. مثلاً این تعداد در مورد انسان ۴۶ است. البته در تمامی جاندارانی که از تولیدمثل جنسی بوجود می‌آیند، سلولها همواره حاوی دو دست کروموزوم هستند. هر دست

کروموزوم نسخه‌ی یکسانی از کروموزومهایی است که از هر یک از والدین به ارث رسیده است. بدین ترتیب هر سلول انسان دقیقاً حاوی ۲۳ جفت کروموزوم است. کروموزومهای متناظر در هر جفت را کروموزومهای همسان (homologous) می‌گویند. مجموع DNA در هر دست از این ۲۳ جفت کروموزوم شامل حدود ۳ میلیارد جفت‌باز می‌شود. جاندارانی که دو نسخه‌ی یک DNA را داشته باشند اصطلاحاً دیپلوئید (diploid) نامیده می‌شوند.

اصطلاح ژنوم به مجموعه‌ی همه‌ی ژنهای یک جاندار اطلاق می‌شود. در موجودات عالی نواحی بسیار وسیعی از DNA فاقد ویژگیهای عمومی ژنها هستند. به عنوان مثال بیش از سه چهارم DNA در انسان را نواحی بین ژنی تشکیل می‌دهند. با این حال امروزه باور بر این است که برای هیچ یک از مناطق DNA به قطعیت نمی‌توان اظهار کرد که هیچ نقشی وجود ندارد. بنابراین رایج بر این است که تمامی DNA یک جاندار را ژنوم آن جاندار بنامند. در این رساله، منظور از ناحیه‌ی ژنومی، قطعه‌ای از کروموزوم است که علاقمند به مطالعه برخی ویژگی‌های آن هستیم. منظور از فنوتیپ خصیصه‌های ظاهری و به سادگی قابل تشخیص در بین افراد مختلف یک گونه از جانداران است، مثل رنگ پوست یا استعداد ابتلا به یک بیماری خاص. در واقع، وجود هر فنوتیپ مرتبط با بروز یک یا چند ژن خاص است. توالی‌های مختلفی که می‌توانند در این ژن یا ژنها دیده شوند را اصطلاحاً ژنوتیپ‌های مرتبط با فنوتیپ می‌نامند.

تنوع حیات ناشی از تنوع ژنوم است. دو رویداد اصلی زمینه‌ی بروز این تنوع هستند: جهش و نوترکیبی. جهش رویدادی است که طی آن یک یا چند باز بر روی رشته‌ی DNA به باز یا بازهای دیگری تغییر می‌یابد یا در مواردی یک نوکلئوتید از رشته حذف می‌شود یا یک نوکلئوتید جدید در رشته درج می‌شود. اگر اطلاعات ژنتیکی در این توالی جدید آنچنان معیوب نشده باشد که به مرگ جاندار منجر شود، توالی جهش یافته به نسلهای بعدی منتقل می‌شود و بسته به شرایط جمعیتی و میزان سازگاری (fitness) با شرایط محیط، سهمی از ژنوتیپ جمعیتی را به خود اختصاص می‌دهد.

در متداولترین شکل رویداد نوترکیبی، بخشهایی مشابهی از دو کروموزوم همسان با یکدیگر مبادله می‌شوند که به آن cross-over می‌گویند. منظور از کروموزومهای همسان، کروموزومهای متناظر با یکدیگر در افراد مختلف یک گونه‌ی ثابت است. بر خلاف رویداد جهش، ژنومی که از چنین نوترکیبی بدست می‌آید به ندرت

ممکن است به مرگ جاندار و حتی به زایش گونه‌ی جدیدی منجر شود. از این رو نرخ رویدادهای نوترکیبی در مقایسه با جهش به مراتب بیشتر است. بین کروموزومهای همسان انسان در هر نسل تا دهها نوترکیبی می‌تواند روی دهد.

باید به یاد داشت که هنگامی که صحبت از ژنوم انسان می‌شود منظور توالی کاملی از DNA است که به طور مشترک در بین تمام افراد مختلف جمعیت انسان دیده می‌شود. در واقع توالی ژنوم در بین افراد مختلف یک گونه‌ی معین به جز اختلافهای بسیار جزئی، یکسان هستند. مثلاً در مورد افراد انسان، تخمین زده می‌شود بیش از ۹۹/۹ درصد ژنوم بین هر دو فرد یکسان باشد^۱. همین اختلافهای ناچیز زمینه‌ی اصلی تفاوت‌های ظاهری یا فنوتیپی در بین افراد مختلف است. این اختلافها به اشکال گوناگونی بر روی ژنوم دیده می‌شوند. رایج‌ترین شکل تفاوت در بین ژنوم افراد مختلف یک جمعیت، چندریختی تک نوکلئوتیدی (Single Nucleotide Polymorphis) یا به اختصار SNP (اسنیپ) است. یک اسنیپ جایگاهی بر روی ژنوم است که در بین افراد مختلف جمعیت بیش از یک نوع نوکلئوتید در آن مشاهده می‌شود. در واقع، هر اسنیپ نتیجه‌ی یک رویداد جهش در زمان بسیار دور بوده است که به نسل‌های بعد منتقل شده است تا در زمان حال که ژنوم کسر معینی از جمعیت، حامل آن است.

به طور کلی اشکال مختلفی که در یک جایگاه ژنی معین دیده می‌شوند را آلل‌های آن ژن می‌گویند. مثال ساده‌ی آن، آللهای ژن مرتبط با گروه خون است که آنها را با A، B و O می‌شناسیم. در مورد اسنیپ‌ها، نوکلئوتیدی را که فراوانتر در جمعیت مشاهده می‌شود آلل اصلی و نوکلئوتیدی که کمتر مشاهده می‌شود را آلل فرعی می‌نامیم. نرخ رویداد جهش، از آن دست که مشکلی برای بقای جاندار ایجاد نکند، بسیار پایین است، آن چنان که احتمال اینکه در یک جفت باز در بین میلیاردها جفت باز ژنوم انسان دو بار جهش روی داده باشد نزدیک به صفر است. بدین ترتیب تقریباً همه اسنیپ‌ها دو آللی هستند و به ندرت اسنیپ سه آللی مشاهده می‌شود. از دیگر سو، بر روی ژنوم هر فرد، نوکلئوتیدهای فراوانی ممکن است وجود داشته باشند که با آنچه در ژنوم دیگر افراد وجود دارد تفاوت داشته باشند. با این حال، بسیاری از آنها جز جهش‌های بدنی (somatic mutation) بیانگر اطلاعات دیگری نیستند و نمی‌توان آنها را اسنیپ به شمار آورد. از اینرو رایج آن است

^۱ جالب است اگر بدانید بین ژنوم انسان و ژنوم موش قریب به ۹۰ درصد شباهت وجود دارد.

که یک کران پایین برای فراوانی مشاهده‌ی آلل فرعی^۲ در نظر می‌گیرند تا این جهش‌های نادر از اسنپ‌های رایج در جمعیت تمایز داده شوند. مثلاً در بسیاری از مطالعات، تنها جایگاه‌هایی را به عنوان اسنپ در نظر می‌گیرند که در آنها فراوانی آلل فرعی بیشتر از یک درصد باشد.

A G G A C T A G A T A A T A G A C C G	0 1 1 0 0
A G G A C C A C A T T A T A G T C C G	0 0 0 1 1
A G G A C C A G A T A A T A G T C C G	0 0 1 0 1
A T G A C C A C A T T A T A G T C C G	1 0 0 1 1
A T G A C T A C A T A A T A G A C C G	1 1 0 0 0

شکل ۲۰۱: آلل‌های متفاوت پنج اسنپ در قطعه‌ی یکسانی از پنج نمونه‌ی متفاوت از یک کروموزوم توالی کروموزومی هر فرد را می‌توان مجموعه‌ای از توالی نوکلئوتیدهای یکسان در بین تمام افراد و توالی نوکلئوتیدهای فرد در جایگاه‌های اسنپ در نظر گرفت. توالی کروموزوم در جایگاه‌های اسنپ را می‌توان با کدهای صفر و یک برای هر نمونه نمایش داد.

وقتی ژنوتیپ یک جاندار دیپلوئید را در یک جایگاه اسنپ بررسی می‌کنیم سه وضعیت متفاوت ممکن است مشاهده شود: یا آلل‌های بر روی دو کروموزوم، متفاوتند که یک حالت هتروزیگوت (heterozygot) نامیده می‌شود یا حالتی را داریم که هر دو کروموزوم آللهای یکسانی دارند یعنی یا هر دو آلل، آلل اصلی هستند یا هر دو، آلل فرعی هستند، که به ترتیب هموزیگوت ماژور (homozygot major) و هموزیگوت مینور (homozygot minor) نامیده می‌شوند. وقتی توالی ژنومی را تنها به جایگاه اسنپ‌ها محدود کنیم، دنباله‌ای از وضعیت‌های سه گانه بر روی اسنپ‌ها بدست می‌آید. به این توالی اسنپ-ژنوتیپ می‌گوییم. یادآوری این نکته ضروری است که در این حالت توالی ژنوم همزمان از مجموعه‌ی دو کروموزوم همسان خوانده می‌شود و به همین جهت در مواردی برای تاکید بر این موضوع، توالی آللهای در این حالت را دیپلوטיפ (diplotype) می‌گویند. اگر توالی ژنوم را تنها بر روی یک کروموزوم به طور مستقل مطالعه کنیم، توالی آللهایی که در موقعیت اسنپ‌ها مشاهده می‌شوند اسنپ-هاپلوטיפ (SNP-haplotype) نامیده می‌شوند. در این رساله به اختصار واژگان هاپلوטיפ و ژنوتیپ را به ترتیب برای اشاره به اسنپ-هاپلوטיפ و اسنپ-دیپلوטיפ به کار می‌بریم.

^۲Minor Allele Frequency (MAF)