

راهنمای عملی
داده‌پردازی زیستی
و پروژه‌های ژنوم

محمد علی سلیمی *

امیر شیمی

تجهیزه لهراسی *

* اعضای هیئت علمی پژوهشگاه ملی مهندسی ژنتیک و زیست فناوری

www.ketab.ir



فهرست

فصل اول

بانک‌های اطلاعاتی اولیه

Primary Databanks ۲۱

پایگاه داده‌ها چیست؟ ۲۳

تاریخچه ۲۳

انواع پایگاه داده‌ها ۲۴

پایگاه داده‌های رابطه‌ای ۲۴

پایگاه داده‌های شی‌گرا ۲۵

پایگاه داده‌ها در علوم زیستی ۲۵

پایگاه‌های اطلاعاتی اولیه داده‌های زیستی ۲۷

پایگاه‌های اطلاعاتی اولیه توالی DNA ۲۸

معرفی GenBank ۳۰

روش‌های تسلیم داده به بانک‌های اطلاعاتی اولیه توالی نوکلئوتیدی ۳۳

اجزای یک رکورد GenBank ۳۴

بخش عنوان یا Header ۳۴

بخش ویژگی‌ها یا Features ۳۶

توالی یا Sequence ۳۷

www.ketab.ir



۳۷	سایر انواع ژنومهای GenBank
۳۸	۱- داده‌های EST
۳۹	۲- داده‌های GSS
۴۰	۳- داده‌های STS
۴۰	۴- داده‌های HTG
۴۱	۵- داده‌های dbSNP
۴۱	۶- داده‌های HTC
۴۱	بانک‌های داده‌های RefSeq
۴۳	بانک‌های اطلاعاتی اولیه نوکلئیدی پروتئینی
۴۳	۱- بانک‌های اطلاعات پروتئینی SwissProt
۴۸	۲- بانک‌های اطلاعات پروتئینی PIR (Protein Information Resource)
۴۹	UniProt
۵۴	iProClass
۵۷	PIRSF (Protein Information Resource-Specific Biological Functions)
۶۱	بانک‌های داده‌های نوکلئیدی پروتئینی (PIR-PSD)
۶۲	بانک‌های داده‌های ساختار فضایی پروتئین‌ها
۶۲	PDB (Protein Data Bank)
۶۶	سامانه‌های بازیابی اطلاعات از بانک‌های داده‌های زیستی
۶۶	۱- Entrez
۶۷	نحوه جستجو در Entrez
۷۰	حجوهی متنی نوکلئیدی نوکلئیدی
۷۵	قالب‌های (Formats): حاشی نوکلئیدی
۷۶	حجوهی متنی نوکلئیدی پروتئینی
۷۸	۲- SRS (Sequence Retrieval System)
۸۱	نتیجه جستجو



فصل دوم

۸۳	همردیفی و جستجوی توالی
۸۵	مقایسه دو توالی
۸۸	Dot plot روش
۸۹	مقایسه همردیفی‌ها
۹۰	امتیازدهی جایگزینی‌ها
۹۳	چند اصطلاح
۹۵	امتیازدهی حذف و اضافه شدن توالی
۹۵	۱- جریمه ایجاد فاصله (Gap Opening Penalty)
۹۵	۲- جریمه طول فاصله (Gap Extension Penalty)
۹۶	انواع همردیفی
۹۶	الف- از نظر تعداد توالی
۹۶	ب- از نظر طول توالی
۹۹	جستجوی یک توالی
۹۹	BLAST چگونه کار می‌کند؟
۱۰۰	مراحل اصلی در الگوریتم BLAST
۱۰۰	انواع BLAST
۱۰۰	BLASTN
۱۰۱	BLASTP
۱۰۱	BLASTX
۱۰۱	tBLASTn
۱۰۱	tBLASTX
۱۰۲	Doolittle , Russell F
۱۰۳	روش جستجوی BLAST
۱۰۳	وارد کردن توالی به منظور جستجو در صفحه BLAST
۱۰۵	تنظیم پارامترهای پیشرفته BLAST



فصل دوم

پایگاه‌های اطلاعاتی ثانویه پروتئین‌ها

۱۱۱ Protein Secondary Databank

۱۱۲ جستجوی پروتئین‌ها در پایگاه‌های اطلاعاتی

۱۱۳ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۱۴ Protein Data Bank

۱۱۵ Protein Structure Database

۱۱۶ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۱۷ Protein Structure Database

۱۱۸ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۱۹ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۰ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۱ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۲ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۳ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۴ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۵ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۶ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۷ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۸ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۲۹ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۳۰ پایگاه‌های اطلاعاتی پروتئین‌ها

۱۳۱ پایگاه‌های اطلاعاتی پروتئین‌ها



۱۵۱	Genome Projects & Access to Their Data
۱۵۵	تهیه نقشه‌های ژنومی
۱۵۵	۱- نقشه‌های ژنتیکی
۱۵۶	۲- نقشه‌های فیزیکی
۱۵۷	تهیه کتابخانه‌های ژنومی
۱۵۸	تعیین توالی‌های ژنومی
۱۵۹	راهِبرد ۱- تعیین توالی براساس نقشه (Map-based sequencing)
۱۶۲	راهِبرد ۲- تعیین توالی براساس BAC (BAC-based sequencing)
۱۶۴	راهِبرد ۳- تعیین توالی در مقیاس بزرگ (Large-Scale Sequencing)
۱۶۵	تکمیل توالی یابی و عرضه نتایج حاصل
۱۶۶	بهره‌برداری از پایگاه‌های اطلاعاتی ژنوم
۱۶۶	۱- استفاده از نرم‌افزار Map Viewer
۱۶۷	حسّجو در Map Viewer
۱۷۴	انواع خروجی‌های برنامه Map Viewer
۱۷۴	۱- Map view
۱۷۵	۲- (SV) Sequence viewer
۱۷۶	۳- (EV) Evidence viewer
۱۷۸	۴- (seq or del) Sequence Download
۱۷۹	۵- (mm)ModelMaker
۱۷۹	۲- پایگاه ژنومی Ensemble
۱۸۳	۳- پایگاه ژنومی UCSC Genome Browser
۱۸۷	مستندسازی توالی‌های ژنومی (Genome annotation)
۱۸۷	پایگاه UniGene
۱۸۹	پایگاه RefSeq



۲۱۷.....	داده‌های توالی‌های رونوشت‌ها
۲۱۷.....	GENE INDICES پایگاه
۲۱۷.....	GENE INDICES نحوه جستجو در پایگاه
۲۲۴.....	داده‌های بیان ژن‌ها
۲۲۴.....	GEO (The Gene Expression Omnibus)
۲۲۴.....	GEO نحوه جستجو در پایگاه

فصل ششم

پروتئومیکس

۲۳۱.....	Proteomics
۲۳۵.....	جداسازی و شناسایی پروتئین‌ها
۲۳۷.....	تعیین ویژگی‌های پروتئینی
۲۳۸.....	شناسایی تغییرات پس از ترجمه
۲۳۹.....	SWISS-2DPAGE پایگاه
۲۴۰.....	SWISS-2DPAGE ساختار داده‌های ورودی پایگاه
۲۴۱.....	SWISS-2DPAGE جستجو در پایگاه
۲۴۴.....	ALDENTE پایگاه
۲۴۵.....	ALDENTE صفحه جستجو در
۲۴۷.....	ALDENTE خروجی جستجوی
۲۴۸.....	Mascot پایگاه
۲۴۸.....	Mascot جستجو در
۲۴۸.....	جستجو با استفاده از جرم مولکولی
۲۴۹.....	Mascot صفحه جستجوی
۲۵۲.....	FindMod پایگاه
۲۵۳.....	FindMod جستجو در
۲۵۴.....	FindMod صفحه جستجو در



۲۵۵..... FindMod خروجی‌های

۲۵۸..... ذخیره‌سازی

۲۵۸..... اصطلاحات و مفهدهای استفاده‌شده در آنجمن‌های پایگاه‌های بیولوژیکی

فصل هفتم

متابولمیکس و زیست‌سامانه‌ها

۲۶۱..... Metabolomics and Systems Biology

۲۶۲..... پایگاه داده‌های متابولیک

۲۶۲..... KEGG پایگاه‌های

۲۶۴..... KEGG برای

۲۶۵..... گوناوری داده

۲۶۷..... ابزارهای

۲۶۸..... تعریف داده‌ها

۲۶۹..... جستجو در مسیرهای KEGG

۲۶۹..... جستجوی در نقشه‌های بیوشیمیایی

۲۸۲..... جدولی از

۲۸۷..... کدبالوگ‌های

۲۸۹..... تحلیل‌های متابولیکی و ارتباط آن‌ها

۲۹۳..... کدبالوگ‌های

۲۹۷..... DBGET LinkDB±

۳۰۰..... پایگاه داده‌های KEGG

۳۰۷..... پایگاه مسیرهای متابولیکی ExPASy

۳۰۹..... پایگاه مسیرهای متابولیکی UniPathway

۳۱۰..... پایگاه داده‌های Reactome



www.ketab.ir



پیشگفتار

ما در دنیایی زندگی می‌کنیم که در آن حدود یک و نیم میلیون موجود زنده شناخته شده است و حدود ده هزار گونه آنها در خشکی زندگی می‌کنند. البته تعداد کل گونه‌های کره زمین ۱۰ تا ۱۰۰ برابر آن تخمین زده می‌شود. برای هر یک از این موجودات می‌توان داده‌های زیادی را جمع‌آوری نمود که شامل داده‌های ژنوم، داده‌های بیان آنها، ساخت پروتئین‌ها و توالی آنها و مانند آن می‌باشد. از پردازش این داده‌ها اطلاعات فراوانی به دست می‌آید که خود حجم زیادی دارند. مجموعه این اطلاعات آن هم برای یکی از موجودات موضوع میلیون‌ها رکورد در انواعی از بانک‌های اطلاعاتی است. چگونگی بازیابی این اطلاعات و بهره‌برداری از آنها در آموزش و پژوهش زیست‌فناوری موضوع این راهنمای عملی است. به عبارت دیگر، در این کتاب ما تنها به بخشی از رشته زیست‌رایانه (Biocomputing) می‌پردازیم که آن را بیوانفورماتیک یا داده‌پردازی زیستی (Bioinformatics) می‌نامیم. در واقع، داده‌پردازی زیستی از اشتراک فعالیت‌های مرتبط با تولید هر گونه اطلاعات زیستی (از جمله پروژه‌های ژنوم)، ایجاد پایگاه‌های اطلاعاتی و ابزارهای مربوط، تولید نرم‌افزارها و ساخت سخت‌افزارهای مورد نیاز حاصل می‌شود. بازیابی اطلاعات حاصل، تجزیه و تحلیل و



دسترسی به سامانه‌های زیستی رایانه‌ای بسیاری از طراحی‌های فناوری و بررسی اثربخشی و مخاطرات آنها با هزینه بسیار کمتر و با سرعتی بسیار بیشتر (۴ تا ۵ برابر) انجام شود. به عبارت دیگر، در مقطعی از زمان قرار گرفته‌ایم که سرنوشت زیست فناوری در آن رقم می‌خورد. بدیهی است در این حرکت بسیار پرشتاب آشنایی با داده‌پردازی زیستی و پروژه‌های ژنوم اساس ورود در این عرصه است.

در کتاب حاضر، سعی شده است با زبانی ساده مطالب به ترتیبی ارائه شود تا خوانندگان ابتدا با برخی مفاهیم اولیه، پایگاه‌های اولیه و امکانات آنها آشنا شوند. سپس به نحوه دسترسی به اطلاعات ژنومی، ماهیت آنها و استفاده از آنها خواهیم پرداخت.

www.ketab.ir

