

مقدمه‌ای بر بیوانفورماتیک کاربردی

تألیف

پل ام. سلزر
ریچارد جی. ماروفر
الیور کوخ

ترجمه

دکتر سید محمد مهدی مرتضویان
دانشیار ژنتیک و به نژادی گیاهی دانشگاه تهران

ویرایش دوم

۲۰۱۸



شماره مسلسل ۱۰۶۸۷

شماره انتشار ۴۳۹۲

انتشارات دانشگاه تهران

سرشناسه	: سلزر، پل.ام. Selzer, P.M.(Paul M)
عنوان و نام پدیدآور	: مقدمه‌ای بر بیوانفورماتیک کاربردی / تألیف پل.ام. سلزر، ریچارد. ماروفر، الیور. کوخ؛ ترجمه سیدمحمد مهدی مرتضویان.
مشخصات نشر	: تهران: دانشگاه تهران، مؤسسه انتشارات، ۱۳۹۹.
مشخصات ظاهری	: ۲۴۰ص.
فروست	: انتشارات دانشگاه تهران؛ شماره انتشار ۴۳۹۲.
شابک	: 978-964-03-0163-0
وضعیت فهرست‌نویسی	: فیبا
یادداشت	: Applied Bioinformatics: an Introduction, 2th ed, c2008.
یادداشت	: کتاب حاضر در سالهای مختلف توسط مترجمان و ناشران متفاوت ترجمه و منتشر شده است.
یادداشت	: نمایه.
موضوع	: زیست انفورماتیک Bioinformatics
موضوع	: زیست انفورماتیک -- مسائل، تمرین‌ها و غیره.
موضوع	: Bioinformatics - Problems, Exercises, etc
شناسه افزوده	: ماروفر، آر.ج. (ریشارد ج.)
شناسه افزوده	: Marhöfer, R.J.(Richard J.)
شناسه افزوده	: کوخ، الیور
شناسه افزوده	: Koch, Oliver
شناسه افزوده	: مرتضویان، سیدمحمد مهدی، ۱۳۵۸- مترجم
شناسه افزوده	: دانشگاه تهران. مؤسسه انتشارات. University of Tehran. Press
رده‌بندی کنگره	: QH۳۲۴/۲ ۱۴۰۰
رده‌بندی دیویی	: ۵۷۲/۸۰۲۸۵
شماره کتابشناسی ملی	: ۷۵۱۹۷۸۲

این کتاب مشمول قانون حمایت از حقوق مؤلفان و مصنفان است. تکثیر کتاب به هر روش اعم از فتوکپی، ریسوگرافی، تهیه فایل‌های pdf، لوح فشرده، بازنویسی در وبلاگ‌ها، سایت‌ها، مجله‌ها و کتاب، بدون اجازه کتبی ناشر مجاز نیست و موجب پیگرد قانونی می‌شود و تمامی حقوق برای ناشر محفوظ است.
(این کتاب با کاغذ حمایتی به چاپ رسیده است.)



عنوان: مقدمه‌ای بر بیوانفورماتیک
تألیف: پل سلزر- ریچارد ماروفر- الیور کوخ
ترجمه: دکتر سیدمحمد مهدی مرتضویان
ویرایش ادبی: نسترن حسین‌پور
نوبت چاپ: اول
تاریخ انتشار: ۱۴۰۰
شمارگان: ۱۰۰ نسخه
ناشر: مؤسسه انتشارات دانشگاه تهران
چاپ و صحافی: مؤسسه انتشارات دانشگاه تهران

«مسئولیت صحت مطالب کتاب با مترجم است»

بها: ۷۲۰۰۰۰ ریال

خیابان کارگر شمالی - خیابان شهید فرشی مقدم - مؤسسه انتشارات دانشگاه تهران
پست الکترونیک: press@ut.ac.ir - تارنما: http://press.ut.ac.ir
پخش و فروش: تلفکس ۸۸۳۳۸۷۱۲



فهرست مطالب

پیشگفتار مترجم.....	ذ
گردش اطلاعات ژنتیکی.....	ش
تاریخچه مختصری از بیوانفورماتیک.....	ش
منابع.....	ط
فصل اول: مبانی زیستی بیوانفورماتیک.....	۱
۱-۱- پروتئین‌ها و اسیدهای نوکلئیک.....	۱
۱-۲- ساختار اسیدهای نوکلئیک DNA و RNA.....	۱
۱-۳- ذخیره اطلاعات ژنتیکی.....	۴
۱-۴- ساختار پروتئین‌ها.....	۶
۱-۴-۱- ساختار اولیه.....	۶
۱-۴-۲- ساختار ثانویه.....	۸
۱-۴-۳- ساختار سوم و چهارم.....	۹
۱-۵- تمرین‌ها.....	۱۱
منابع.....	۱۲
منابع برای مطالعه بیشتر.....	۱۲
فصل دوم: پایگاه‌های داده زیستی.....	۱۳
۱-۲- اطلاعات زیستی ذخیره شده در پایگاه‌های اطلاعاتی Global.....	۱۳
۲-۲- پایگاه‌های داده اولیه.....	۱۴
۲-۲-۱- پایگاه‌های داده توالی نوکلئوتیدی.....	۱۴
۲-۲-۱-۱- بانک ژن یا GenBank.....	۱۴
Entrez.....	۱۵
EMBL و DDBJ.....	۱۶
فراخوانی یا جست‌وجوی آنلاین ENA.....	۱۷
۲-۲-۲- پایگاه‌های توالی پروتئین.....	۲۱

۲۱	UniProt - ۱ - ۲ - ۲ - ۲
۲۴	NCBI پایگاه داده پروتئین - ۲ - ۲ - ۲ - ۲
۲۵	پایگاه‌های داده ثانویه - ۳ - ۲
۲۵	Prosites - ۱ - ۳ - ۲
۲۶	PRINTS - ۲ - ۳ - ۲
۲۷	Pfam - ۳ - ۳ - ۲
۲۷	Interpro - ۴ - ۳ - ۲
۲۸	پایگاه‌های داده فنوتیپ و ژنوتیپ - ۴ - ۲
۲۸	PhenomicDB - ۱ - ۴ - ۲
۳۰	پایگاه‌های داده ساختار مولکولی - ۵ - ۲
۳۰	بانک داده پروتئین - ۱ - ۵ - ۲
۳۱	SCOP - ۲ - ۵ - ۲
۳۲	CATH - ۳ - ۵ - ۲
۳۳	PubChem - ۴ - ۵ - ۲
۳۴	تمرین‌ها - ۶ - ۲
۳۷	منابع
۳۷	منابع برای مطالعه بیشتر
۳۹	فصل سوم: مقایسه‌های توالی و جست‌وجوی پایگاه‌های داده توالی
۳۹	۱ - ۳ - مقایسه‌های چندتایی و دوتایی توالی‌ها
۴۷	۲ - ۳ - جست‌وجو در پایگاه اطلاعات توالی‌های پروتئینی و نوکلئوتیدی
۵۱	۳ - ۲ - ۱ - الگوریتم‌های مهم در جست‌وجوی پایگاه داده‌ها
۵۲	۳ - ۳ - نرم‌افزار تجزیه و تحلیل توالی
۵۴	۴ - ۳ - تمرین‌ها
۵۵	منابع
۵۶	منابع برای مطالعه بیشتر
۵۷	فصل چهارم: رمزگشایی ژنوم یوکاریوتی
۵۷	۴ - ۱ - توالی‌یابی ژنوم کامل
۵۷	۴ - ۲ - تعیین ویژگی‌های ژنوم با استفاده از توالی‌های STS و EST

فهرست مطالب □ ج

۴-۲-۱	نواحی نشاندار ژنوم یا sequence tagged sites مشخصه‌های اصلی در ژنوم انسانی هستند	۵۷
۴-۲-۲	دنباله توالی‌های بیان شده یا Expressed sequence tags	۵۸
۴-۳	اجرای پروژه EST	۶۰
۴-۴	شناسایی ژن‌های ناشناخته	۶۲
۴-۵	کشف واریانت‌های پیرایشی	۶۶
۴-۶	علل ژنتیکی تفاوت‌های فردی	۶۷
۴-۶-۱	Pharmacogenetics	۶۹
۴-۶-۲	پزشکی شخصی و نشانگرهای زیستی	۷۳
۴-۶-۳	توالی‌یابی نسل جدید یا Next-Generation Sequencing (NGS)	۷۵
۴-۶-۴	پروتئوژنومیکس Proteogenomics	۷۷
۴-۷	تمرین‌ها	۷۸
۴-۱۴	تمرین	۸۰
	منابع	۸۰
	منابع برای مطالعه بیشتر	۸۱
فصل پنجم: ساختارهای پروتئینی و طراحی هوشمندانه داروی ساختاربنیان		
۵-۱	ساختار پروتئین	۸۳
۵-۲	پپتیدهای راهنما یا سیگنال پپتیدها	۸۴
	پیش‌بینی ۴.۱-SignalP (گرم - شبکه): توالی	۸۵
۵-۳	پروتئین‌های گذرنده از غشا	۸۶
	احتمال پسین TMHMM برای 5HT1B_SPAEH P۵۶۴۹۶ sp	۸۷
۵-۴	آنالیز ساختارهای پروتئینی	۸۷
۵-۴-۱	مدل‌سازی پروتئین	۸۷
۵-۴-۲	تعیین ساختارهای پروتئین با روش‌هایی با کارایی زیاد	۸۸
۵-۵	طراحی هوشمند داروی ساختاربنیان	۹۰
۵-۵-۱	نمونه‌ای از شبیه‌سازی محاسباتی اتصال با استفاده از برنامه DOCK	۹۰
۵-۵-۲	نمونه‌ای از شبیه‌سازی محاسباتی اتصال با استفاده از برنامه GOLD	۹۴
۵-۵-۳	جست‌وجو و مدل‌سازی فارماکوفور	۹۵
۵-۵-۴	موفقیت طراحی داروی ساختاربنیان	۹۶
۵-۶	تمرین‌ها	۹۸

منابع.....	۹۹
منابع برای مطالعه بیشتر.....	۱۰۱
فصل ششم: تجزیه و تحلیل کارکردی ژنوم.....	۱۰۳
۱-۶ شناسایی عملکردهای سلولی محصولات ژن.....	۱۰۳
۱-۱-۶ ترنسکریپتومیکس.....	۱۰۴
۱-۱-۶ ریزآرایه‌های DNA.....	۱۰۵
اجرای یک آزمایش تعیین نیمرخ بیان با cDNA.....	۱۰۸
تفسیر آزمایش تهیه نیمرخ بیان.....	۱۰۹
۱-۱-۶ آنالیز سریال بیان ژن.....	۱۱۴
۱-۲-۶ پروتئومیکس.....	۱۱۵
۱-۲-۶ پروتئومیکس کلاسیک.....	۱۱۵
۱-۲-۶ پروتئومیکس عملکردی.....	۱۲۰
۱-۲-۶ آرایه‌های پروتئین.....	۱۲۴
۱-۳-۶ متابولومیکس.....	۱۲۵
۱-۴-۶ فنومیکس.....	۱۲۸
۲-۶ زیست‌شناسی سیستم‌ها.....	۱۳۱
۳-۶ تمرین‌ها.....	۱۳۵
منابع.....	۱۳۷
منابع برای مطالعه بیشتر.....	۱۳۹
فصل هفتم: تجزیه و تحلیل مقایسه‌ای ژنوم.....	۱۴۱
۱-۷ عصر ژنوم مقایسه‌ای.....	۱۴۱
۲-۷ تحقیقات دارویی در زمینه پروتئین هدف.....	۱۴۱
۳-۷ آنالیز مقایسه‌ای ژنوم اطلاعاتی در مورد زیست‌شناسی موجودات فراهم می‌کند.....	۱۴۴
۱-۳-۷ ساختار ژنوم.....	۱۴۴
۲-۳-۷ نواحی کدکننده.....	۱۴۵
۳-۳-۷ نواحی غیرکدکننده.....	۱۴۶
۴-۷ تجزیه و تحلیل مقایسه‌ای متابولیک.....	۱۴۶
۱-۴-۷ دایره‌المعارف ژن‌ها و ژنوم کیوتو (KEGG).....	۱۵۱

فهرست مطالب □ خ

۱۵۳	۷-۵- گروه‌بندی پروتئین‌های ارتولوگ.....
۱۵۷	۷-۶- تمرین‌ها.....
۱۵۸	منابع.....
۱۵۹	منابع برای مطالعه بیشتر.....
۱۶۱	پاسخ تمرین‌ها.....
۱۶۱	فصل اول.....
۱۶۲	فصل دوم.....
۱۶۹	فصل سوم.....
۱۷۳	فصل چهارم.....
۱۷۷	فصل پنجم.....
۱۸۰	فصل ششم.....
۱۸۳	فصل هفتم.....
۱۸۹	فرهنگ اصطلاحات.....
۲۱۷	نمایه.....

پیشگفتار مترجم

در چند دهه اخیر به مدد پیشرفت‌های سریع و شگرف در زیست‌شناسی مولکولی و ابزارهای مطالعاتی با توان بالا دستیابی به اطلاعات خام اولیه از ماده ژنتیکی حیات یعنی ژنوم موجودات زنده و نیز محصولات آن‌ها یعنی پروتئین‌ها و متابولیت‌های سلولی آسان‌تر و نسبت به گذشته سهل‌الوصول شده است. تولید حجم زیاد داده در زمان اندک لزوم روش‌های دقیق ارزیابی خطاهای احتمالی در تولید آن‌ها، سرهم‌بندی توالی‌های به‌دست‌آمده و نیز بررسی کارکرد و نقش آن‌ها در سلول را نیازمند توسعه و طراحی الگوریتم‌های آماری و واکاوی داده کرده است. از همین رو بیوانفورماتیک یا داده‌پردازی زیستی به‌عنوان یکی از دانش‌های نوظهور که از علوم ریاضی و کامپیوتر برای تجزیه و تحلیل و بررسی داده‌ها و پیش‌بینی نتایج کمک می‌گیرد، بیش از پیش در مرکز توجه محققان علوم زیستی قرار دارد.

با توجه به اهمیت این رشته علمی، درسی با عنوان بیوانفورماتیک در بسیاری از حوزه‌های علوم زیستی در مقاطع کارشناسی، کارشناسی ارشد و دکتری به فراخور گرایش تحصیلی اضافه شده است. همانند بسیاری از علوم، بیوانفورماتیک حوزه‌های مختلفی را شامل می‌شود. آشنایی با مبانی بیوانفورماتیک با نیم‌نگاهی به اصول زیستی به همراه ارائه تمرین‌های عملی، دانشجویان و علاقه‌مندان به بیوانفورماتیک را با کاربردهای آن آشنا می‌کند. علی‌رغم در دست بودن کتاب‌های بیوانفورماتیک متعدد، به‌خصوص در سال‌های اخیر، کتاب حاضر در دنیا استقبال فراوان دیده و از همین رو چاپ دوم این کتاب به دو زبان آلمانی و انگلیسی و به‌طور هم‌زمان در سال ۲۰۱۸ منتشر شده است. همان‌طور که مؤلفان تأکید کرده‌اند، این کتاب از این نظر که شامل تمرین‌ها و پاسخ سؤال‌های آن‌هاست، کتابی منحصربه‌فرد است که آن را برای تدریس کلاسی کاربردی‌تر می‌کند. کتاب حاضر که نتیجه سال‌ها تحقیق و مطالعه نگارندگان است، به زبان‌های آلمانی و انگلیسی به رشته تحریر درآمده و دانشجویان و محققان دنیا از آن استفاده می‌کنند. نسخه پیش رو که ترجمه ویرایش دوم کتاب با عنوان *Applied Bioinformatics* است و در سال ۲۰۱۸ به رشته تحریر درآمده است، به زبانی ساده نگارش یافته و در انتهای هر فصل تمرین‌های عملی برای ارزیابی فراگیران ارائه شده است. در انتهای کتاب نیز راه‌حل تمرین‌ها ارائه شده که می‌تواند علاقه‌مندان به سایت‌های بیوانفورماتیک و کاربرد بیوانفورماتیک در حل مسائل زیستی را به‌صورت گام به گام با خود همراه کند.

هریک از هفت فصل کتاب موضوعات مهم در حوزه بیوانفورماتیک کاربردی را شرح داده و منابع و لینک‌های اینترنتی را در اختیار قرار می‌دهد. تمرین‌ها و راه‌حل‌های مفصل آن برای ترغیب خواننده به تمرین و یادگیری موضوع است و سبب افزایش مهارت و تجربه در استفاده از نرم‌افزارهای مرتبط

می‌شود. فصل نخست، مبانی زیستی بیوانفورماتیک شامل ساختار ماکرومولکول‌های زیستی را شرح می‌دهد. فصل دوم، ضمن معرفی پایگاه‌های داده اولیه، ثانویه و اختصاصی اقدام به معرفی پایگاه‌های مهم و کاربرد آن‌ها می‌کند. فصل سوم، هم‌ردیفی توالی‌ها و الگوریتم‌های مهم در هم‌ردیفی مورد استفاده در برنامه‌های مختلف جست‌وجو از جمله BLAST را شرح می‌دهد. فصل چهارم، روش‌های توالی‌یابی کامل ژنوم، EST، اصول فارماکوژنتیک و اهمیت بیوانفورماتیک در این شاخه مهم پزشکی، پزشکی شخصی و نسل بعدی توالی‌یابی را شرح می‌دهد. فصل پنجم، به یکی از موضوعات کاربردی مهم در بیوانفورماتیک یعنی طراحی هوشمندانه ساختارینیان به کمک داده‌های بیوانفورماتیک و روش‌های محاسباتی آن می‌پردازد. فصل ششم، بررسی ژنومیکس کارکردی شامل ترنسکریپتومیکس، پروتئومیکس، متابولومیکس و فنومیکس و در نهایت ارتباط آن‌ها با یکدیگر در زیست‌شناسی سیستم‌ها را بررسی می‌کند و به فراخور به معرفی پایگاه‌های داده مرتبط می‌پردازد. فصل هفتم، ژنومیکس مقایسه‌ای و پایگاه‌های داده متابولومیکس را معرفی می‌کند. در ادامه پاسخ تمرین‌های ارائه‌شده در انتهای هر فصل به‌طور کامل شرح داده می‌شود و در فصل پایانی فرهنگ اصطلاحات یا دایره‌المعارف بیوانفورماتیک ارائه می‌شود.

دانشجویان رشته‌های بیوانفورماتیک، زیست‌شناسی، بیوتکنولوژی، ژنتیک مولکولی، بیوتکنولوژی کشاورزی، مهندسی ژنتیک و مهندسی کشاورزی در مقاطع کارشناسی ارشد و دکتری و همچنین محققان علاقه‌مند به مبانی بیوانفورماتیک مخاطبان اصلی این کتاب هستند. تدریس در دروسی نظیر بیوانفورماتیک مقدماتی (۲ واحد) در دوره کارشناسی ارشد به‌نژادی گیاهی و بیوانفورماتیک پیشرفته (۲ واحد) در مقطع دکتری ژنتیک و به‌نژادی گیاهی و بیوتکنولوژی کشاورزی، درس بیوانفورماتیک (۲ واحد) برای دانشجویان مقطع دکتری رشته بیوانفورماتیک و سایر رشته‌ها می‌توانند از مباحث کتاب استفاده کنند.

در پایان امید است ترجمه و ارائه کتاب حاضر که مصادف با سالروز ولادت باسعادت حضرت فاطمه زهرا(س) به پایان رسیده است، مورد توجه علاقه‌مندان و دانشجویان عزیز قرار گیرد.

بهمن‌ماه ۱۳۹۸

پیشگفتار مؤلفان

اگرچه بیوانفورماتیک یک رشته نسبتاً جدید است، در بسیاری از حوزه‌های علوم زیستی از جمله زیست‌شناسی، بیوشیمی، پزشکی و شیمی روزبه‌روز اهمیت بیشتری پیدا می‌کند. موفقیت بیوانفورماتیک از زمان شروع آن در اواخر دهه ۱۹۸۰، با پیشرفت سریع در علوم رایانه، به‌ویژه در زمینه سخت‌افزاری و نرم‌افزاری همراه بوده است. علاوه‌براین، پیشرفت‌های بیوتکنولوژی مانند پیشرفت‌هایی که در زمینه توالی‌یابی ژنوم، ریزآرایه و پروتئومیکس رخ داده است، به رونق بیوانفورماتیک بسیار کمک کرده‌اند. درنهایت، پیشرفت شگرف و گسترش شبکه جهانی وب، توزیع جهانی و دسترسی آسان به ابزارهای بیوانفورماتیک در سراسر جهان را تسهیل کرده است.

امروزه، تکنیک‌های بیوانفورماتیک مانند الگوریتم BLAST، مقایسه جفتی و چندتایی توالی‌ها، جست‌وجو در پایگاه‌های داده زیستی و تجزیه و تحلیل فیلوژنتیک، ابزارهایی معروف برای محققان علوم زیستی است. بسیاری از محصولات نرم‌افزاری که در ابتدا ناقص و مبهم بود، به بلوغ رسیده‌اند و به برنامه‌هایی نسبتاً ساده و کاربرپسند تبدیل شده‌اند که به راحتی از طریق اینترنت در دسترس قرار می‌گیرند. دیگر نیاز نیست که محقق، یک متخصص علوم رایانه‌ای باشد تا بتواند ابزارهای بیوانفورماتیک را با توجه به سؤالات پیچیده علمی به کار گیرد. البته مسئله مهمی که هنوز هم باقی است، درک مبانی اساسی زیستی به همراه اطلاع از ابزارهای مناسب بیوانفورماتیک و چگونگی به کارگیری آن‌هاست. مسئله مهم بعدی اطمینان از به کارگیری صحیح این ابزارها برای دستیابی به نتایج معتبر است.

کتاب حاضر، ویرایش دوم نسخه انگلیسی کتاب است که به‌طور کامل براساس سلسله سخنرانی‌های پل ام. سلزر^۱، استاد بیوشیمی انستیتوی بین‌رشته‌ای بیوشیمی در دانشگاه ایبرهارد-کارلس Tübingen آلمان و همچنین چندین دوره آموزشی در مجامع بین‌المللی در چارچوب برنامه‌های EU FP7 و برنامه‌های افق ۲۰۲۰ بازنگری شده است. این کتاب از این نظر که شامل تمرین‌ها و پاسخ سؤال‌های آن‌ها می‌شود و لذا آن را برای استفاده در کلاس درس مناسب می‌کند، منحصربه‌فرد است. با توجه به موفقیت چشمگیر چاپ اول نسخه آلمانی کتاب از سال ۲۰۰۴ و متعاقباً موفقیت چشمگیر بین‌المللی چاپ اول آن به زبان انگلیسی از سال ۲۰۰۸، نویسندگان تصمیم گرفتند نسخه دوم آلمانی و انگلیسی را در کنار یکدیگر منتشر کنند. کار با همان تیم اولیه یعنی سه نگارنده کتاب براساس سال‌ها تجربه در زمینه تحقیق و توسعه در صنعت داروسازی، به‌ویژه در حوزه بیوانفورماتیک و کمواینفورماتیک و قبل از اینکه به سمت موقعیت‌های شغلی مختلف بروند تا مهارت و تجربیات علمی و صنعتی خود را گسترش دهند، صورت گرفت. هدف از تهیه این کتاب، معرفی کاربردهای متداول انواع ابزارهای بیوانفورماتیک و

1- Paul M. Selzer

مروری بر حوزه‌های پیچیده مرتبط است. هدف، تشریح یا حتی ارائه فرمول یا الگوریتم‌ها نیست، بلکه تسهیل دسترسی سریع و ساختاری به بیوانفورماتیک کاربردی به دانشجویان و محققان علاقه‌مند به حوزه بیوانفورماتیک است. بنابراین، اطلاع عمیق از برنامه‌نویسی کامپیوتری برای فهم یا استفاده از مطالب این کتاب لازم نیست.

هریک از هفت فصل کتاب موضوعات مهم در حوزه بیوانفورماتیک کاربردی را شرح داده و منابع و لینک‌های اینترنتی را در اختیار قرار می‌دهد. تمرین‌ها و پاسخ سؤال‌های مفصل آن برای ترغیب خواننده به تمرین و یادگیری موضوع است و سبب افزایش مهارت و تجربه در استفاده از نرم‌افزار مرتبط می‌شود. هر جا مقدور بوده، تمرینات به گونه‌ای انتخاب شده‌اند که مثال‌هایی از توالی پروتئین یا نوکلئوتیدی قابل‌جابه‌جایی باشند. این امر به خوانندگان این امکان را می‌دهد که براساس درک صحیح از مبانی اصلی، مثال‌هایی را انتخاب کنند که به اهداف علمی آن‌ها نزدیک‌تر باشد. لازم است کاربر، به‌صورت متنی یا با فشار دادن دکمه‌ها اطلاعات را وارد کند. درنهایت، این کتاب با یک واژه‌نامه مفصل از تعاریف عمومی و اصطلاحات مورد استفاده در بیوانفورماتیک کاربردی به پایان می‌رسد.

نگارندگان از همکار سابق در چاپ اول کتاب، دکتر آندریاس روور، برای کمک‌های وی که همچنان در چاپ دوم از اهمیت ویژه‌ای برخوردار است تشکر می‌کنند. از خانم کریستین اهرت و خانم لینا هومبک — TU Dortmund، آلمان — بابت مطالعه دقیق کتاب و بررسی همه تمرین‌ها و پاسخ سؤال آن‌ها بسیار سپاسگزاریم. مایلیم از دکتر ساندرا نواک به‌خاطر کمک‌های سازنده‌اش تشکر کنیم. در پایان، از خانم استفانی گرگ و خانم سابین شوارتز از انتشارات اشپرینگر برای حمایت مستمرشان در چاپ نسخه دوم تشکر می‌کنیم.

پل ام. سلزر (انگلهایم راین، آلمان)

ریچارد جی. ماروفر (ورمز، آلمان)

الیور کوخ (دورتموند، آلمان)

می ۲۰۱۸

گردش اطلاعات ژنتیکی

اطلاعات ژنتیکی با الفبای چهارحرفی رمزگذاری می‌شوند که آن‌ها هم خود به الفبای بیست‌حرفی پروتئینی ترجمه می‌شوند. پروتئین‌ها به‌صورت ساختارهای سه‌بعدی درمی‌آیند که وظایف مهمی در موجودات تک‌سلولی یا چندسلولی به‌عهده دارند. این موجودات به‌طور مداوم تحت فشار انتخابی قرار دارند که سبب تغییر در اطلاعات ژنتیکی آن‌ها می‌شود.

تاریخچه مختصری از بیوانفورماتیک

اولین الگوریتم، برای مقایسه توالی‌های پروتئین یا DNA را نیدلمن و وانچ^۱ در سال ۱۹۷۰ (فصل ۳) منتشر کردند. بنابراین، بیوانفورماتیک فقط یک سال جوان‌تر از ARPANET^۲ یا اینترنت اولیه و یک سال قدیمی‌تر از پست الکترونیکی است که ری تاملینسون^۳ در سال ۱۹۷۱ ابداع کرد. البته اصطلاح بیوانفورماتیک در سال ۱۹۷۸ (Hogeweg, 1978) ارائه و به‌صورت «مطالعه فرایندهای اطلاعاتی در سیستم‌های زنده یا بیوتیک» تعریف شد. بانک اطلاعات پروتئینی Brookhaven (PDB^۴) هم در سال ۱۹۷۱ تأسیس شد. PDB یک پایگاه داده برای ذخیره داده‌های کریستالوگرافی پروتئین‌هاست (فصل ۷). توسعه بیوانفورماتیک در ابتدا به‌آرامی آغاز شد تا زمانی که توالی کامل ژن ویروس باکتریوفاژ φX174 در سال ۱۹۷۷ منتشر شد (Sanger *et al.*, 1977). اندکی پس از آن (۱۹۸۰) IntelliGenetics Suite، اولین بسته نرم‌افزاری برای تجزیه و تحلیل توالی‌های DNA و پروتئین، استفاده شد. در سال بعد، اسمیت و واترمن^۵ یک الگوریتم دیگر برای مقایسه توالی منتشر کردند و IBM اولین کامپیوتر شخصی را به بازار عرضه کرد (فصل ۷). در سال ۱۹۸۲، یک گروه از دانشگاه ویسکانسین – گروه کامپیوتر ژنتیک^۶ – یک بسته نرم‌افزاری برای زیست‌شناسی مولکولی به نام Wisconsin Suite عرضه کردند. هردو نرم‌افزار ابتدایی یعنی IntelliGenetics Suite و Wisconsin Suite بسته‌هایی از برنامه‌های واحد و نسبتاً کوچک بودند که از طریق خط فرمان^۷ کنترل می‌شدند. بعدها یک رابط گرافیکی برای Wisconsin Suite ارائه شد که سبب اجرای راحت‌تر دستورات در برنامه شد. از آن زمان IntelliGenetics suite از بازار حذف شد، اما Wisconsin Suite با نام GCG تا سال ۲۰۰۰ در دسترس بود.

1- Needleman and Wunsch

۲- ARPANET: مخفف سازمان پروژه‌های تحقیقاتی پیشرفته شبکه (به انگلیسی: Advanced Research Projects Agency Network). اولین شبکه سوئیچینگ بسته‌های اطلاعات در دنیا بود و پس از مدتی تبدیل به اینترنت جهانی شد.

3- Ray Thomson

4- Brookhaven Protein Data Bank

5- Smith and Waterman

6- Genetics Computer Group

7- command line

معرفی واکنش زنجیره‌ای پلیمراز (PCR) که مولیس و همکارانش در سال ۱۹۸۶ انجام دادند، نقطه عطفی در زیست‌شناسی مولکولی و همین‌طور بیوانفورماتیک بود (Mullis *et al.*, 1986). در همان سال، پایگاه داده SWISS-PROT تأسیس شد و توماس رادریک^۱ اصطلاح ژنومیک^۲ که حوزه توالی‌یابی و توصیف کل ژنوم را دربرمی‌گرفت، معرفی کرد (Kuska, 1998). دو سال بعد، مرکز ملی بیوتکنولوژی اطلاعات (NCBI)^۳ تأسیس شد که امروزه یکی از مهم‌ترین پایگاه‌های داده اولیه را مدیریت می‌کند (شکل ۱؛ به فصل ۲ نگاه کنید). در همان سال شروع Human Genome Initiative و انتشار الگوریتم FASTA اتفاق افتاد (فصل ۳). در سال ۱۹۹۱، CERN پروتکل‌هایی را ارائه کرد که وب‌سایت جهانی را در دسترس قرار می‌داد.^۴

وب برای اولین بار امکان دسترسی آسان به ابزارهای بیوانفورماتیک را فراهم کرد. البته چندین سال طول کشید تا این ابزارها به‌طور واقعی در دسترس قرار گیرد. همچنین، در سال ۱۹۹۱، گرگ ونتر^۵ استفاده از برچسب‌های توالی بیان‌شده یا EST را منتشر کرد (فصل ۴). سال بعد ونتر و همسرش کلر فریسر^۶ مرکز تحقیقات ژنومیک^۷ (TIGR) را تأسیس کردند. با انتشار GeneQuiz در سال ۱۹۹۴، از یک نرم‌افزار تجزیه و تحلیل جامع یکپارچه رونمایی شد که در سال ۱۹۹۶، در پروژه GeneCrunch برای اولین تجزیه و تحلیل خودکار بیش از ۶۰۰۰ پروتئین مخمر نان (*Saccharomyces cerevisiae*) استفاده شد (گوفو و همکاران، ۱۹۹۶). در همان سال، راه‌اندازی پایگاه اطلاعاتی Prosite اعلام شد (فصل ۲). یک سال پس از اجرای موفقیت‌آمیز بسته GeneQuiz برای تجزیه و تحلیل خودکار توالی، LION Biosciences AG در Heidelberg آلمان تأسیس شد. اساس و مبنای یکی از محصولات اصلی LION Biosciences، به نام بسته تجزیه و تحلیل توالی یکپارچه bioSCOUT، GeneQuiz بود. LION Biosciences AG همراه سایر بسته‌های سیستم فراخوانی توالی (SRS)^۸ به سرعت به یک شرکت بیوانفورماتیک بسیار موفق جهانی تبدیل شد. با این حال، این تفوق خیلی به طول نینجامید و در سال ۲۰۰۶ بخش Bioinformatics به BioWisdom فروخته شد که همچنان به بهینه‌سازی و فروش SRS ادامه می‌دهد. در آن زمان، SRS قطعاً یکی از مهم‌ترین سیستم‌های نمایه‌سازی یا ایندکسینگ و مدیریت پایگاه داده‌های فایل تخت^۹ بود. اهمیت SRS در سال‌های اخیر پیوسته کاهش یافته است؛ با وجود این، هنوز هم برنامه‌های نصبی‌ای در وب یافت می‌شوند که از این سیستم استفاده می‌کنند.

1- Thomas Roderick

2- genomics

3- National Center for Biotechnology Information

4- <https://home.cern/topics/birth-web>. <https://timeline.web.cern.ch/timelines/The-birth-of-the-World-Wide-Web>

5- Greg Venter

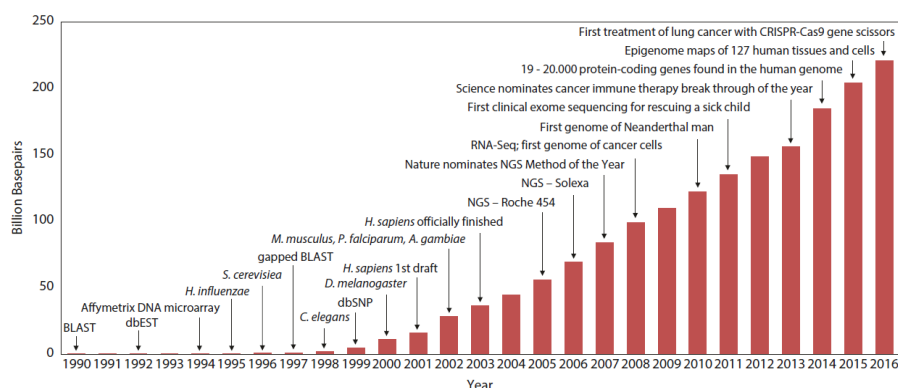
6- Claire Fraser

7- The Institute for Genomics Research

8- Sequence-Retrieval System

9- flat file databases

گردش اطلاعات ژنتیکی □ ض



شکل ۱ توسعه پایگاه داده بانک ژن NCBI و انطباق آن با برخی از نقاط عطف بیوانفورماتیک.

بیست سال بعد از اینکه اصطلاح بیوانفورماتیک ارائه شد، اصطلاح دیگری به نام chemoinformatics مطرح شد (براون، ۱۹۹۸). تا آن زمان، اصطلاحات chemometrics، شیمی رایانه^۱ و شیمی محاسباتی^۲ رایج بود و هنوز هم استفاده می‌شوند. اصطلاح chemoinformatics که گاهی cheminformatics هم نامیده می‌شود به صورت یک اصطلاح کلی استفاده می‌شود که گاهی اوقات حتی شامل واژگان دیگر مانند مدل‌سازی مولکولی یا molecular modeling می‌شود. توجه داشته باشید که برخی هنوز از این اصطلاح فقط برای نشان دادن و کار با ساختارهای شیمیایی در پایگاه‌های داده استفاده می‌کنند.

دهه ۱۹۹۰ شاهد نقاط عطف دیگری در بیوانفورماتیک و زیست‌شناسی مولکولی بود. ژنوم‌های سه موجود مدل مهم منتشر شد: *Haemophilus influenzae* (فلایشن و همکاران، ۱۹۹۵)، *S. cerevisiae* (1996) و *Caenorhabditis elegans* (C. Elegans Sequencing Consortium, 1998). همچنین، در سال ۱۹۹۸، گرگ ونتور کمپانی Celera را تأسیس کرد و در سال ۲۰۰۰ ژنوم دو موجود مدل دیگر، *Arabidopsis thaliana* و *Drosophila melanogaster* را هدف قرار داد. سال بعد، اولین پیش‌نویس ژنوم انسان منتشر شد و به صورت رسمی اعلام شد که در سال ۲۰۰۳ کامل می‌شود. در سال ۲۰۰۲، سه مؤسسه مهم، مؤسسه بیوانفورماتیک اروپایی^۳ (EMBL-EBI)، مؤسسه بیوانفورماتیک سوئیس^۴ (SIB) و منبع اطلاعات پروتئین^۵ (PIR)، کنسرسیوم UniProt را تأسیس کردند و پایگاه داده‌های خود شامل Swiss-Prto، TrEMBL و PIR-PSD را در پایگاه داده UniProt (فصل ۲ را

1- computer chemistry

2- computational chemistry

3- European Bioinformatics Institute

4- Swiss Institute of Bioinformatics

5- Protein Information Resource

ببینید) ادغام کردند. در همان سال انتشار ژنوم موش (*mus musculus*)، ژنوم عامل مالاریا در انسان (*Plasmodium falciparum*) و ناقل آن یعنی پشه *Anopheles gambiae* اتفاق افتاد. مدت کوتاهی پس از آن، در سال ۲۰۰۴، ژنوم موش قهوه‌ای (*Rattus norvegicus*) منتشر شد و ژنوم شامپانزه (*Pan troglodytes*) در سال ۲۰۰۵ منتشر شد. توالی‌یابی ژنوم‌های دیگر در حال انجام است و فهرست کردن همه آن‌ها فراتر از محدوده این بررسی اجمالی است. خلاصه‌ای از پروژه‌های ژنومی تکمیل‌شده و در حال انجام در پایگاه داده GOLD (Genome OnLine Database) قابل بررسی است.^۱

در سال ۲۰۰۵، توالی‌یابی ۴۵۴ – اولین روش توالی‌یابی نسل بعدی (NGS)، به فصل ۴ نگاه کنید) – ارائه شد و مدت کوتاهی بعد یعنی در سال ۲۰۰۶ توالی‌یابی نسل دوم به نام Solexa ارائه شد. NGS را حدوداً یک سال بعد مجله *Nature Methods* نام‌گذاری کرد. یک سال بعد، در سال ۲۰۰۸، RNA-Seq که مبتنی بر NGS است، معرفی و منجر به تشکیل رشته‌های جدیدی مانند proteogenomics و pharmacogenetics شد (فصل ۴). NGS نقش مهمی در پزشکی هم دارد که از آن به‌طور گسترده در زمینه پزشکی شخصی استفاده می‌شود. البته سرویس‌های وب جدید و پایگاه‌های داده جدید، دائم و پیوسته برای اهداف بسیار تخصصی منتشر می‌شوند. فهرست کردن همه این اهداف موضوعی است که بسیار فراتر از حوصله این کتاب است. باین‌حال، فهرست جامعی از پایگاه‌های داده، یک‌بار در سال در شماره ژانویه مجله *Nucleic Acids Research* (شماره پایگاه‌های داده) منتشر می‌شود و لیستی از خدمات وب هم یک‌بار در سال در جولای منتشر می‌شود (شماره نرم‌افزاری): NAR.^۲

منابع

- Brown (1998) Chemoinformatics: what is it and how does it impact drug discovery. *Annu Rep Med Chem* 33:375–384.
- C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018.
- Fleischmann et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.
- Goffeau et al. (1996) Life with 6000 genes. *Science* 274:546–567.
- Hogeweg (1978) Simulation of cellular forms. In: Zeigler BP (ed) *Frontiers in system modelling*. Simulation Councils, Inc., pp 90–95.
- Kuska (1998) Beer, Bethesda, and biology: how “genomics” came into being. *J Nat Cancer Inst* 90:93.
- Mullis et al. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 51(Pt 1):263–273.
- Sanger et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687–695.

1- <http://www.genomesonline.org>

2- <https://nar.oxfordjournals.org>