

به نام خدا

اصول علم بیوانفورماتیک

علیرضا عباسی تهریزی

محمد قلی پور

عضو هیات علمی

عضو هیات علمی

دانشگاه پیام نور

دانشگاه پیام نور



انتشارات شمس دانش

شابک	: ۹۷۸-۶۰۰-۹۸۳۳۵-۰۰-۴
شماره کتابشناسی ملی	: ۵۲۳۵۰۲۶
عنوان و نام پدیدآور	: اصول علم بیوانفورماتیک/ علیرضا عبادی تبریزی، محمد قلی پور.
مشخصات نشر	: تهران: شمس دانش، ۱۳۹۷.
مشخصات ظاهری	: ۳۱۷ ص.: تصویر، جدول، نمودار.
یادداشت	: واژه نامه.
موضوع	: زیست انفورماتیک
موضوع	: Bioinformatics
رده بندی دیویی	: ۵۷۳/۸۰۲۸۵
رده بندی کنگره	: ۱۳۹۷ الف۶/ع۲/۳/۴۲۳۲۴ QH۳۲۴
سرشناسه	: عبادی تبریزی، علیرضا، ۱۳۵۹ -
شناسه افزوده	: قلی پور، محمد، ۱۳۵۲ -
وضعیت فهرست نویسی	: فیبا



انتشارات شمس دانش

اصول علم بیوانفورماتیک

علیرضا عبادی تبریزی - محمد قلی پور

ویراستار علمی: دکتر امیر هوشنگ تاجفر

چاپ اول: ۱۳۹۷

شمارگان: ۱۰۰۰ نسخه

قیمت: ۲۹۵۰۰۰ ریال

شابک: ۹۷۸-۶۰۰-۹۸۳۳۵-۰۰-۴

حق چاپ محفوظ است و متعلق به انتشارات شمس دانش می باشد

<http://shamse-danesh.ir/>

shamse.danesh@gmail.com

تلفن: ۰۹۳۹۱۸۸۱۳۸۸

۰۲۱-۶۶۱۹۸۹۳۳

فهرست مطالب

پیشگفتار.....	۱
۱ مقدمه.....	۳
بیوانفورماتیک: تصویر کلی.....	۶
نرم افزارهای بیوانفورماتیک: دوحیط مجزا.....	۱۰
نرم افزارهای مبتنی بر وب.....	۱۱
مهم‌ترین مبتنی بر خط فرمان.....	۱۲
اتصال دو محاسبه.....	۱۵
الگوی جدید برای نگه‌داری برنامه نویسی در علم بیوانفورماتیک.....	۱۷
پژوهش قابل تکرار در بیوانفورماتیک.....	۱۸
بیوانفورماتیک و دیگر شاخه‌های انفورماتیک.....	۲۰
پیشنهادها.....	۲۰
کتابهای پیشنهادی.....	۲۱
برخی از مفاهیم پایه در زیست‌شناسی.....	۲۱
۲ دسترسی به داده‌های توالی و منابع علمی آن.....	۳۹
معرفی پایگاه‌های داده زیستی.....	۲۴
GenBank: مشهورترین پایگاه‌های داده.....	۲۵
حجم داده‌های توالی.....	۲۵
ارگانیزم‌های موجود در GenBank.....	۲۸
انواع داده‌ها در GenBank.....	۳۱
پایگاه‌های داده‌های DNA ژنومی.....	۳۲
پایگاه‌های داده cDNA مرتبط با ژن‌های بیان شده.....	۳۳
برچسب‌های توالی بیان شده (ESTs).....	۳۴
ESTها و UniGene.....	۳۵
جایگاه‌های توالی برچسب خورده (STSs).....	۳۶
توالی‌های بررسی ژنوم (GSSs).....	۳۷

۳۸.....	توالی یابی ژنومی با توان عملیاتی بالا (HTGS)
۳۸.....	توالیهای پروتئینی
۳۹.....	مرکز ملی اطلاعات بیوتکنولوژی
۳۹.....	معرفی صفحه خانگی NCBI
۳۹.....	PubMed
۴۰.....	Entrez
۴۰.....	BLAST
۴۰.....	OMIM
۴۱.....	کد دبا
۴۱.....	رده بندی (تاکسونومی)
۴۱.....	ساختمان
۴۲.....	موسسه یونانفورمیک اروپا (EBI)
۴۳.....	دسترسى به اطلاعات: شماره های دستیابی برای برچسب گذاری و شناسایی توالیها
۴۵.....	پروژه توالی مرجع (RefSeq)
۴۷.....	پروژه توالی کدکننده اجماع (CCLE)
۴۷.....	دسترسى به اطلاعات از طریق Entrez Gene در NCBI
۴۸.....	رابطه Entrez Gene، Entrez Nucleotide و Entrez Protein
۴۸.....	مقایسه Entrez Gene و UniGene
۴۹.....	Entrez Gene و HomoloGene
۵۰.....	دسترسى به اطلاعات پایگاه دادهای پروتئینها
۵۱.....	UniProt
۵۲.....	سامانه بازیابی توالی در ExPASy
۵۳.....	دستیابی به اطلاعات: سه مرورگر اصلی ژنومی
۵۳.....	مثالی از نحوه دسترسى به داده های توالی
۵۴.....	HIV pol
۵۵.....	دسترسى به منابع علمى زیست پزشکی NLM
۵۶.....	PubMed مرکزی و دسترسى رایگان به مجلات علمى
۵۷.....	همردیفی دوگانه توالی ها

۵۸.....	اساس تکاملی همردیفی.....
۵۹.....	همولوژی توالی در مقایسه با شباهت توالی.....
۶۱.....	تشابه توالی ها در مقایسه با یکسان بودن توالی ها.....
۶۲.....	روش های همردیفی دوگانه توالی ها.....
۶۳.....	همردیف کردن کلی و همردیف کردن محلی.....
۶۴.....	الگوریتم های همردیفی توالی.....
۶۵.....	روش ماتریس نقطه ای.....
۶۹.....	روش برنامه ریزی دینامیک.....
۷۰.....	Gap penalties(gp).....
۷۱.....	برنامه ریزی دینامیک برای همردیف کردن کلی.....
۷۳.....	برنامه ریزی دینامیک برای همردیف کردن محلی.....
۷۴.....	ماتریس های مره ۵۰.....
۷۶.....	ماتریس های مره ۱۰.....
۷۷.....	ماتریسهای PAM.....
۷۹.....	ماتریس های BLOSUM.....
۸۱.....	مقایسه بین PAM و BLOSUM.....
۸۲.....	ارزیابی آماری همردیف های توالی.....
۸۵.....	۴ ابزار جستجوی اصلی همردیفی محلی (BLAST).....
۸۸.....	مراحل جستجوی BLAST.....
	استفاده الگوریتم BLAST از همردیفی محلی به عنوان راهکار
۱۰۵.....	جستجو.....
۱۰۷.....	بخش های الگوریتم BLAST : فهرست، اسکن، بسط.....
۱۱۱.....	الگوریتم BLAST : آمار و مقدار E جستجوی همردیفی محلی.....
۱۱۲.....	فهم امتیازهای خام از طریق امتیازهای بیت.....
۱۱۲.....	الگوریتم BLAST: رابطه بین مقدار E و p.....
۱۱۳.....	پارامترهای جستجوی BLAST.....
۱۱۴.....	راهکارهای جستجوی BLAST.....
۱۱۵.....	اصول جستجوی BLAST.....
۱۲۲.....	نحوه مدیریت نتایج انبوه.....

توالی یابی ژنومی با توان عملیاتی بالا (HTGS).....	۳۸
توالیهای پروتئینی.....	۳۸
مرکز ملی اطلاعات بیوتکنولوژی.....	۳۹
معرفی صفحه خانگی NCBI.....	۳۹
PubMed.....	۳۹
Entrez.....	۴۰
BLAST.....	۴۰
OMIM.....	۴۰
کتابها.....	۴۱
داده های (تاکسونومی).....	۴۱
باختمان.....	۴۱
موسسه یونفورماتیک اروپا (EBI).....	۴۲
دسترسی به اطلاعات: شماره های دستیابی برای برجسب گذاری و شناسایی توالیها.....	۴۳
پروژه توالی مرجع (RefSeq).....	۴۵
پروژه توالی کدکننده اجماعی (CCLEF).....	۴۷
دسترسی به اطلاعات از طریق Entrez Gene در NCBI.....	۴۷
رابطه Entrez Gene، Entrez Nucleotide و Entrez Protein.....	۴۸
مقایسه Entrez Gene و UniGene.....	۴۸
Entrez Gene و HomoloGene.....	۴۹
دسترسی به اطلاعات پایگاه دادهای پروتئینها.....	۵۰
UniProt.....	۵۱
سامانه بازیابی توالی در ExPASy.....	۵۲
دستیابی به اطلاعات: سه مرورگر اصلی ژنومی.....	۵۳
مثالی از نحوه دسترسی به داده های توالی.....	۵۳
HIV pol.....	۵۴
دسترسی به منابع علمی زیست پزشکی NLM.....	۵۵
PubMed مرکزی و دسترسی رایگان به مجلات علمی.....	۵۶
همردیفی دوگانه توالی ها.....	۵۷

۵۸	اساس تکاملی همردیفی
۵۹	همولوژی توالی در مقایسه با شباهت توالی
۶۱	تشابه توالی ها در مقایسه با یکسان بودن توالی ها
۶۲	روش های همردیفی دوگانه توالی ها
۶۳	همردیف کردن کلی و همردیف کردن محلی
۶۴	الگوریتم های همردیفی توالی
۶۵	روش ماتریس نقطه ای
۶۹	روش برنامه ریزی دینامیک
۷۰	Gap penalties(qp)
۷۱	برنامه ریزی دینامیک برای همردیف کردن کلی
۷۳	برنامه ریزی دینامیک برای همردیف کردن محلی
۷۴	ماتریس های همردیفی
۷۶	ماتریس های نیمه آمینو اسیدها
۷۷	ماتریسهای PAM
۷۹	ماتریس های BLOSUM
۸۱	مقایسه بین PAM و BLOSUM
۸۲	ارزیابی آماری همردیف های توالی

۴ ابزار جستجوی اصلی همردیفی محلی (BLAST)

۸۸ مراحل جستجوی BLAST

استفاده الگوریتم BLAST از همردیفی محلی به عنوان راهکار

۱۰۵ جستجو

بخش های الگوریتم BLAST: فهرست، اسکن، بسط

الگوریتم BLAST: آمار و مقدار E جستجوی همردیفی محلی

فهم امتیازهای خام از طریق امتیازهای بیت

الگوریتم BLAST: رابطه بین مقدار E و p

پارامترهای جستجوی BLAST

راهکارهای جستجوی BLAST

اصول جستجوی BLAST

نحوه مدیریت نتایج انبوه

- ۱۲۲..... نحوه مدیریت داده های بسیار کم
- ۱۲۳..... جستجوی BLAST با پروتئین چند دومینی HIV-1 pol
- ۵ ۱۲۷..... جستجوی پیشرفته پایگاه داده
- ۱۲۸..... وبگاه های تخصصی BLAST
- ۱۲۹..... سایت های BLAST مبتنی بر ارگانیسم
- ۱۲۹..... Ensembl BLAST
- ۱۳۱..... الگوریتم های اختصاصی مرتبط با BLAST
- ۱۳۱..... WUBLAST ۲
- ۱۳۲..... بسته وی بیوانفورماتیک اروپا (EBI)
- ۱۳۲..... سایت های NCBI BLAST اختصاصی
- ۱۳۳..... یافتن پروتئین هایی با خویشاوندی دور
- ۱۳۳..... BLAST برای مختص مکان (PSI-BLAST)
- ۱۴۰..... ارزیابی کارایی PSI-BLAST
- ۱۴۱..... خطاهای PSI-BLAST: مدل انحراف
- ۱۴۲..... BLAST مختص مکان معکوس
- ۱۴۲..... BLAST آغاز شده با الگو (PH-BLAST)
- ۱۴۵..... جستجوی پروتئینی: مدل های پنهان مارکوف
- ۶ ۱۵۱..... همردیفی چندگانه توالی
- ۱۵۲..... تابع امتیاز دهی یا نمره دهی
- ۱۵۳..... الگوریتم های جامع (Exhaustive)
- ۱۵۴..... الگوریتم های ابتکاری
- ۱۵۵..... روش همردیفی تصاعدی
- ۱۵۸..... نقصهای الگوریتم جامع و راه حلهای آن
- ۱۵۹..... همردیفی تکراری
- ۱۶۰..... همردیفی بلوکی
- ۱۶۰..... ویرایش نتایج همردیفی
- ۱۶۱..... تبدیل فرمت

۷	فیلورنی مولکولی و تکامل	۲۹۷
	تکامل مولکولی و فیلورنتیک مولکولی	۱۶۴
	مفروضات اصلی	۱۶۵
	اصطلاحات علمی	۱۶۶
	فیلورنی ژنی در مقابل فیلورنی گونه	۱۶۹
	اشکال نمایش درخت فیلورنتیک	۱۶۹
	چرا یافتن یک درخت حقیقی دشوار است؟	۱۷۳
	روش ترسیم درخت فیلورنتیک	۱۷۵
	برنامه ها و روش های ترسیم درخت فیلورنتیک	۱۸۲
	تکامل درخت فیلورنتیک	۲۱۳
	برنامه های فیلورنتیک	۲۱۹
۸	DNA و کروموزوم و ریوت	۲۲۱
۹	داده های نسل جدید	۲۲۷
	اصول توالی یابی نسل جدید	۲۲۸
	روشهای NGS	۲۳۴
	توالی یابی اگزوم	۲۳۵
	توالی یابی یک سکانس جدید (از نو)	۲۳۶
	توالی یابی هدفمند	۲۳۶
	توالی یابی کل mRNA ها یا کل mRNA	۲۳۸
	توالی یابی هدفمند RNA	۲۳۸
	توالی یابی Noncoding RNA و smallRNA	۲۳۹
	توالی یابی متیلاسیون	۲۳۹
	توالی یابی CHIP	۲۴۰
	نمایه سازی ریبوزوم	۲۴۰
	داده های NGS	۲۴۱
۱۰	بیان ژن	۲۴۳
	آنالیز داده های ریزآرایه ای	۲۴۳

۲۴۶	نرم افزارهای آنالیز داده‌های ریزآرایه‌ای و مجموعه‌های داده‌ای
۲۴۸	کسب داده‌های ریزآرایه‌ای از GEO و ArrayExpress
۲۴۸	تکرارپذیری آزمایش‌های ریزآرایه‌ای
۲۴۹	تفسیر عملکردی داده‌های ریزآرایه‌ای

۲۹۷	۱۱ تحلیل پروتئین و پروتئومیکس
۲۵۲	تکنولوژی آنالیز بیان پروتئین
۲۵۳	الکتروفورز دو بعدی با ژل پلی آکریل آمید (۲-D PAGE)
۲۵۴	تیین پروتئین‌ها با اسپکترومتری جرمی
۲۵۵	شناسایی پروتئین از طریق جستجوی پایگاه داده‌ها
۲۵۷	ژل الکتروفرز تمازی
۲۵۸	ریز آرایه‌ها- پروتئین (Protein Microarrays)
۲۵۹	تغییرات پس از ترجمه در پروتئین‌ها
۲۶۰	پیشگویی پل‌های نوین
۲۶۱	شناسایی تغییرات بعد از ترجمه در آنالیزهای پروتئومیک
۲۶۲	طبقه بندی پروتئین‌ها
۲۶۵	برهم کنش‌های پروتئین-پروتئین
۲۶۵	تعیین برهم‌کنشهای پروتئین-پروتئین با کمک روش‌های آزمایشگاهی
۲۶۶	پیش بینی برهم کنش‌های پروتئین-پروتئین
۲۶۶	پیش بینی برهم کنش‌های پروتئین-پروتئین بر اساس ترکیب دومین
۲۶۸	پیش بینی برهم کنش‌های پروتئین-پروتئین بر اساس مجامعت ژنها
۲۶۸	پیش بینی برهم کنش‌های پروتئین-پروتئین بر اساس هموازی توالی‌ها
۲۶۹	پیش بینی برهم کنش‌های پروتئین-پروتئین بر اساس ادغام فیلدنتیک
۲۷۱	پیش بینی برهم کنش‌های پروتئین-پروتئین با استفاده از روشهای هیبرید یا ترکیبی

۲۷۳	۱۲ مقدمه‌ای بر ژنومیکس عملکردی
۲۷۷	رابطه ژنوتیپ و فنوتیپ
۲۷۸	هشت ارگانسیم مدل ژنومیکس عملکردی
۲۷۹	باکتری Escherichia coli

۲۸۰	Saccharomyces cerevisiae مخمر
۲۸۲	Arabidopsis thaliana گیاه آرابیدوپسیس
۲۸۲	Caenorhabditis elegans نماتد
۲۸۳	Drosophila melanogaster مگس میوه
۲۸۴	Danio rerio ماهی مخطط
۲۸۴	Mus musculus موش
۲۸۵	Homo sapiens تنوع در انسان
۲۸۶	ژنومیکس عملکردی با استفاده از ژنتیک معکوس و ژنتیک مستقیم
۲۸۸	ژنتیک معکوس: غیرفعال سازی های ژنی در موش و ژن بتاگلوبین
۲۹۲	ژنتیک معکوس: غیرفعال سازی ژن ها در مخمر با استفاده از بارکدهای مولکولی
۲۹۶	ژنتیک معکوس: جهش زایی ناشی از جایگذاری تصادفی (به دام اندازی ژن)
۲۹۹	ژنتیک معکوس: جهش زایی جایگذاری در مخمر

۲۹۷	۱۳ مقدمه ای بر زیست شیمیایی (بیولوژی سیستمها)
۳۰۴	موضوعات به روز در بیوانفورماتیک
۳۰۴	بیولوژی سیستمها چیست
۳۰۵	مسیر تغییرات بیولوژی سیستمها
۳۰۷	موضوعات به روز در بیولوژی سیستمها
۳۰۷	چرا باید از بیولوژی سیستمها استفاده کنیم
۳۰۸	راهکارهای مدل سازی در بیولوژی سیستمها
۳۰۹	بیولوژی سیستمها در پروژه های تحقیقاتی علوم زیستی
۳۰۹	انواع خروجی تحقیقاتی
۳۱۱	روشها
۳۱۳	حوزه های پژوهشی مورد آنالیز
۳۱۳	ابعاد تحقیقات علوم زیستی
۳۱۳	انتخاب حوزه تحقیقات
۳۱۴	اندازه صحیح علم
۳۱۵	پژوهش مشارکتی
۳۱۵	طبیعت پژوهش مشارکتی
۳۱۷	مراجع

در گذشته‌ای نه‌چندان دور رابطه بین علم و فناوری به‌گونه‌ای بود که علم کشف می‌کرد و فناوری با اختراعات خود، اکتشافات علمی را با صنعت عجین می‌کرد. در این حیطه، مثال‌های متعددی موجود است که یکی از بارزترین آن‌ها کشف لیزر و به‌تبع آن شروع کاربرد آن در بخش‌های مختلف صنعت مثل پزشکی، نظامی را شاهد بودیم. ولی چندی است که جهت این رابطه تغییر کرده است و فناوری با سرعت زائدالوصفی در برخی از حوزه‌ها از علم پیشی گرفته است. در سال ۲۰۰۰ که پروژه ژنوم انسان تکمیل شد برای توالی‌یابی کل ژنوم انسان به زمانی حدود ۴ دهه نیاز بود و حدود ۳۰ میلیون دلار هزینه می‌شد. کار بر روی تقویت توان تکنیک‌های توالی‌یابی همچنان ادامه داشت تا اینکه تکنیکی به نام توالی‌یابی نسل جدید^۱ وارد عرصه تجاری گردید. بر اساس گزارش^۲ موسسه ملی سلامت آمریکا^۳ در سال ۲۰۱۵ توالی‌یابی کامل ژنوم انسان به کمک این فناوری نوظهور، مدت زمانی کمتر از ۲۴ ساعت و با هزینه‌ای حدود ۱۰۰۰ دلار قابل اجرا شد و این کاهش در زمان و هزینه همچنان ادامه دارد. ظهور این فناوری تازه ابتدای راه بود زیرا که در کمتر از یک روز، حجم عظیمی از داده‌های ژنومی فراهم می‌شود که باید آنالیز شود. در نظر داشته باشید که ژنوم انسان از نظر حجم معادل ۱ میلیون و پانصد هزار صفحه کتاب یا تقریباً ۳۰۰۰ جلد کتاب ۵۰۰ صفحه‌ای است. در گذشته، علم توان استخراج اطلاعات را از این حجم داده در مدت زمان کم نداشت. اینجا بود که داده‌ای و تولید رایانه‌های پرتوان همراه با نرم‌افزارهای هدفمند در حوزه ژنومیکس رونق گرفت و زمینه بیوانفورماتیک به‌عنوان زمینه‌ای متعالی در تحقیقات علوم زیستی به ابزاری منحصربه‌فرد و توانمند تبدیل شد.

بیوانفورماتیک حوزه‌ای بین رشته است که نقطه عطف تخصصی علوم زیستی (پزشکی، کشاورزی و زیست‌شناسی)، علوم کامپیوتر (سخت‌افزار و نرم‌افزار)، ژنتیک (سنتی و بیزی) و ریاضیات به شمار می‌رود. ولی اینکه قرار است یک متخصص کامپیوتر وارد حوزه بیوانفورماتیک شود یا اینکه یک متخصص علوم زیستی چنین قصدی را دارد امری است علی‌حده و قطعاً فرایند

^۱ Next Generation Sequencing (NGS)

^۲ National Institute of Health (NIH)

متفاوتی دارد. کتاب حاضر به قصد تربیت مقدماتی متخصصان کامپیوتر که می‌خواهند در حوزه بیوانفورماتیک فعالیت داشته باشند نگارش شده است.

ماهیت بیولوژی بر این اصل^۱ استوار است که ژن طی فرایند نسخه‌برداری به RNA و مولکول اخیر طی فرایند ترجمه به پروتئین تبدیل می‌شود. بنابراین هر موجود زنده دارای مجموعه‌ای از ژن‌ها به همراه توالی‌های کنترل‌کننده آن‌ها، مجموعه‌ای از RNAها و مجموعه‌ای از پروتئین‌ها می‌باشد که حوزه‌های منحصر به آن‌ها به ترتیب ژنومیکس، ترانسکریپتومیکس و پروتئومیکس^۲ نام دارد. با تأسی از این تقسیم‌بندی، علم بیوانفورماتیک هم دارای تقسیم‌بندی این‌چنینی می‌باشد و این کتاب قصد دارد به‌طور مقدماتی به معرفی تئوری و عملی بیوانفورماتیک در آن‌ها بپردازد.

از آنجایی که ژن، ترانسکریپتها و پروتئین‌ها در سلول و بدن موجود زنده به‌صورت ایزوله نبوده و این سه بخش با هم و دیگر متابولیتهای سلول روابط متقابل^۳ دارند استخراج روابط آن‌ها در حوزه بیولوژی سیستم‌ها^۴ می‌گیرد که در فصل آخر کتاب تشریح خواهد شد.

نویسنده: مهدی تبرزی - محمد قلی‌پور

۱۳۹۷

^۱ Central Dogma

^۲ Genomics, Transcriptomics & Proteomics

^۳ Interaction

^۴ Systems Biology