بسم الله الرحمن الرحیم

جهاد دانشگاهی
واحد چهارمحال و بختیاری

بیوانفورماتیک کاربردی

نویسندگان:

پاول سلرز
ریچارد مارهوفر
آلیور کوچ

مترجمان:

دکتر امین صادقی
دکتر محسن مدرس‌کیا
مهندس سحر شجاعی

بهار ۱۳۹۸

سرشناسه : سلرز، پاول ام.
Selzer, P. M. (Paul M.)
عنوان و نام پدیدآور : بیوانفورماتیک کاربردی/ نویسندگان پاول سلرز، ریچارد مارهوفر، آلیور کوچ- مترجمان امین صادقی، محسن مدرس‌کیا، سحر شجاعی.
مشخصات نشر : شهرکرد: جهاد دانشگاهی، واحد استان چهارمحال و بختیاری، انتشارات، ۱۳۹۸
مشخصات ظاهری : ۲۰۵ ص.
شابک : 978-622-720503-9
وضعیت فهرست نویسی : فیپا
یادداشت: عنوان اصلی : Applied bioinformatics: an introduction, 2th ed, c2008.
موضوع : زیست انفورماتیک—مسائل، تمرین‌ها و غیره.
موضوع : Bioinformatics—Problems, exercise, etc.
شناسه افزوده : مارهوفر، آر، ج. (دیچارد ج) Marhofer, R, J. (Richard J.)
شناسه افزوده : کوچ آلیور Koch, Oliver
شناسه افزوده : صادقی، امین، ۱۳۶۸ مهر- مترجم
شناسه افزوده : مدرس‌کیا، محسن، ۱۳۶۱ بهمن- مترجم
شناسه افزوده : شجاعی، سحر، ۱۳۷۲ آبان- مترجم
رده‌بندی کنگره : QH۳۲۴/۲
رده‌بندی دیویی : ۵۷۲/۸۰۲۸۵
شماره کتابشناسی ملی : ۶۰۷۶۵۱۵

واحد چهارمحال و بختیاری

نام کتاب: بیوانفورماتیک کاربردی
نویسندگان: پاول سلرز، ریچارد مارهوفر، آلیور کوچ
مترجمان: امین صادقی، محسن مدرس‌کیا، سحر شجاعی
ناشر: انتشارات جهاد دانشگاهی، واحد چهارمحال و بختیاری، شهرکرد
نوبت و سال چاپ: اول، ۱۳۹۸
تیراژ: ۱۰۰۰ نسخه
شابک: ۹۷۸-۶۲۲-۷۲۰۵۰۳-۹
طراحی جلد: امین صادقی
ویراستار ادبی: دکتر مهدی منصوری
قیمت: ۱۰۰۰۰ تومان

تقدیم به

پدر و مادر عزیزم

مقدمه مترجمان

با پیشرفت دانش و تجهیزات در علوم زیست‌شناسی به ویژه در علوم سلولی و مولکولی و به تبع آن با افزایش حجم عظیمی از داده‌های استخراج شده از سلول‌های موجودات مختلف، مواجه هستیم. افزایش این حجم از داده‌ها و به سبب آن نیاز به ذخیره، بازیابی و تحلیل مناسب این داده‌ها، موجب پیدایش علم بیوانفورماتیک گردید. این دانش نوظهور، به عنوان یک دانش بین‌رشته‌ای، تلاش می‌کند تا با استفاده از تکنیک‌های موجود در علوم کامپیوتر، ریاضیات، شیمی، فیزیک و علوم مرتبط دیگر، مسائل مختلف زیست‌شناختی را که معمولاً در سطح مولکولی هستند حل نماید. امروزه علم بیوانفورماتیک کاربردهای فراوانی در بخش‌های مختلف مانند:زیست‌شناسی، ژنتیک، طراحی دارو، علوم جنایی، علوم کشاورزی، دامپزشکی و پزشکی دارد.

کتاب حاضر که تحت عنوان "بیوانفورماتیک کاربردی "در سال ۲۰۱۸ در نشریه معتبر علمی اشپرینگر به چاپ رسیده است، منبع ارزشمندی برای محققین، اساتید و دانشجویان حوزه ژنتیک، پزشکی، دارویی، کشاورزی و علوم زیستی فراهم آورده است.

در ترجمه کتاب حاضر سعی بر رعایت امانت‌داری بدون دخل و تصرف در اصل موضوع و استفاده از واژگان تخصصی و ساده مطالب بوده است. با این وجود، هیچ اثری خالی از ایراد نیست، لذا از همه اساتید و دانشجویان عزیز تقاضا داریم در جهت رفع نقایص، ما را در چاپ های بعدی یاری نمایند.

با تشکر گروه مترجمان

فهرست مطالب

عنوان **صفحه**

فصل اول: مبانی بیولوژی در علم بیوانفورماتیک

فصل دوم: داده‌های زیستی

فصل سوم: جستجوی پایگاه‌ها و مقایسه‌ی توالی‌ها

فصل چهارم: رمزگشایی ژنوم یوکاریوتی

فصل پنجم: ساختارهای پروتئینی و طراحی دارو مبتنی بر ساختار

فصل ششم: تجزیه و تحلیل عملکردی ژنوم

فصل هفتم: تحلیل مقایسه‌ای ژنوم

مقدمه

تاریخچه علم بیوانفورماتیک

نخستین الگوریتم جهت مقایسه توالی‌های DNA یا پروتئین توسط دو دانشمند به نام‌های نیدلمن و وانچ[1] در سال ۱۹۷۰ میلادی منتشر شد (فصل ۳). علم بیوانفورماتیک تنها یک‌سال پس از ظهور اینترنت پیشرو ARPANET و یک‌سال قبل از ظهور نخستینE-mail، در سال ۱۹۷۱ توسط ری تاملینسون[2] ابداع شد. با این وجود، اصطلاح "بیوانفورماتیک[3]" در سال ۱۹۸۷ توسط هوگ‌وگ[4] معرفی و تحت عنوان "علم پردازش اطلاعات در سیستم‌های زیستی[5]" تعریف گردید. پایگاه داده‌های پروتئینی بروکهاون (PDB)[6] نیز در سال ۱۹۷۱ تأسیس شد. PDB یک پایگاه داده جهت ذخیره ساختار سه‌بعدی یا بلوری پروتئین می‌باشد (فصل ۲). پیشرفت علم بیوانفورماتیک در ابتدای امر با کندی بسیاری همراه بود تا اینکه در سال ۱۹۷۷ توالی کامل ژن باکتریوفاژ ϕX174 منتشر شد (Sanger et al. 1977). کمی بعد در سال ۱۹۸۰ میلادی، نرم‌افزار IntelliGenetics، نخستین بسته نرم‌افزاری جهت آنالیز توالی‌های DNA و پروتئین مورد استفاده قرار گرفت. پس از آن اسمیت و واترمن[7] الگوریتم دیگری را جهت مقایسه توالی‌ها منتشر نمودند و شرکت IBM نیز نخستین یارانه شخصی را وارد بازار مصرف نمود (فصل ۳). در سال ۱۹۸۲، گروه علوم ژنتیک در کامپیوتر در دانشگاه ویسکونسین[8]، بسته نرم‌افزاری علوم بیولوژی مولکولی را به بازار عرضه نمودند. در ابتدا، هر دو شرکت IntelliGenetics و Wisconsin بسته‌های نرم‌افزاری را طراحی می‌نمودند که فقط برنامه‌های کوچک مبتنی بر خطوط دستوری را مدیریت می‌کردند. اولین صفحه گرافیکی جهت استفاده کاربران، توسط شرکت Wisconsin توسعه یافت که کارکرد بسیار راحت‌تری جهت استفاده از برنامه‌های بیوانفورماتیکی بشمار می‌رفت. چند سال بعد شرکت IntelliGenetics از بازار فروش به کلی ناپدید شد اما شرکت Wisconsin تا دهه ۲۰۰۰ میلادی با نام جدید GCG در بازار فعالیت داشت.

---

[1] Needleman and Wunsch

[2] Ray Thomlinson

[3] Bioinformatics

[4] Hogeweg

[5] Studyof informatic processes in biotic systems

[6] Brookhaven Protein Data Bank(PDB)

[7] Smith and Waterman

[8] University ofWisconsin

انتشار نحوه عملکرد واکنش زنجیره‌ای پلیمراز (PCR)[۱] توسط مولیس و همکاران در سال ۱۹۸۶ گامی بزرگ در عرصه بیولوژی مولکولی و متعاقباً در علم بیوانفورماتیک را سبب شد (Mullis et al. 1986). در همان سال پایگاه داده SWISS-PROT احداث شد و توماس رودریک[۲] اصطلاحی جدید به نام "ژنومیکس[۳]"، که به توصیف و توالی‌یابی کلی ژنوم معطوف می‌شود را، مطرح نمود (Kuska 1998). دو سال بعد (۲۰۰۰)، مرکز ملی اطلاعات بیوتکنولوژی (NCBI)[۴] به بهره‌برداری رسید و امروزه به عنوان مهمترین پایگاه جهت استفاده کاربران واقع شده است (شکل ۱- فصل ۲). همچنین در سال ۲۰۰۰ نیز پروژه ژنوم انسان و الگوریتم FASTA نیز منتشر شدند (فصل ۳). در سال ۱۹۹۱ شرکت CERN نخستین پروتکل‌های مبتنی بر وب را منتشر نمودند و اولین صفحه وب[۵] جهت دسترسی به ابزارهای بیوانفورماتیکی در جهان رونمایی شد. با این وجود، چند سالی به طول انجامید تا بصورت واقعی چنین ابزاری در دسترس عموم کاربران قرار گیرد. علاوه بر این، در سال ۱۹۹۱ شخصی به نام گرگ ونتر[۶] استفاده از برچسب‌های توالی بیان‌شده (ESTs)[۷] را منتشر نمود (فصل ۴). یک‌سال پس از آن آقای ونتر به همراه همسرش خانم کلایر فارسر[۸] موسسه تحقیقاتی ژنومیک (TIGR)[۹] را بنا نهادند. در سال ۱۹۹۶ ابزار کامل آنالیز توالی‌های تلفیقی توسط شرکت GeneQuiz نیز وارد بازار شد و در پروژه GeneCrunch جهت اولین آنالیز خودکار بیش از ۶ هزار پروتئین مخمر نان *Saccharomyces cerevisiae* مورد استفاده قرار گرفت (Goffeau et al. 1996). در همان سال (۱۹۹۶) پایگاه داده Prosite نیز شروع فعالیت‌هایش را اعلام نمود. یک‌سال پس از موفقیت بسته آنالیز خودکار توالی شرکت GeneQuiz، شرکت بیوتکنولوژی LION Biosciences AG نیز در کشور آلمان احداث گردید. محصولات شرکت LION که اصطلاحاً با نام BioScout معروف هستند در واقع همان بسته‌های آنالیز توالی‌های تلفیقی شرکت GeneQuiz می‌باشند. شرکت LION با تولید بسته سیستم بازیابی توالی

---

[1] Polymerase Chain Reaction (PCR)
[2] Thomas Roderick
[3] Genomics
[4] National Center for BiotechnologyInformation (NCBI)
[5] https://home.cern/topics/birth-web; 7 https://timeline.web.cern.ch/timelines/The-birth-of-the-World-Wide-Web
[6] GregVenter
[7] Expressed Sequence Tags (ESTs)
[8] Claire Fraser
[9] The Institute for Genomics Research (TIGR)

(SRS)[1] به سرعت تبدیل به یکی از شرکت‌های موفق در زمینه بیوانفورماتیک در سطح جهان شد. اما این موفقیت زیاد بطول نیانجامید، بطوری که در سال ۲۰۰۰ بخش بیوانفورماتیکی آن به شرکت BioWisdom فروخته شد و این شرکت فروش بسته‌های تغییریافته SRS را ادامه داد. در طول این مدت، بسته‌های SRS به عنوان یکی از مهم‌ترین سیستم‌های شاخص‌گذاری و مدیریت فایل‌های اطلاعاتی خام بشمار می‌رفت. در سال‌های اخیر اهمیت بسته‌های SRS رو به کاهش بوده است اما هنوز برخی از تنظیمات صفحات وب را شامل می‌شود.

بیست سال پس از آغاز اصطلاحی به نام بیوانفورماتیک، اصطلاح دیگری به نام "کموانفورماتیک[2]" نیز نشر یافت (Brown1998). از آن زمان تاکنون، اصطلاحاتی مانند: کمومتریکس[3]، شیمی‌کامپیوتر[4] و شیمی محاسباتی[5] بصورت متداول مورد استفاده قرار می‌گیرند.اصطلاح کموانفورماتیک، و یا کم‌انفورماتیک، یه صورت یک اصطلاح چندمعنا بکار گرفته می‌شود و گاهی تحت عنوان مدل‌سازی مولکولی نیز مطرح می‌گردد. باید خاطر نشان شود که هنوز برخی از سنت‌گرایان از این عبارت‌ها فقط برای توصیف و بررسی ساختارهای شیمیایی استفاده می‌کنند.

در دهه ۱۹۹۰ میلادی شاهد گام‌های بزرگ دیگری در جهان در عرصه علم بیوانفورماتیک و بیولوژی مولکولی بودیم. ساختار ژنومی سه جاندار مدل نیز انتشار یافت (Fleischmann et al. 1995): هموفیلوس آنفولانزا (۱۹۹۵)[6]، مخمر نان (۱۹۹۶)[7] و نماتد الگانس (۱۹۹۸)[8]. علاوه بر این، در سال ۱۹۹۸ آقای گرگ ونتر شرکت Celera را احداث کرد و پس از آن در سال ۲۰۰۰ نیز توالی ژنوم دو جاندار مدل با نام‌های مگس سرکه[9] و گیاه آرابیدوپسیس تالیانا[10] را نیز منتشر نمود. در سال بعد (۲۰۰۱) اولین پیش‌نویس توالی ژنوم انسان نیز منتشر شد که البته نسخه رسمی آن در سال ۲۰۰۳ به شکل کامل انتشار یافت. در سال ۲۰۰۲ سه موسسه مهم به نام‌های: موسسه بیوانفورماتیک اروپا (EMB-EBI)، موسسه بیوانفورماتیک سوئیس

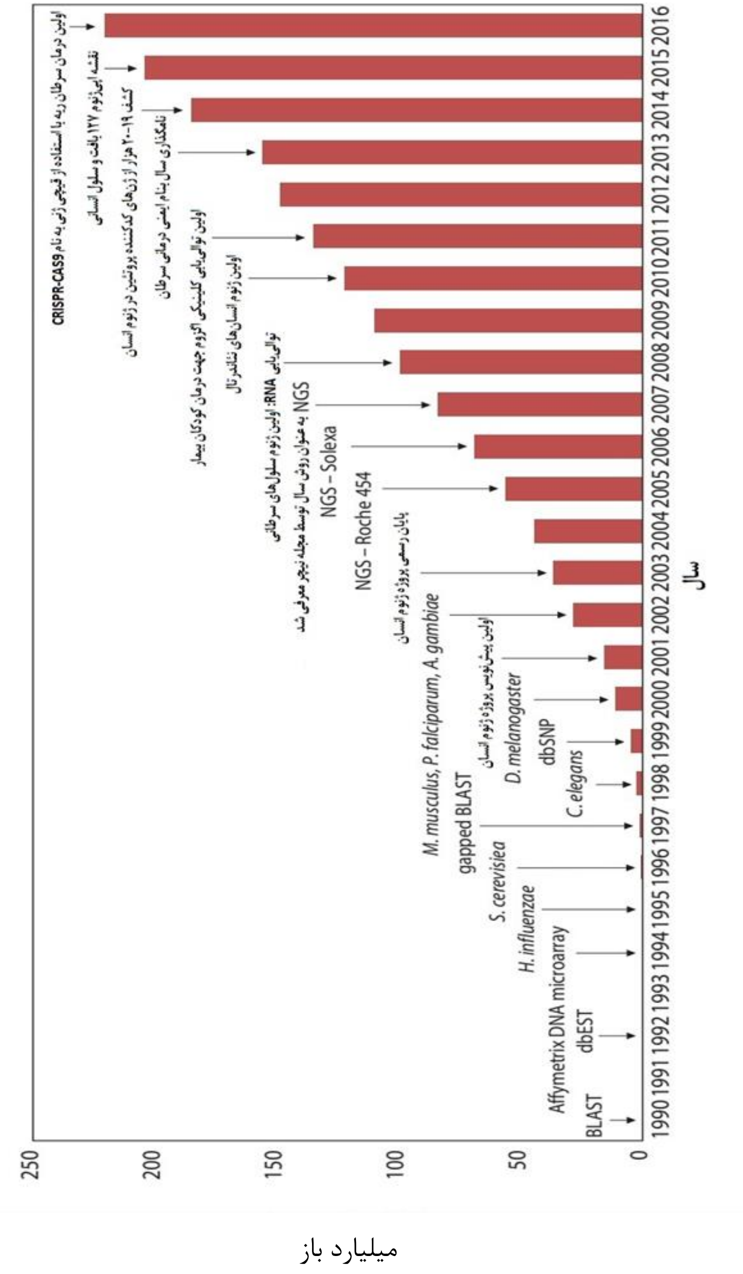
---

[1] Sequence-Retrieval System (SRS)
[2] Chemoinformatics
[3] Chemometrics
[4] Computer Chemistry
[5] Computational Chemistry
[6] *Haemophilus influenzae*
[7] *S. cerevisiae*
[8] *Caenorhabditis elegans*
[9] *Drosophila melanogaster*
[10] *Arabidopsis thaliana*

(SIB) و پایگاه منبع اطلاعات پروتئین (PIR) در کنسرسیوم UniProt تأسیس شدند و داده‌های پایگاه‌های Swiss-Prto، TrEMBL و PIR-PSD نیز در پایگاه داده UniProt تلفیق شدند (فصل ۲). در همان سال (۲۰۰۲) نیز جهان شاهد انتشار توالی ژنوم موش[۱]، مالاریای انسانی[۲] و پشه آنوفل گامبیا[۳] بود. کمی بعد در سال ۲۰۰۴ نیز ژنوم موش صحرایی قهوه‌ای[۴] نشر یافت و پس از آن در سال ۲۰۰۵ توالی ژنوم شامپانزه[۵] شناسایی شد. توالی‌یابی ژنوم سایر جانداران نیز در حال پردازش می‌باشد. جهت مشاهده سایر پروژه‌های ژنومی خاتمه‌یافته و یا در حال اجرا نیز می‌توان به پایگاه اطلاعاتی GOLD[۶] مراجعه نمایید.

---

[1] *mus musculus*
[2] *Plasmodium falciparum*
[3] *Anopheles gambiae*
[4] *Rattus norvegicus*
[5] *Pan troglodytes*
[6] Genomes OnLine Database GOLD: http://www.genomesonline.org/


اولین درمان سرطان ریه با استفاده از قیچی ژنی به نام CRISPR-CAS9
نقشه اپی‌ژنوم ۱۲۷ بافت و سلول انسانی
کشف ۱۹-۲۰ هزار از ژن‌های کدکننده پروتئین در ژنوم انسان
نامگذاری سال بنام ایمنی درمانی سرطان
اولین توالی‌یابی کلینیکی اگزوم جهت درمان کودکان بیمار
اولین ژنوم انسان‌های نئاندرتال
توالی‌یابی RNA: اولین ژنوم سلول‌های سرطانی
NGS به عنوان روش سال توسط مجله نیچر معرفی شد
NGS – Solexa
NGS – Roche 454
پایان رسمی پروژه ژنوم انسان
M. musculus, P. falciparum, A. gambiae
اولین پیش‌نویس پروژه ژنوم انسان
D. melanogaster
dbSNP
C. elegans
gapped BLAST
S. cerevisiea
H. influenzae
Affymetrix DNA microarray
dbEST
BLAST
250
200
150
100
50
0
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
سال
میلیارد باز

در سال ۲۰۰۵ نخستین تکنیک توالی‌یابی نسل بعد (NGS)[۱] به نام "توالی‌یابی ۴۵۴"[۲] عرضه شد و کمی پس از آن در سال ۲۰۰۶ روش "توالی‌یابی Solexa[3]" نیز معرفی گردید (فصل ۴). توالی‌یابی NGS به عنوان روش سال در علوم بیولوژی در مجله نیچر[۴] نام‌گذاری شد. در سال ۲۰۰۸، توالی‌یابی RNA که بر اساس روش NGS انجام می‌شد، معرفی و سبب پیدایش گرایش‌های جدیدی در علم با نام‌های فارماکوژنتیکس[۵] و پروتئوژنومیکس[۶] گردید (فصل ۴).همچنین توالی‌یابی NGSنقش مهمی را بطور گسترده در بخش پزشکی انفرادی[۷] ایفا نموده است. امروزه سرویس‌های اطلاعاتی وب و پایگاه‌های داده جدیدی نیز گسترش یافته که نام بردن تمامی این پایگاه‌ها در این کتاب خارج از عهده نویسندگان و مترجمان می‌باشد. لیست کاملی از پایگاه‌های داده بصورت یکبار در سال، در ماه ژانویه (دی ماه) در مجله *NucleicAcidsResearch* و لیست کاملی از سرویس‌های وب نیز بصورت یکبار در سال، در ماه جولای (تیر ماه) در وب‌سایت NAR[۸] منتشر می‌گردند.

منابع

1. Brown (1998) Chemoinformatics: what is it and how does it impact drug discovery. Annu RepMed Chem 33:375–384
2. *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* aplatform for investigating biology. Science 282:2012–2018.
3. Fleischmann et al. (1995) Whole-genome random sequencing and assembly of *Haemophilusinfluenzae Rd*. Science 269:496–512
4. Goffeau et al. (1996) Life with 6000 genes. Science 274:546–567.
5. Hogeweg (1978) Simulation of cellular forms. In: Zeigler BP (ed) Frontiers in system modelling. Simulation Councils, Inc., pp 90–95.
6. Kuska (1998) Beer, Bethesda, and biology: how “genomics” came into being. J Nat Cancer Inst90:93.

---

[1] Next-Generation Sequencing (NGS)

[2] 454 Sequencing

[3] Solexa Sequencing

[4] Journal *Nature Methods*

[5] Pharmacogenetics

[6] Proteogenomics

[7] Personalized Medicine

[8] https://nar.oxfordjournals.org/

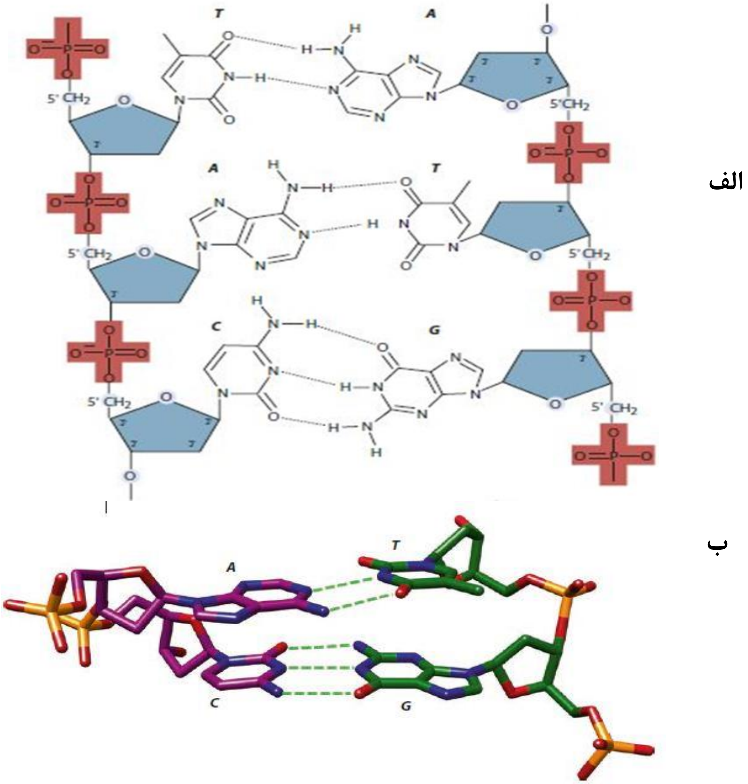
7. Mullis et al. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chainreaction. Cold Spring Harb Symp Quant Biol 51(Pt 1):263–273.
8. Sanger et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695.

فصل اول

مبانی بیولوژی در علم بیوانفورماتیک

۱-۱ نوکلئیک اسیدها و پروتئین‌ها

نوکلئیک اسیدها و پروتئین‌ها دو رده مهم از درشت مولکول‌ها هستند که نقش اساسی در طبیعت دارند و اساس حیات را تشکیل می‌دهند. دئوکسی ریبونوکلئیک اسید (DNA) حامل اطلاعات ژنتیکی است و ریبونوکلئیک اسید (RNA) در بیوسنتز پروتئین‌هایی دخیل است که فرایندهای سلولی حیات را کنترل می‌کنند. اجزای مونومری اصلی نوکلئیک اسیدها، نوکلئوتیدها هستند، در حالی که مونومرهای پروتئین‌ها، آمینواسیدها هستند.



شکل ۱- ۱) ترکیب اسیدهای نوکلئیک: الف) نمای کلی اسیدهای نوکلئیک ب) مارپیچ دوگانه و اتصال بازهای آدنین با تیمین و گوانین با سیتوزین

۲-۱ ساختار نوکلئیک اسیدهای DNA و RNA

ساختار نوکلئوتیدها در DNA و RNA یکسان است (Alberts et al. 2014). نوکلئوتیدها از یک پنتوز، یک واحد فسفریک اسید و یک باز هتروسیکلیک تشکیل شده‌اند. در یک رشته DNA یا RNA، نوکلئوتیدها از طریق پیوند شیمیایی بین قند پنتوز یک نوکلئوتید و واحد اسید فسفریک بعدی به هم متصل می‌شوند (شکل ۱-۱). طبق آن، چارچوب اصلی نوکلئیک اسیدها یک پلی‌نوکلئوتید است که اسید فسفریک یک پیوند استری بین گروه $3'$ OH واحد قند یک نوکلئوتید و گروه $5'$ OH گروه قند نوکلئوتید بعدی تشکیل می‌دهد. بنابراین در یک انتهای زنجیره پلی نوکلئوتیدی، یک گروه فسفات به اکسیژن $5'$ قند پنتوز متصل می‌شود، درحالی که در انتهای دیگر، یک گروه هیدروکسیل آزاد $3'$ حضور دارد (شکل ۱-۱).

هر واحد از ساختار ریبوز و فسفریک اسید، یک نوکلئوباز هتروسیکل دارد که به واحد قند از طریق یک اتصال N- گلیکوزیدی متصل می‌شود. نوکلئیک اسیدها از پنج باز مختلف تشکیل شده‌اند (سیتوزین، اوراسیل، تیمین، آدنین و گوانین) که اوراسیل فقط در RNA و تیمین فقط در DNA وجود دارد. نوکلئوتیدها ممکن است با استفاده از حرف اول باز مربوطه به اختصار مطرح شوند و توالی آنها توالی نوکلئوتیدی رشته نوکلئیک اسید را نشان می‌دهد. DNA و RNA نه تنها در بازها متفاوت هستند، بلکه واحد قند آنها نیز از نظر ترکیب شیمیایی متفاوت است. در RNA، قند ریبوز است در حالی که DNA از ۲- دئوکسی ریبوز تشکیل شده است.

DNA شامل دو رشته نوکلئوتیدی است که در جهت موازی ولی مخالف هم ترکیب شده‌اند بنابراین پیوندهای هیدروژنی بین بازهای هر رشته تشکیل می‌شود که منجر به ساختار نردبانی می‌شود. وقتی بازها جفت می‌شوند یک حلقه پورینی روی یک رشته با حلقه پیریمیدنی روی رشته مقابل برهمکنش می‌کنند. دو پیوند هیدروژنی بین A و T و سه پیوند هیدروژنی بین G و C وجود دارد. دو رشته نوکلئوتیدی که DNA را می‌سازند مکمل همدیگر هستند. بنابراین، توالی بازها روی یک رشته، توالی باز روی رشته دیگر را معین می‌کند. تحت شرایط فیزیولوژیک، DNA به صورت مارپیچ دو رشته‌ای وجود دارد که دو رشته پلی نوکلئوتید به صورت راست گرد دور یک محور عمومی‌می‌چرخد (شکل ۱-۲). قطر مارپیچ دو رشته‌ای ۲ نانومتر است. در طول مارپیچ دو رشته‌ای، بازهای مقابل ۰/۳۴ نانومتر از هم فاصله دارند و با زاویه ۳۶ درجه نسبت به یکدیگر چرخش دارند. ساختار مارپیچ به ازای هر ۳/۴ نانومتر به جای اول برمی‌گردد که شامل ۱۰ جفت باز می‌شود (Watson and Crick 1953a, b).

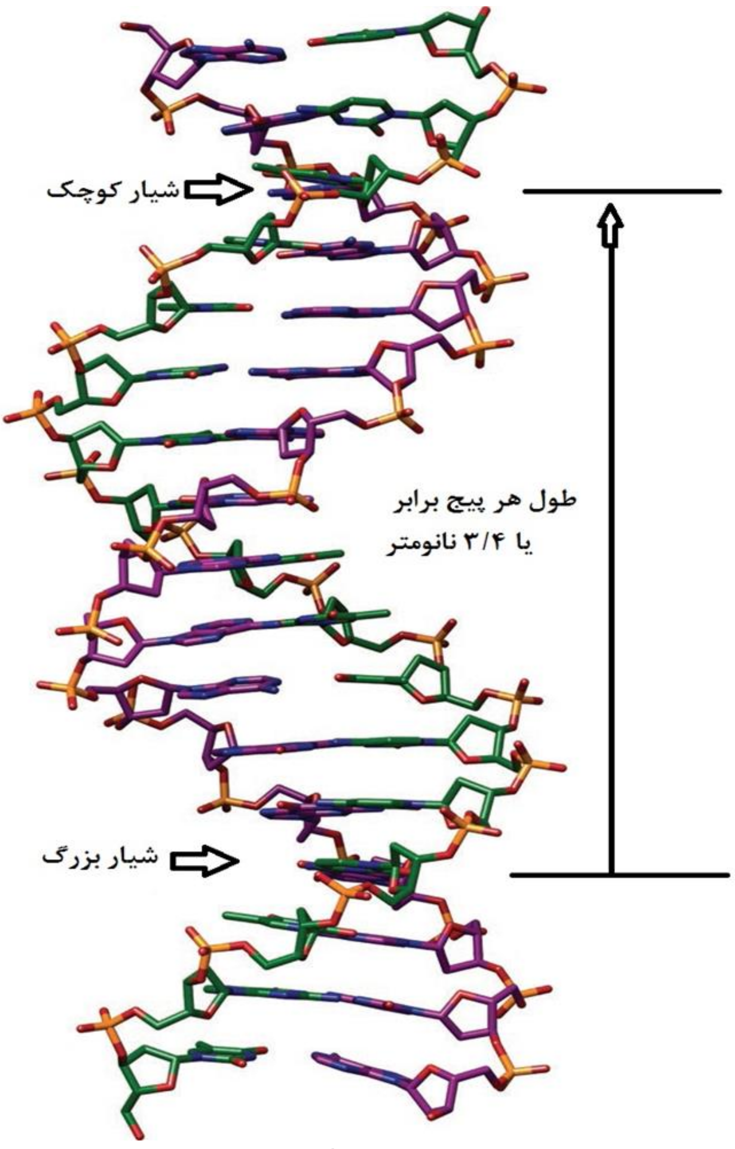
۱-۳ ذخیره اطلاعات ژنتیکی

DNA حاوی ۴ نوکلئوتید است که اطلاعات ژنتیکی را ذخیره می‌کند. توالی باز تنها عنصر متغیر روی رشته نوکلئوتید است و بنابراین، اطلاعات ضروری را رمز می‌کند تا پروتئین تولید شود. پروتئین‌ها از مقادیر متغیر تا ۲۰ آمینواسید تشکیل شده‌اند و هر آمینواسید با سه باز متوالی رمز می‌شود، که کدون نام دارد. اگر کدون‌های دو بازی استفاده شود که پروتئین را رمز کنند، ۱۶= $۴^۲$ ترکیب احتمالی حاصل نمی‌تواند ۲۰ آمینواسید را رمز کند. ازطرف دیگر، کدون‌های سه گانه ۶۴= $۴^۳$ احتمال است، که ترکیبات بیشتری را به همراه دارد که برای رمز کردن ۲۰ آمینواسید ضروری نیست. از این محاسبات نظری ممکن است تصور شود که هر آمینواسید توسط بیش از یک کدون رمز می‌شود. بنابراین، کد ژنتیکی حاصل به صورت هرز[1] تشریح می‌شود. کد ژنتیکی که در شکل ۱-۳ نشان داده شده است به صورت کلی برای همه موجودات زنده استفاده می‌شود؛ ولی، استثنائاتی در میتوکندری یافت می‌شود.

رابطه بین DNA، RNA و پروتئین به صورت اصل مرکزی زیست‌شناسی مولکولی توصیف شده است (کریک ۱۹۷۰) (شکل ۱-۴). اطلاعات ژنتیکی در DNA به صورت توالی بازهایش رمز می‌شوند. این اطلاعات به RNA پیامبر (mRNA) طی فرایند رونویسی منتقل می‌شوند، که انتقال واضح اطلاعات با جفت شدن بازهای مکمل تضمین می‌شود. فرایند نهایی ساختن پروتئین‌ها از mRNA ترجمه نامیده می‌شود. در کل، ترکیب آمینواسیدی پروتئین‌ها با اطلاعات ژنتیکی توالی DNA تعیین می‌شود. بنابراین، جریان اطلاعات عموماً از ژنوم به ترانسکریپتوم و به پروتئوم پیش می‌رود. ولی ویروس‌های RNAدار استثنا هستند. آنها با کمک یک ترانسکریپتاز معکوس، RNA خودشان را به DNA رونویسی می‌کنند و RNA را با استفاده از یک رپلیکاز همانندسازی می‌کنند. تمام DNA ژنومی‌در هر موجود به عنوان ژنوم شناخته می‌شود، و کل مخزن mRNA در هر موجودی ترانسکریپتوم نام دارد. به همین ترتیب کل مخزن پروتئین در هر موجودی پروتئوم نام دارد.

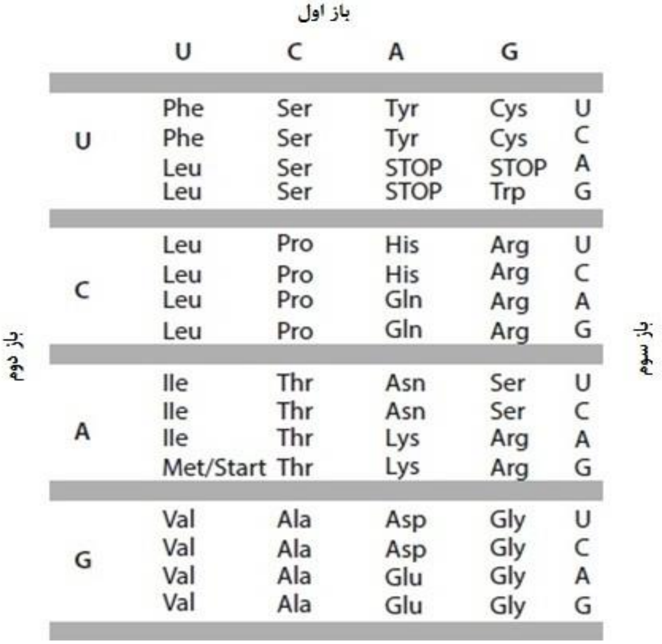
---

[1] Degenerate



شکل ۱- ۲ مشخصات مارپیچ دوگانه: فرم B در ساختار DNA حاوی شیار بزرگ و کوچک که نشان‌دهنده جفت بازها در سطح می‌باشد

بنابراین، یک ژنوم از ژن‌هایی تشکیل شده است که حاوی اطلاعاتی است که پروتئین‌ها را می‌سازند. سازماندهی یک ناحیه ژن در پروکاریوت‌ها نسبت به یوکاریوت‌ها متفاوت است (شکل ۵-۱). تفاوت برجسته این است که اطلاعات ژن پروکاریوتی در امتداد DNA رمز می‌شوند در حالی که در یوکاریوت‌ها، اگزون‌های رمز کننده توسط اینترون‌های غیررمزکننده از هم جدا می‌شوند (Krebs et al. 2014). رونویسی یوکاریوتی از DNA به mRNA بالغ (حاوی اطلاعاتی که فقط از اگزون‌ها به دست می‌آید) به چند مرحله نیاز دارد. اینترون‌ها طی فرایند splicing (ویرایش) حذف می‌شوند. در اثر Alternative splicing (حذف و اتصال اینترون‌ها و اگزون‌ها)، mRNA مختلف و در نتیجه پروتئین‌های مختلف از یک ژن حاصل می‌شوند (شکل ۴-۷). Alternative splicing یا ویرایش جایگزینی در بین سایر مکانیزم‌ها، توضیح می‌دهد که چرا تعداد نسبتاً کمتری از ژن‌ها در ژنوم انسان در مقایسه با تعداد فراوان‌تر پروتئین‌هایی که تولید شده‌اند، وجود دارند (Claverie 2001; Venter et al. 2001).



	باز اول				
باز دوم	U	C	A	G	باز سوم
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

شکل ۱- ۳) کدهای ژنتیکی

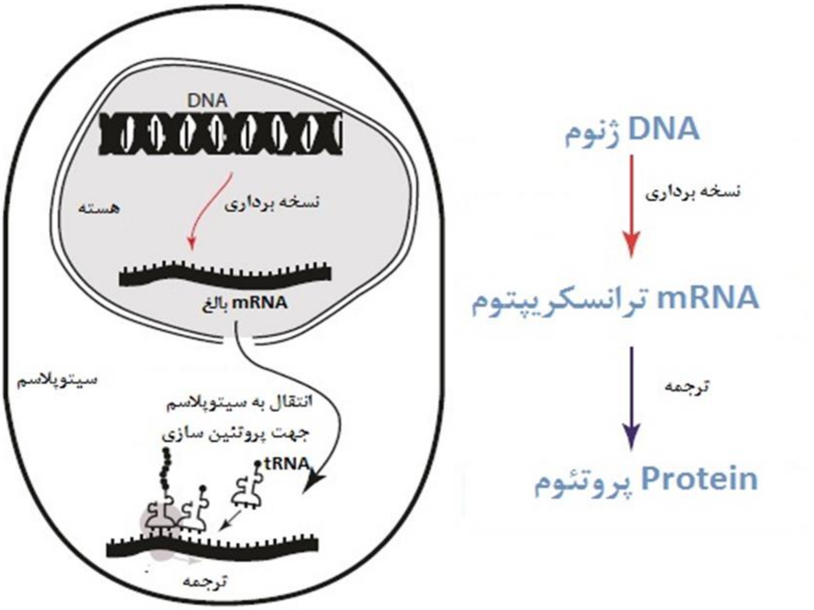
۱-٤ ساختار پروتئین‌ها

۱-٤-۱ ساختار اولیه

همانطور که اشاره شد، پروتئین‌ها درشت مولکول‌هایی هستند که از ۲۰ آمینواسید طبیعی تشکیل شده‌اند (شکل ۱-۶). ساختار اولیه توالی آمینواسید است. تحت شرایط فیزیولوژیکی، پروتئین‌ها به ساختارهای سه بعدی مشخص تا می‌خورد که بیانگر ویژگی‌ها و اعمال زیستی آنهاست (Berg et al. 2015). ترکیب عمومی آمینواسیدهای طبیعی با یک گروه آمینو و یک گروه کربوکسیل اطراف اتم α-کربن شناخته می‌شود.

زنجیره کناری مربوطه از هر آمینواسید ویژگی‌های شیمیایی مانند آب گریزی، قطبیت، اسیدی یا بازی بودن را مشخص می‌کند (شکل ۱-۷). به خاطر محدودیت فقط ۲۰ آمینواسید، پروتئین‌های غیرطبیعی (تا نخورده) ویژگی‌های بسیار مشابهی دارند که اصولاً به سطح مقطع همگن زنجیره‌های کناری پراکنده تصادفی مربوط می‌شود. ویژگی‌های مختلف پروتئین‌های عملکردی بر اساس ساختار سه بعدی پروتئین (تا خوردن) هستند. با این وجود ساختار اولیه برای تعیین ساختارهای دوم و سوم و تا خوردن سه بعدی ضروری است.

پیوندهای پپتیدی، آمینواسیدهای انفرادی را در یک زنجیره پلی پپتید به هم وصل می‌کنند. هر آمینواسید از طریق پیوند اسید گروه α-کربوکسیل به گروه α-آمینو آمینواسید بعدی، متصل می‌شود. در نتیجه، پلی پپتیدها انتهای N و C آزاد دارند. اتصال این قسمت اصلی آمینواسیدها، اسکلت پروتئینی نامیده می‌شود. ساختار اولیه پلی پپتید، یعنی توالی آمینواسید از انتهای N به C، می‌تواند بین ۳ تا چند صد آمینواسید را در بربگیرد. هر آمینواسید در زنجیره پلی پپتید با کد یک یا سه حرفی خلاصه می‌شود.

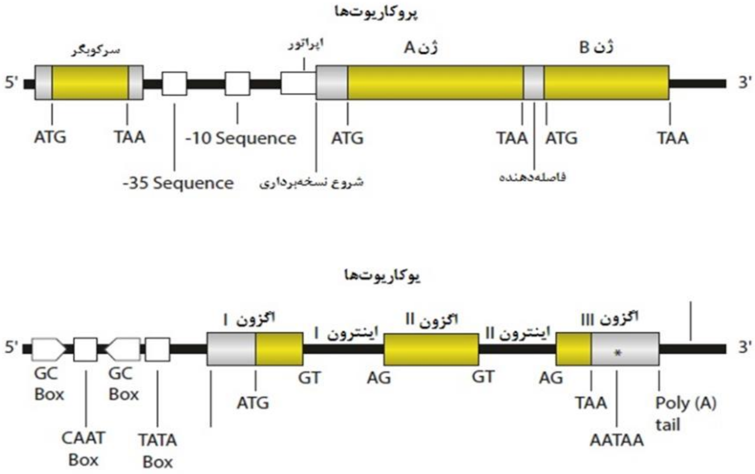


شکل ۱- ۴) اصل مرکزی در زیست‌شناسی مولکولی. جریان اطلاعات همیشه از ژنوم به پروتئوم انتقال میابد نه برعکس. استثنائاتی در ژنوم ویروس وجود دارد که بوسیله آنزیم نسخه‌برداری معکوس اطلاعات از ترانسکریپتوم شروع می‌شود و به ژنوم می‌رسد.

۱-٤-۲ ساختار دوم

ساختار دوم،شکل اسکلتی هر پلیمر را تشریح می‌کند. در مورد پروتئین‌ها، ساختار دوم الگوهای پیچش مرتب زنجیره پلی‌پپتید به صورت مارپیچ منظم (مارپیچ آلفا) و ساختارهای ورقه‌ای (رشته بتا) و چرخش‌های نامنظماست. چرخش‌ها از سه تا شش آمینواسید تشکیل شده‌اند و فضای کنفورماسیون عظیمی‌از اسکلت پروتئینی را پوشش می‌دهند. این عناصر ساختار دومسه‌تایی، بلوک‌های ساختمان الگوی پیچش سه‌بعدی پروتئین‌ها را نشان می‌دهد (Koch and Klebe 2009). حلقه‌ها عنصر ساختاری دیگری است که شامل پیچش‌های چندگانه و مارپیچ‌هاو ورقه‌های متصلمی‌شود.

کلید فهم این ساختارهای پیچیده‌تر در ویژگی‌های هندسی گروه پپتیدی قرار دارد. لینوس پاولینگ[1] و رابرت کوری[2] در دهه‌های ۱۹۳۰ و ۱۹۴۰ بیان کردند که پیوند پپتیدی یک ساختار محکم و مسطح است که به ۴۰ درصد ویژگی پیوند دوگانه پیوند پپتیدی مربوط می‌شود. بر همین اساس، یک زنجیره پلی پپتید به صورت یک زنجیره متصل مرتب از گروه‌های پپتیدی محکم و مسطح است. کنفورماسیون زنجیره پلی پپتید را می‌توان توسط زوایای پیچش اطراف اتصال Cα–N (ϕ) و اتصال Cα–C (ψ) از واحدهای آمینواسید سازنده تعیین کرد. در کنفورماسیون مسطح و کاملاً کشیده (همگی ترانس)، همه زوایا ۱۸۰ درجه است. با توجه به اتم Cα، زوایا با چرخش ساعتگرد افزایش می‌یابند. همه مقادیر متصور برای ϕ و ψممکن نیستند که اساساً به خاطر ممانعت فضایی است که در اثر زنجیره کناری آمینواسید ایجاد می‌شود. رسم راماکاندران (Ramachandran) یک جدول کنفورماسیون از مقادیری است که برای ϕ و ψ محتمل است (شکل ۲-۶). سطوح رسم راماچانداران که به مقادیر ممکن زوایای ϕ و ψ مربوط است که سطوح مجاز نامیده می‌شود؛ آن مقادیری که محتمل نیست سطوح ممنوعه نام دارند (شکل ۱-۸).



شکل ۱- ۵) محل قرارگیری ژن‌ها در پروکاریوت‌ها و یوکاریوت‌ها

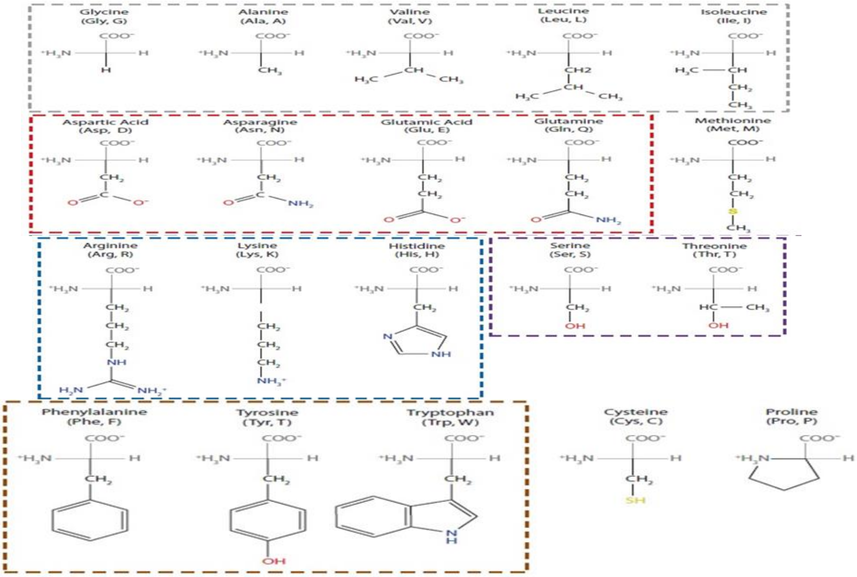
---

[1] Linus Pauling
[2] Robert Corey

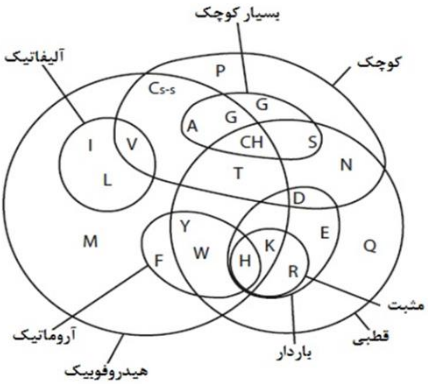
همانطور که هم اکنون اشاره شد، سه جز در ساختار دوم پروتئین‌ها مشخص شده است: مارپیچ α، رشته β و چرخش‌ها (شکل ۱-۹). زنجیره پلی پپتید یک مارپیچ α یک گام ۰/۵۴ نانومتر با ۳/۶ واحد در هر گردش را نشان می‌دهد. مانند مارپیچ α، رشته β هم توسط پیوندهای هیدروژنی تثبیت می‌شود. ولی مانند مارپیچ، پیوندها درون یک بخش منطقه‌ای زنجیره پلی پپتید یافت نمی‌شوند، بلکه بین رشته‌های همسایه قرار دارند. چنین رشته‌های β که در هر دو شکل موازی و غیرموازی قرار دارند به جهت زنجیره پلی پپتید تعلق دارند. در رشته‌های β، هر زنجیره کناری متوالی در جهت مخالف صفحه ورقه قرار دارد، که از واحدهای تکراری دو آمینواسیدی و در فواصل ۰/۷ نانومتر تشکیل شده است. به صورت میانگین، نیمی‌از پروتئین کروی از مارپیچ‌های α و نیم دیگر از رشته‌های β تشکیل شده است. بقیه پروتئین از چرخش‌های غیرتکراری تشکیل شده است. آنها مسئول کروی بودن پروتئین هستند زیرا آنها باعث حجم فراوانی از کنفورماسیون‌های مختلف می‌شوند. در کل، ۱۵۸ کنفورماسیون مختلف از اسکلت پروتئینی برای چرخش‌ها گزارش شده است (Koch and Klebe 2009).

۳-۴-۱- ساختار سوم و چهارم

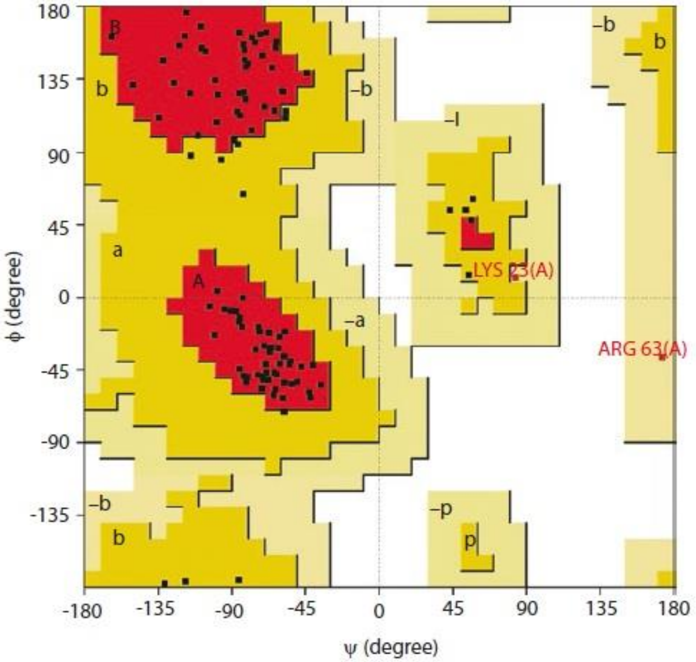
ساختار سوم آرایش سه بعدی و جاگیری عناصر ساختار دوم را تشریح می‌کند. زنجیره‌های بلند پلی پپتیدی (بیش از ۲۰۰ آمینواسید) اغلب درون واحدهای مختلف تا می‌خورند که دمین نامیده می‌شود. معمولاً چنین دمین‌هایی از ۲۰۰-۱۰۰ آمینواسید با قطر حدود ۲/۵ نانومتر تشکیل شده‌اند. ساختار سوم ویژگی‌های پروتئین را مشخص می‌کند، مثلاً اینکه پروتئین به عنوان آنزیم یا پروتئین ساختاری عمل کند. با فشردگی عناصر ساختار دوم و برهمکنش بین آمینواسیدهای آن عناصر، ساختار پروتئین تثبیت می‌شود. برهمکنش‌های آمینواسید شامل پیوندهای هیدروژنی بین گروه‌های پپتیدی، پیوندهای دی سولفیدی بین آمینواسیدهای سیستئین، پیوندهای یونی بین گروهای باردار زنجیرهای کناری آمینواسید و برهمکنش‌های آب گریز می‌شود. ساختار چهارم آرایش چند زیرواحد پلی پپتیدی است. این موارد به هندسه خاصی مربوط هستند که یک کمپلکس متقارن را تشکیل می‌دهند. مونتاژ زیرواحدهای منفرد از طریق برهمکنش‌های غیرکووالانت صورت می‌گیرد.



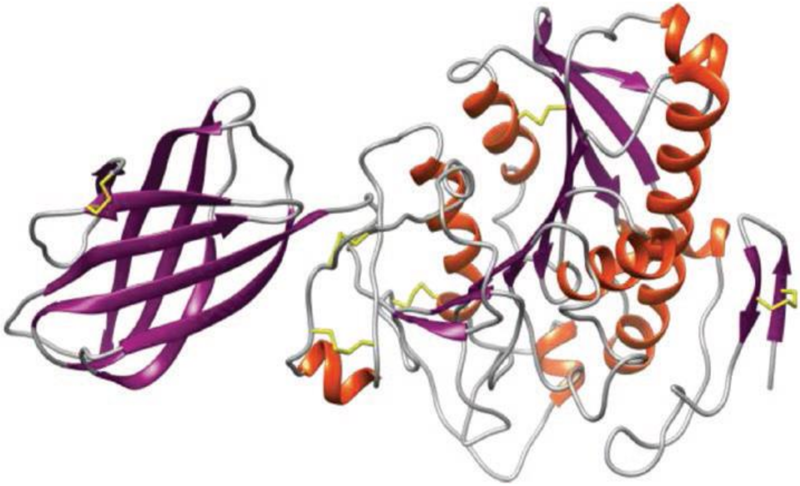
شکل ۱- ۶) اسیدهای‌آمینه به همراه کدهای سه حرفی و یک حرفی. هر قاب رنگی بیانگر اسیدهای‌آمینه با ویژگی‌های مشترک می‌باشد. زنجیره جانبی آلیفاتیک (خاکستری)، اسیدها و آمیدهای آن‌ها (قرمز)، زنجیره‌هایجانبی پایه‌ای (آبی)، با گروه هیدروکسی (قرمز) و زنجیره‌هایآروماتیک (نارنجی)



شکل ۱- ۷) نمودار ون (Venn) ویژگی‌های اسیدهای‌آمینه



شکل ۱- ۸) طرح راماکاندران پروتئین تنظیم‌کننده نسخه‌برداری GAL4 از باکتری ساکارومایسز سرویزیه. اسیدهای‌آمینه به صورت مربع‌های کوچک سیاه نشان داده شده است. تقریبا تمام اسیدهای‌آمینه در مناطق مجاز واقع شده‌اند (قرمز و زرد). دو اسیدآمینه (ARG63و LYS23) در نواحی غیرمجاز از طرح راماکاندران واقع شده‌اند. این به این معنی است که به لحاظ نظری ترکیبی از مقادیرψ و ϕ به علت منع زنجیره‌های جانبی، امکان پذیر نیست. اما در عمل، می توان مشاهده نمود.



شکل ۱- ۹) ساختار ثانویه (دوم) آنزیم لیپاز از پانکراس اسب، دارای دو دامین. مارپیچ نارنجی بیانگر مارپیچ آلفا و پیکان بنفش بیانگر صفحه بتا می‌باشد. ارتباط بین این دو گروه بوسیله ساختارهای حلقه‌ای صورت می‌پذیرد. پل دی‌سولفید بوسیله نوار زرد رنگ نشان داده شده است.

سایت‌های مفید

Amino acids.https://en.wikipedia.org/wiki/Amino_acid
Biochemistry.https://en.wikipedia.org/wiki/Biochemistry
NCBI Books.http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books
Protein structures.http://www.rcsb.org/

منابع

1. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P (2014) Molecular Biology of the Cell. Garland Science, New York.
2. Berg JM, Tymoczko JL, Gatto GJ, Stryer L (2015) Biochemistry, 8th edn. W. H. Freeman Claverie JM (2001) What if there are only 30000 human genes Science 291:1255–1256.
3. Crick F (1970) Central dogma of molecular biology. Nature 227:561–563.
4. Koch O, Klebe G (2009) Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. Proteins 74:353–367.
5. Krebs JE, Goldstein ES, Kilpatrick ST (2014) Lewins Genes XI. Jones & Bartlett Learning, Burlington.

6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291.
7. Rullmann JAC (1996) AQUA, Computer program. Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University.
8. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291:1304–1351.
9. Watson JD, Crick FHC (1953a) Molecular structure of nucleic acids. Nature 171:737–738.
10. Watson JD, Crick FHC (1953b) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967.

فصل دوم

داده‌های زیستی

۱-۲ دانش زیستی در پایگاه‌های داده جهانی ذخیره شده است

مهمترین اساس بیوانفورماتیک کاربردی، جمع‌آوری داده‌های توالی و اطلاعات زیستی مربوط به آن است. برای مثال، با توالی‌یابی ژنوم، هر روز حجم بسیار بزرگی از داده‌ها در سراسر جهان تولید می‌شوند. برای استفاده بهتر علاقه‌مندان از این داده‌ها، یک سامانه بایگانی سازمان‌یافته لازم است در دسترس باشد. هر سال، مجله تحقیقات نوکلئیک اسید[1] یک سر مقاله کامل (اولین مقاله در ژانویه) را به همه پایگاه‌های داده‌زیستی اختصاص می‌دهد که فهرست‌وار با URL مربوطه، ثبت می‌شود. به علاوه، برای تعدادی از پایگاه‌های داده، مقالات اصلی کارآیی آن‌ها را توضیح می‌دهند. مقاله این پایگاه داده، که به صورت رایگان روی وب در دسترس است، یک نقطه شروع برای کار با پایگاه‌های داده‌زیستی است. بسته به نوع داده‌های موجود، طبقه‌بندی‌های مختلف پایگاه‌های داده‌زیستی مشخص شده‌اند. پایگاه‌های داده اولیه، حاوی اطلاعات توالی اولیه (نوکلئوتید یا پروتئین) هستند و شامل اطلاعات مفصل مربوط به عملکرد، کتابشناسی، مراجع متقابل به سایر پایگاه‌های داده و غیره می‌شوند. اما پایگاه‌های داده ثانویه نتایج تحلیل‌های پایگاه‌های داده پروتئین اولیه را خلاصه می‌کنند. هدف از این تحلیل‌ها، استخراج ویژگی‌های عمومی برای رده‌های پروتئینی است، که در عوض برای طبقه‌بندی توالی‌های ناشناخته قابل استفاده است. سایر پایگاه‌های داده دیگر که اطلاعات زیستی یا پزشکی را ذخیره می‌کنند، مانند: پایگاه‌های منابع، اغلب به عنوان پایگاه داده ثانویه طبقه‌بندی می‌شوند.

با توجه اینکه برخی سامانه‌های داده رابطه‌ای تاکنون در زمینه پایگاه‌های داده‌زیستی پذیرفته نشده‌اند، اما بکارگیری این سامانه‌ها (مانند: Oracle، MS Access، Informax، DB2) و استفاده از توانایی آن‌ها در مدیریت داده‌های بزرگ، آن‌ها را برای بایگانی سازمان‌یافته، ایده‌آل ساخته است. داده‌های توالی و اطلاعات این سامانه‌ها معمولاً به شکل پایگاه‌های داده‌ی فایل‌های تک سطحی نوشته شده‌اند، یعنی، فایل‌های متن ASCII سازمان‌یافته. این مورد علل تاریخی دارد، زیرا فایل‌های حاوی متن ASCII این قابلیت را دارند که داده‌ها را بدون نیاز به سامانه‌های پیچیده پایگاه داده، تغییر دهند. همچنین فایل‌های حاوی متن ASCII تبادل داده بین دانشمندان را نیز نسبتاً ساده می‌کند. یکی از موانع در استفاده از چنین سامانه‌ها، این است که جستجو برای کلیدواژه‌های درست درون سری داده، هم دشوار و هم زمانبر است. برای به

---

[1] *Nucleic Acids Research* [nar]

حداقل رساندن این موانع، سامانه‌های مختلفی ایجاد شده‌اند که می‌توانند پایگاه‌های داده بر اساس فایل تک سطحی را فهرست‌بندی کنند، که مانند ثبت شاخص‌های یک کتاب می‌باشد، بنابراین جستجو بر اساس کلیدواژه را تسریع می‌کنند.

۲-۲ پایگاه‌های داده اولیه

۲-۲-۱ پایگاه‌های داده توالی نوکلئوتید

۲-۲-۱-۱ GenBank

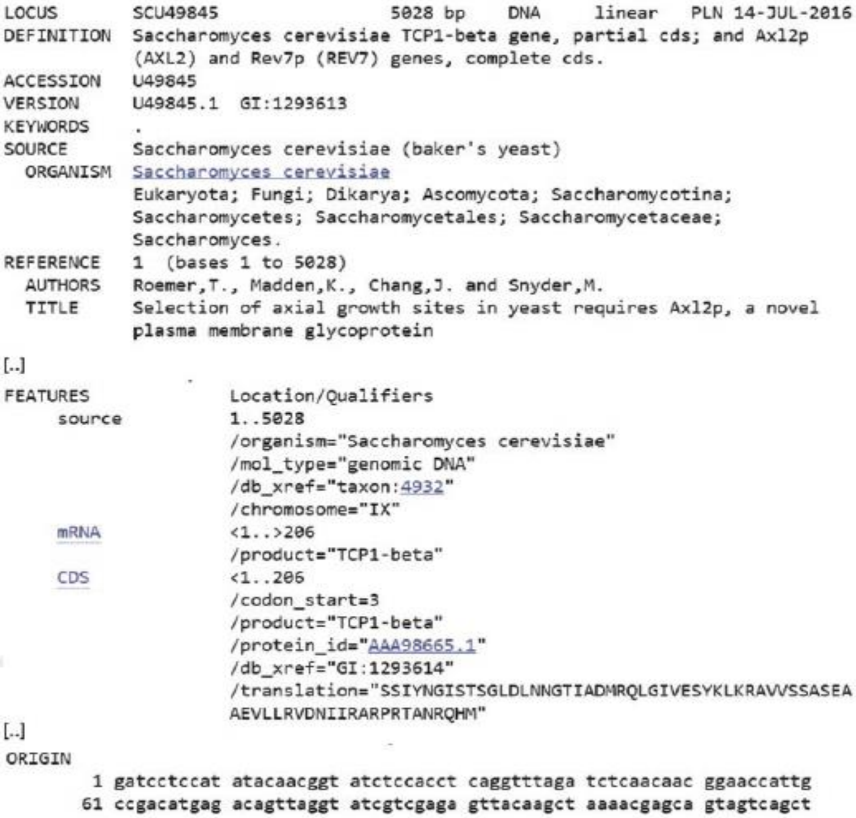
پایگاه داده GenBank شناخته‌شده‌ترین پایگاه داده توالی نوکلئوتید موجود در مرکز ملی اطلاعات بیوتکنولوژی[۱] (NCBI) آمریکاست [ncbi]. GenBank یک پایگاه داده عمومی است، که در نسخه حاضر (217.00, December 2016) حاوی تقریباً ۱۹۹ میلیون ورودی[۲] توالی است. هر شخص هنگام کار با سری‌های توالی بزرگ‌تر می‌تواند توالی‌ها را از طریق صفحه وب [bankit] یا رایانامه [sequin] به GenBank وارد کند. ورودی قبلی داده‌های توالی به GenBank یا هر کدام از پایگاه‌های داده مربوطه، برای مثال آرشیو نوکلئوتید اروپایی (ENA) یا پایگاه داده DNA ژاپن (DDBJ)، پیش نیازی برای انتشار توالی‌های جدید در هر کدام از مجلات علمی است. هر ورودی پایگاه داده با برچسب طبقه‌بندی خاصی فراهم می‌شود که شماره دسترسی (AN) نام دارد. AN[۳]یک شماره ثبت دائمی می‌باشد که حتی اگر تغییراتی در ثبت پایگاه داده ایجاد شود، بدون تغییر باقی می‌ماند. در برخی موارد، یک AN جدید را می‌توان به یک شماره موجود استناد کرد، بهعنوان مثال، اگر یک محقق ثبت داده جدیدی را به پایگاه اضافه کند، AN قدیمی به عنوان شماره ثانویه باقی می‌ماند. AN تنها روشی است که به صورت کامل ماهیت توالی یا ورودی پایگاه داده را تأیید می‌کند.

---

[1] U.S. National Center for Biotechnology Information

[2] Entery

[3] Accession Number



```
LOCUS       SCU49845                 5028 bp    DNA     linear   PLN 14-JUL-2016
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
[..]
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
     mRNA            <1..>206
                     /product="TCP1-beta"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
[..]
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
```

شکل ۱-۲) اطلاعات ثبت شده از پایگاه GenBank

شکل ۲-۱ یک ورودی GenBank را نشان می‌دهد. ورودی در بعضی نقاط کوتاه شده است و این موارد با [...] نشان داده شده است. ساختاربندی لازم ثبت پایگاه داده از طریق کلیدواژه‌های تعریف شده اجرا شده است. هر ورودی با کلید واژه LOCUS شروع می‌شود و بعد از آن نام جایگاه می‌آید. مانند AN، نام جایگاه هم منحصر به فرد است. ولی بر خلاف AN، ممکن است بعد از بازنگری پایگاه داده تغییر کند. نام جایگاه از هشت حرف تشکیل شده است که شامل اولین حرف نام جنس و گونه به علاوه شش حرف AN می‌شود. ورودی‌های جدیدتر یک AN هشت حرفی دارند. در چنین مواردی، نام جایگاه برای AN منحصر به فرد است. روی همان خط که در ادامه نام جایگاه می‌آید، طول توالی داده می‌شود. یک توالی باید حداقل ۵۰ جفت باز داشته باشد که درون GenBank وارد شود. این الزام فقط همین اخیراً وارد شده

است و بنابراین بعضی از ورودی‌های قدیمی‌تر این ویژگی را ندارند. ستون سوم به نوع مولکول ورودی توالی اختصاص دارد. هر ورودی GenBank باید حاوی اطلاعات توالی مربوط به نوع مولکول باشد، بدین معنی که یک ورودی نمی‌تواند همزمان حاوی اطلاعات توالی DNA ژنومی و RNA باشد. ستون آخر در LOCUS تاریخ آخرین تغییر ورودی را نشان می‌دهد. انتهای ثبت پایگاه داده با کلیدواژه ORIGIN شروع می‌شود. در ورودی‌های جدیدتر، این قسمت خالی مانده است. اطلاعات دقیق توالی روی خط بعدی شروع می‌شود و ممکن است حاوی خطوط فراوانی باشد. تشریح دقیق‌تر همه کلیدواژه‌ها روی صفحه نمونه GenBank یافت می‌شود [gb-sample].

جدول ۱-۲) IDهای مرتبط برای محدود کردن شرایط جستجو برخی از بخش‌های پایگاه داده Entrez

IDهای مرتبط	اطلاعات بخش
ACC	شماره دسترسی (Accession Number)
AU	نام نویسنده (Author Name)
DP	تاریخ انتشار (Publication Date)
GENES	نام ژن (Gene Name)
ORGN	نام علمی و عمومی ارگانیسم هدف
PT	نوع انتشار (مانند: مقاله مروری، گزارش علمی، خلاصه)
TA	نام مجله، وابستگی‌های رسمی، شماره ISSN

Entrez

درخواست[1] پایگاه داده GenBank از طریق سیستم EntrezNCBI انجام می‌شود [entrez]، که برای درخواست همه پایگاه‌های داده مرتبط با NCBI استفاده می‌شود (NCBI Resource Coordinators 2016). از آنجا که واژه‌های تحقیق به وسیله عملگر منطقی (AND، OR، NOT) ترکیب می‌شوند و واژه‌های تکی به زمینه‌های خاص پایگاه داده محدود می‌شوند، Entrez یک ابزار مهم و مفید برای اجرای جستجوهای هم ساده و هم پیچیده است. محدودیت واژه‌های جستجو به حوزه‌های پایگاه داده تکی عموماً از طریق ID اجرا می‌شود و

---

[1] Query

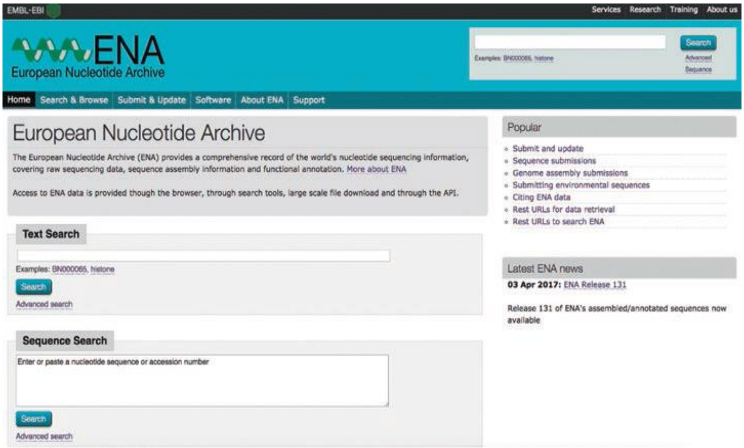
بعد از واژه: واژه جستجو [field-id] قرار گرفته‌اند. برای مثال، جستجو برای یک توالی از *Saccharomyces cerevisiae* با طول بین ۳۲۶۰ و ۳۲۷۰ جفت باز به ترکیب[1] ذیل نیاز دارد: (Saccharomyces cerevisiae[ORGN]) AND 3260:3270[SLEN]. IDهای مرتبط برای اجزای تحقیق در GenBank در جدول ۲-۱ آمده است. دستورالعمل کامل برای استفاده از Entrez روی صفحه کمک Entrez[entrez-help] آمده است. برای ساده شدن ساختار استخراج داده پیچیده، جستجوی پیشرفته ایجاد شده است. برای استفاده از این جستجو، خط زیر حوزه جستجوی Entrez را دنبال کنید. IDهای مربوطه و عملگرهای منطقی را می‌توان از جعبه‌های فهرست انتخاب کرد و درخواست مربوطه به صورت خودکار ساخته شده و به بخش جستجو وارد می‌شود. برای خوانش بهتر در این مورد، IDهای مرتبط با نام کامل وارد می‌شوند. مورد اخیر هم در جستجوی ژنریک (دارویی) کار می‌کند؛ بنابراین نیاز بیشتری به یادآوری IDهای مربوطه مخفف نیست.

---

[1] Syntax

```
ID   U49845; SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
XX
AC   U49845;
XX
DT   07-MAY-1996 (Rel. 47, Created)
DT   25-MAR-2010 (Rel. 104, Last updated, Version 5)
XX
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and
DE   Rev7p (REV7) genes, complete cds.
XX
KW   .
XX
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
XX
RN   [1]
RP   1-5028
RX   PUBMED; 8846915.
RA   Roemer T., Madden K., Chang J., Snyder M.;
RT   "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT   membrane glycoprotein";
RL   Genes Dev. 10(7):777-793(1996).
XX
RN   [2]
RP   1-5028
RA   Roemer T.;
RT   ;
RL   Submitted (22-FEB-1996) to the INSDC.
RL   Biology, Yale University, New Haven, CT 06520, USA
XX
DR   MD5; f152907ff924e11e159c909e145a77dd.
DR   Ensembl-Gn; YIL139C; saccharomyces cerevisiae.
[..]
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5028
FT                   /organism="Saccharomyces cerevisiae"
FT                   /chromosome="IX"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:4932"
FT   mRNA            <1..>206
FT                   /product="TCP1-beta"
FT   CDS             <1..206
FT                   /codon_start=3
FT                   /product="TCP1-beta"
FT                   /db_xref="GOA:P39076"
FT                   /db_xref="InterPro:IPR002194"
FT                   /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAA
FT                   EVLLRVDNIIRARPRTANRQHM"
[..]
XX
SQ   Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1609 T; 0 other;
     gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg        60
     ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct       120
     ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa       180
     gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg       240
[..]
     tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc                   5028
//
```

شکل ۲-۲) اطلاعات ثبت شده در پایگاه داده EMBL



European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA

Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.

Text Search

Examples: BN000065, histone

Sequence Search

Enter or paste a nucleotide sequence or accession number

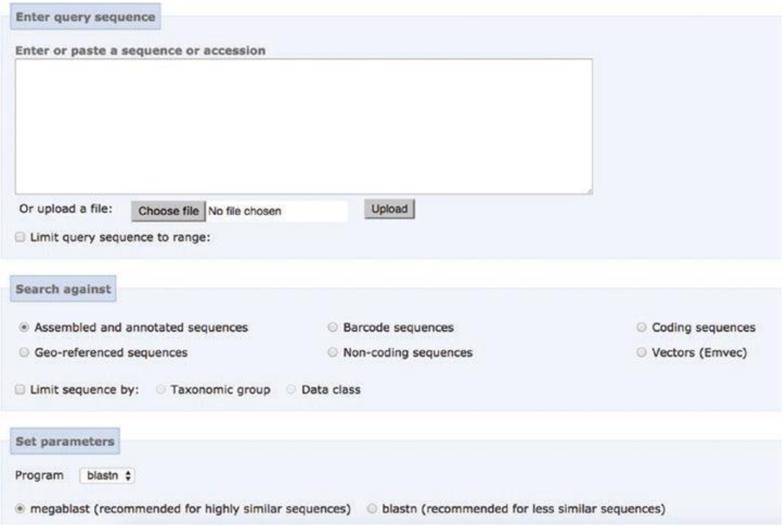
Popular
- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- Rest URLs to search ENA

Latest ENA news

03 Apr 2017: ENA Release 131

Release 131 of ENA's assembled/annotated sequences now available

شکل ۳-۲) صفحه اصلی پایگاه داده ENA به همراهبخش‌های جستجوی ساده برای بازیابی متن و توالی


Enter query sequence

Enter or paste a sequence or accession

Or upload a file: Choose file No file chosen

☐ Limit query sequence to range:

Search against

● Assembled and annotated sequences ○ Barcode sequences ○ Coding sequences

○ Geo-referenced sequences ○ Non-coding sequences ○ Vectors (Emvec)

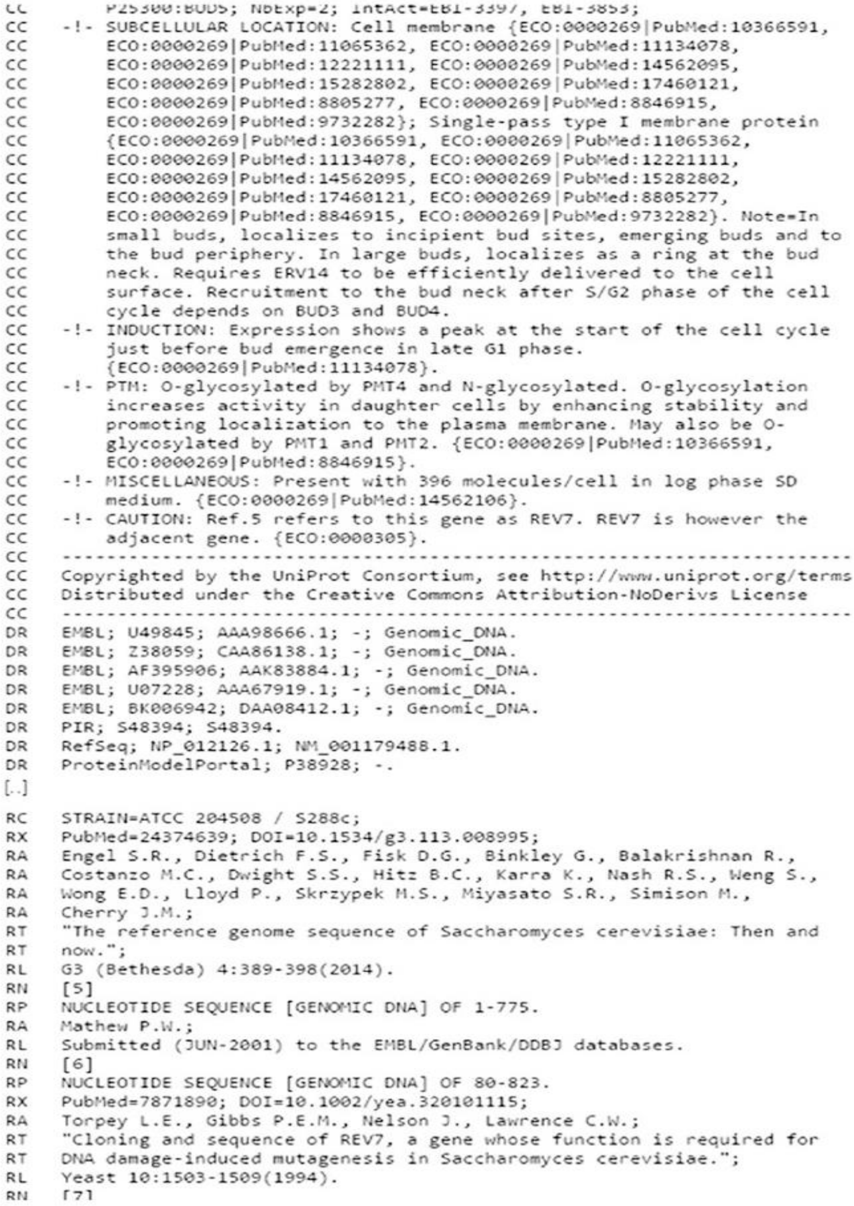
☐ Limit sequence by: ○ Taxonomic group ○ Data class

Set parameters

Program blastn

● megablast (recommended for highly similar sequences) ○ blastn (recommended for less similar sequences)

شکل ۴-۲) جستجوی پیشرفته پایگاه داده ENA



```
CC       P25300:BUD5; NbExp=2; IntAct=EBI-3397, EBI-3853;
CC   -!- SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:10366591,
CC       ECO:0000269|PubMed:11065362, ECO:0000269|PubMed:11134078,
CC       ECO:0000269|PubMed:12221111, ECO:0000269|PubMed:14562095,
CC       ECO:0000269|PubMed:15282802, ECO:0000269|PubMed:17460121,
CC       ECO:0000269|PubMed:8805277, ECO:0000269|PubMed:8846915,
CC       ECO:0000269|PubMed:9732282}; Single-pass type I membrane protein
CC       {ECO:0000269|PubMed:10366591, ECO:0000269|PubMed:11065362,
CC       ECO:0000269|PubMed:11134078, ECO:0000269|PubMed:12221111,
CC       ECO:0000269|PubMed:14562095, ECO:0000269|PubMed:15282802,
CC       ECO:0000269|PubMed:17460121, ECO:0000269|PubMed:8805277,
CC       ECO:0000269|PubMed:8846915, ECO:0000269|PubMed:9732282}. Note=In
CC       small buds, localizes to incipient bud sites, emerging buds and to
CC       the bud periphery. In large buds, localizes as a ring at the bud
CC       neck. Requires ERV14 to be efficiently delivered to the cell
CC       surface. Recruitment to the bud neck after S/G2 phase of the cell
CC       cycle depends on BUD3 and BUD4.
CC   -!- INDUCTION: Expression shows a peak at the start of the cell cycle
CC       just before bud emergence in late G1 phase.
CC       {ECO:0000269|PubMed:11134078}.
CC   -!- PTM: O-glycosylated by PMT4 and N-glycosylated. O-glycosylation
CC       increases activity in daughter cells by enhancing stability and
CC       promoting localization to the plasma membrane. May also be O-
CC       glycosylated by PMT1 and PMT2. {ECO:0000269|PubMed:10366591,
CC       ECO:0000269|PubMed:8846915}.
CC   -!- MISCELLANEOUS: Present with 396 molecules/cell in log phase SD
CC       medium. {ECO:0000269|PubMed:14562106}.
CC   -!- CAUTION: Ref.5 refers to this gene as REV7. REV7 is however the
CC       adjacent gene. {ECO:0000305}.
CC   -----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   -----------------------------------------------------------------------
DR   EMBL; U49845; AAA98666.1; -; Genomic_DNA.
DR   EMBL; Z38059; CAA86138.1; -; Genomic_DNA.
DR   EMBL; AF395906; AAK83884.1; -; Genomic_DNA.
DR   EMBL; U07228; AAA67919.1; -; Genomic_DNA.
DR   EMBL; BK006942; DAA08412.1; -; Genomic_DNA.
DR   PIR; S48394; S48394.
DR   RefSeq; NP_012126.1; NM_001179488.1.
DR   ProteinModelPortal; P38928; -.
[..]

RC   STRAIN=ATCC 204508 / S288c;
RX   PubMed=24374639; DOI=10.1534/g3.113.008995;
RA   Engel S.R., Dietrich F.S., Fisk D.G., Binkley G., Balakrishnan R.,
RA   Costanzo M.C., Dwight S.S., Hitz B.C., Karra K., Nash R.S., Weng S.,
RA   Wong E.D., Lloyd P., Skrzypek M.S., Miyasato S.R., Simison M.,
RA   Cherry J.M.;
RT   "The reference genome sequence of Saccharomyces cerevisiae: Then and
RT   now.";
RL   G3 (Bethesda) 4:389-398(2014).
RN   [5]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 1-775.
RA   Mathew P.W.;
RL   Submitted (JUN-2001) to the EMBL/GenBank/DDBJ databases.
RN   [6]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 80-823.
RX   PubMed=7871890; DOI=10.1002/yea.320101115;
RA   Torpey L.E., Gibbs P.E.M., Nelson J., Lawrence C.W.;
RT   "Cloning and sequence of REV7, a gene whose function is required for
RT   DNA damage-induced mutagenesis in Saccharomyces cerevisiae.";
RL   Yeast 10:1503-1509(1994).
RN   [7]
```

شکل ۵-۲) اطلاعات خام موجود در پایگاه داده UniProtKB/SwissProt

EMBL و DDBJ

بخش اروپایی GenBank، ENA هست [ena]، که در مؤسسه بیوانفورماتیک اروپایی (EBI) [ebi] قرار گرفته است. پایگاه داده اولیه دیگر توالی نوکلئوتید، DDBJ[ddbj]، توسط مؤسسه ملی ژنتیک (NIG) [nig] در ژاپن اجرا می‌شود و مربوط به آسیا است. سه مجری پایگاه داده، NCBI، EBI و NIG، مشارکت بین المللی پایگاه داده توالی نوکلئوتیدی[1] را تشکیل می‌دهند و در این پایگاه‌ها داده‌ها هر ۲۴ ساعت با هم منطبق می‌شوند. هر داده از همه سه پایگاه داده قابل استخراج است و بنابراین لازم نیست که هر توالی جدید نوکلئوتید را به همه سه پایگاه داده وارد کنیم.

در حالی که فرمت پایگاه داده DDBJ مشابه NCBI است، ENA تا حدودی متفاوت است. شکل ۲-۲ یک ورودی در پایگاه داده EMBL را نشان می‌دهد. مشخص‌ترین تفاوت، استفاده از کدهای دو حرفی بجای کلیدواژه کامل است. همچنین، تغییرات جزئی در سازماندهی بخش‌های هر داده وجود دارد. توصیف کامل فرمت EMBL روی صفحه راهنما ENA[ebi-manual] مطرح شده است.

بازیابی آنلاین ENA

ENA چندین شکل برای جستجوی صحیح را ارائه می‌دهد. ابتدا جستجوی ساده است، که به جستجوی متن یا بازیابی توالی منجر می‌شود (شکل ۲-۳). برای جستجوی متن، امکان جستجوی شماره‌های دسترسی و متن‌های ساده رایگان می‌باشد. جستجو به حوزه‌های پایگاه داده محدود نمی‌شود و اجازه محدود کردن جستجو برای حوزه‌های متن را نمی‌دهد اما سامانه Entrez اینگونه عمل می‌کند. در عوض همه ورودی‌های پایگاه داده که به صورت تصادفی حاوی واژه جستجو هستند، بازیابی می‌شوند. برای استفاده از این نوع پارامتر، برای جستجوی توالی از *S. cerevisiae*با توالی به طول ۳۲۷۰ جفت باز برای مثال، جستجوی پیشرفته باید استفاده شود. برای این امر باید لینک مربوطه زیر حوزه جستجوی ساده متن را دنبال کنید. شکل جستجوی پیشرفته (شکل ۲-۴) با چند ویژگی دیگر حوزه‌های پایگاه داده شروع می‌شود. زمانی که یکی از این ویژگی‌ها انتخاب شود، بخش‌های متن و جعبه‌های گزینه بیشتری نمایش داده می‌شود که امکان محدود کردن جستجو برای حوزه‌های پایگاه داده منفرد یا گروه‌های را

---

[1] International Nucleotide Sequence Database Collaboration

فراهم می‌کند. برای بازیابی توالی *S. cerevisiae* مزبور، ما باید قسمت *Sequence* را انتخاب کنیم و واژه جستجو Saccharomyces cerevisiae را درون حوزه *Taxon* وارد کنیم. عامل مقایسه بر روی معادل[1] تنظیم شده است. استفاده از دو عامل دیگر، البته، فقط در صورتی مفید است که ما مقادیر شمارشی را مقایسه کنیم. در حوزه *Base count*، ۳۲۷۰ را وارد می‌کنیم و عامل مقایسه بر روی مساوی یا کمتر تنظیم می‌شود (=>). هنگامی که وارد می‌کنیم، همه ورودی‌ها همزمان به یک درخواست ترجمه می‌شوند، که در حوزه متن خاکستری در بالای صفحه نمایش داده می‌شود. بازیابی با فشار دادن دکمه *Search* شروع می‌شود. متأسفانه، این شکل جستجو اجازه نمی‌دهد که دامنه‌ای که در مثال EntrezNCBI برای طول توالی انجام دادیم، را جستجو کنیم. ولی این امکان هست که درخواست را در سازنده درخواست[2] ابتدا بدون دامنه بسازیم و سپس درخواست حاصل را به صورت دستی ویرایش کنیم. برای انجام این کار، ما بر روی زیرلینک *EditQuery* سمت راست بخش جستجوی متن کلیک می‌کنیم. هم اکنون ما می‌توانیم درخواست پیش ساخته را تغییر دهیم و محدودیت بیشتری را برای ID بخش *base_count* با *AND* اضافه کنیم. هم اکنون درخواست حاصل tax_eq(4932) AND (base_count > = 3260 AND base_count <= 3270) است. گاهی لازم است که از پرانتز استفاده کنیم تا بر برتری عملگر منطقی اثر بگذاریم. در اینجا ممکن است ضروری نبوده باشد؛ ولی ما از پرانتز برای علل خوانایی استفاده کردیم. اگر ما توالی S. cerevisiae را مد نظر قرار داده بودیم که کوتاه‌تر از ۳۲۶۰ جفت باز یا بلندتر از ۳۲۷۰ جفت باز باشد، ما باید از پرانتز استفاده می‌کردیم تا برتری عملگر منطقی را لغو کنیم. درخواست ممکن است به tax_eq(4932) AND (base_count <= 3260 OR base_count >= 3270) ختم شود.

علاوه بر جستجوی متن، ENA اجازه جستجوی توالی را با استفاده از مقایسه توالی می‌دهد. اساساً، این یک جستجوی BLAST است، که می‌توان با استفاده از پارامترهای BLAST استاندارد انجام دهیم یا این امکان هست که پارامترهای BLAST را بر صفحه جستجوی پیشرفته تنظیم کنیم. جستجوهای BLAST با جزئیات در فصل بعدی بحث می‌شود، بنابراین ما این مورد را بررسی نمی‌کنیم.

---

[1] Equal

[2] Query builder

۲-۲-۲ پایگاه‌های داده توالی پروتئین

۲-۲-۲-۱ UniProt

اطلاعات موجود برای پروتئین به سرعت درحال رشد است. علاوه بر اطلاعات توالی، پروفایل‌های بیان را می‌توان بررسی کرد، ساختارهای ثانویه پیش‌بینی می‌شوند و اعمال زیستی یا بیوشیمیایی تحلیل می‌شوند. همه این داده‌ها در پایگاه‌های داده ذخیره می‌شوند، که بعضی از آن‌ها کاملاً اختصاصی هستند. بنابراین، جمع‌آوری همه اطلاعات مرتبط در مورد هر پروتئین مشخص زمانبر است. به همین دلیل، EBI، مؤسسه سوئیسی بیوانفورماتیک[1] (SIB)، و دانشگاه جورجتاون مشارکتی با هدف ایجاد کاتالوگ مرکزی برای اطلاعات پروتئین را تشکیل داده‌اند. نتیجه منبع جهانی پروتئین (UniProt) [uniprot] (UniProt Consortium 2016)، که اطلاعات را در سه پایگاه داده پروتئین Swissprot، TrEMBL و منبع اطلاعات پروتئین[2] (PIR) متحد کرده‌اند. UniProt از سه قسمت تشکیل شده است، شامل: پایگاه UniProt (UniProtKB)، پایگاه داده خوشه‌های مرجع UniProt (UniRef) و آرشیو UniProt (UniPArc)، که مجموعه‌ای از توالی‌های پروتئین و تاریخچه آن‌ها را دارا می‌باشد.

توالی‌های پروتئین و تفسیر آن‌ها در پایگاه UniProt ذخیره می‌شوند (UniProtKB)، که به دو بخش تقسیم می‌شود. بخش اول UniProtKB/TrEMBL است، که حاوی توالی‌های تفسیر شده خودکار است، و بخش دوم UniProtKN/SwissProt است، که به صورت دستی توالی‌های کنترل و تفسیر شده، ذخیره می‌شوند. UniProtKB/TrEMBL هم اکنون (ژوئن ۲۰۱۶) حاوی حدود ۶۵ میلیون ورودی است و بنابراین حدود ۱۲۰ مرتبه بزرگتر از حوزه UniProtKB/SwissProt است، که حاوی حدود ۵۵۰۰۰۰ ورودی است. به دلیل کنترل دستی، بخش UniprotKB/SwissProt یکی از مهم‌ترین پایگاه‌های داده پروتئینی محسوب می‌شود. که اغلب به عنوان استاندارد طلایی تفسیر پروتئین هم ارجاع داده می‌شود.

پایگاه داده SwissProt قبل از پایگاه داده UniProt تأسیس شد و در SIB قرار گرفت. از آنجا که گروه متخصصین در SIB با توده توالی‌های جدید که در پایگاه‌های داده وارد شد، غوطه ور شدند، ضمیمه‌ای بر پایگاه داده SwissProt، پایگاه داده TrEMBL وارد شد.

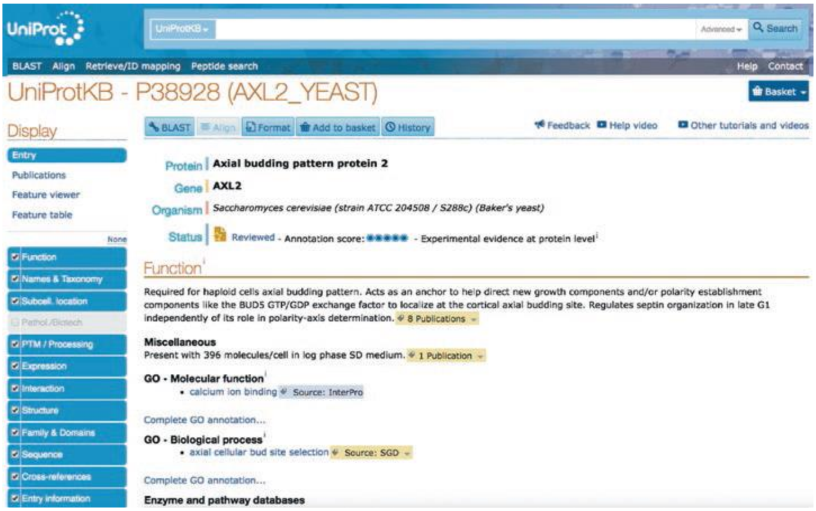
---

[1] SwissInstitute of Bioinformatics

[2] Protein Information Resource

TrEMBL در واقع EMBL ترجمه شده است که هنوز به صورت دستی سرپرستی نشده است. پایگاه داده EMBL پیشگام ENA است. همه ورودی‌ها در TrEMBL (امروزه UniProtKB/TrEMBL) به صورت خودکار تفسیر شده، یعنی کیفیت تفسیرها نسبت به تفسیرهای UniProtKB/SwissProt قابل مقایسه است.

شکل ۲-۵ یک ورودی در پایگاه داده UniProtKB/SwissProt را نشان می‌دهد. در نگاه اول، ورودی مشابه ورودی ENA است. در واقع، هر دو پایگاه داده به هم مرتبط هستند. هر دو پایگاه داده از شناسه‌های[1] دو حرفی استفاده می‌کنند و بیشتر شناسه‌ها برای هر دو پایگاه داده یکسان هستند. ولی بعضی از شناسه‌ها، برای UniProtKB تغییر یافته‌اند و بعضی اضافه شده‌اند. ورودی خام پایگاه داده که در شکل ۲-۵ نشان داده شده است به ندرت یافت می‌شود. اغلب، یک نسخه گرافیکی توسط UniProtKB نشان داده می‌شود، همانطور که در شکل ۲-۶ آمده است.



UniProtKB - P38928 (AXL2_YEAST)

Protein | **Axial budding pattern protein 2**

Gene | **AXL2**

Organism | *Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

Required for haploid cells axial budding pattern. Acts as an anchor to help direct new growth components and/or polarity establishment components like the BUD5 GTP/GDP exchange factor to localize at the cortical axial budding site. Regulates septin organization in late G1 independently of its role in polarity-axis determination. 8 Publications

Miscellaneous
Present with 396 molecules/cell in log phase SD medium. 1 Publication

GO - Molecular function
- calcium ion binding Source: InterPro

Complete GO annotation...

GO - Biological process
- axial cellular bud site selection Source: SGD

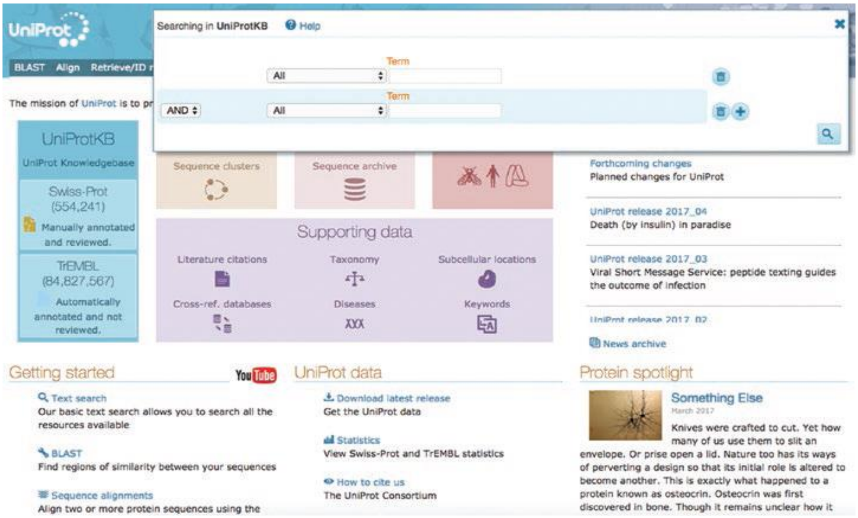
Complete GO annotation...

Enzyme and pathway databases

شکل ۶-۲) داده های ورودی پایگاه **UniProtKB/SwissProt**

---

[1] Identifiers



The mission of UniProt is to pr

Manually annotated and reviewed.

Automatically annotated and not reviewed.

Our basic text search allows you to search all the resources available

Find regions of similarity between your sequences

Align two or more protein sequences using the

Get the UniProt data

View Swiss-Prot and TrEMBL statistics

The UniProt Consortium

Planned changes for UniProt

Death (by insulin) in paradise

Viral Short Message Service: peptide texting guides the outcome of infection

Knives were crafted to cut. Yet how many of us use them to slit an envelope. Or prise open a lid. Nature too has its ways of perverting a design so that its initial role is altered to become another. This is exactly what happened to a protein known as osteocrin. Osteocrin was first discovered in bone. Though it remains unclear how it

شکل ۷-۲) صفحه اصلی پایگاه دادهUniProtبه همراه جستجوی ساده متن

UniProtKB با استفاده از جستجوی متن کامل یا استفاده از درخواست‌های پیچیده با عملگرهای منطقی مورد جستجو قرار می‌گیرند (شکل ۲-۷). برای یک جستجوی ساده، عبارت مورد جستجو به سادگی در قسمت متن در بالای صفحه وارد می‌شود. برای جستجوهای پیچیده، شکل جستجوی پیشرفته استفاده می‌شود. جستجو با کلیک روی لینک Advanced در گوشه سمت راست قسمت متن شروع می‌شود. در شکل جستجوی پیشرفته، IDهای بخش و عملگرهای منطقی مربوطه از فهرست‌های کشویی انتخاب می‌شوند. وقتی شروع می‌شود، درخواست جستجو در قسمت متن به نمایش در می‌آید و در صورت لزوم به صورت دستی قابل تنظیم است.

UniRef یک پایگاه داده غیرتکراری است که برای بررسی سریع جستجوهای مشابه است. این پایگاه داده در سه نسخه موجود است: UniRef100، UniRef90 و UniRef50. هر کدام از پایگاه‌های داده مناسب جستجوی توالی‌های با همسانی ۱۰۰ درصد، بالای ۹۰ درصد یا بالای ۵۰ درصد هستند. اندازه پایگاه داده هم به همین ترتیب تغییر می‌کند، که جستجوی همسانی را برای مثال با BLAST سریع‌تر می‌سازد.

۲-۲-۲-۲ پایگاه داده پروتئین NCBI

پایگاه داده توالی پروتئین شناخته شده دیگر در NCBI نگهداری می‌شود. ولی این پایگاه داده فقط یک پایگاه داده نیست بلکه جمع‌آوری ورودی‌های یافت شده در سایر پایگاه‌های داده توالی پروتئین است. برای مثال، پایگاه داده NCBI حاوی ورودی‌ها از Swissprot، پایگاه داده PIR[pir]، پایگاه بانک داده‌های پروتئین[1] (PDB)[pdb]، ترجمه‌های پروتئین از پایگاه‌های داده GenBank و چند پایگاه داده دیگر است. فرمت آن به GenBank مربوط است و درخواست‌ها همانند پایگاه GenBank از طریق سامانه Entrez از سایت NCBI انجام می‌شود.

۲-۳ پایگاه‌های داده ثانویه

۲-۳-۱ Prosite

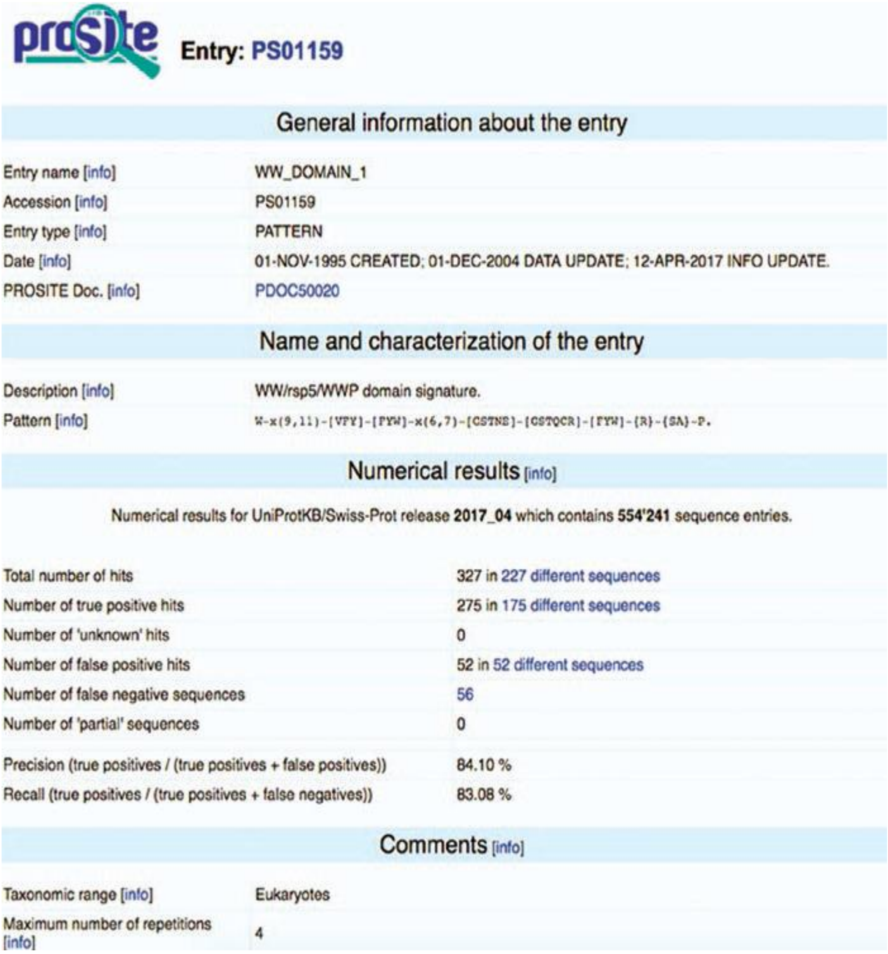
یک پایگاه داده ثانویه زیستی Prosite[prosite] است (Sigrist et al. 2012)، که درSIB واقع شده است [expasy]. طبقه بندی پروتئین‌ها در Prosite با استفاده از موتیف‌های محافظت شده تکی تعیین می‌شود، یعنی، نواحی کوتاه توالی (۲۰-۱۰ اسیدآمینه) که در پروتئین‌های مربوطه محافظت شده هستند و معمولاً نقش کلیدی در عمل پروتئین دارند. جستجو برای چنین موتیف‌های توالی در پروتئین‌های ناشناخته اولین نشانه نسبت به خانواده یا عملکرد پروتئین را فراهم می‌کند. یک موتیف از هم ترازی[2]های مختلف مشتق می‌شود (فصل ۳) و در پایگاه‌های داده به صورت بیان منظم ذخیره می‌شود (شکل ۸-۲). این یک الگوی رسمی برای توصیف ویژگی‌های توالی‌ها می‌باشد. در یک بیان منظم در Prosite، هر اسیدآمینه با یک کد یک حرفی نمایش داده می‌شود و با خط فاصله از هم جدا می‌شوند. اگر یک جایگاه بیش از یک اسیدآمینه داشته باشد، داخل قلاب‌های گوشه‌دار قرار می‌گیرد. جایگاه‌هایی که می‌توانند با یک اسیدآمینه پر شوند با حرف کوچک *x* مشخص می‌شوند. تکرارهای یک اسیدآمینه در پرانتز قرار می‌گیرند، که بعد از آن شماره تکرار می‌آید. نحوه نوشتن بیان منظم در پایگاه Prosite مطابق شکل مقابل می‌باشد: -[GSTNE]-[GSTQCR] [FYW]-{ANW}-x(2)-P. این بیان منظم سه موقعیت اسیدآمینه را دارد. اولین اسیدآمینه

---

[1] Protein Data Bank

[2] Alignment

می‌تواند گلیسین، سرین، ترئونین، آسپارژین یا گلوتامات باشد؛ دومین موقعیت گلیسین، سرین، ترئونین، گلوتامین، سیستئین یا آرژنین است؛ و سومین موقعیت فنیل‌آلانین، تیروزین یا تریپتوفان است. موقعیت چهارم می‌تواند هر کدام از چهار اسیدآمینه به جز آلانین، آسپارژین و تریپتوفان باشد. در موقعیت‌های پنجم و ششم، هر اسید آمینه‌ای می‌تواند باشد موقعیت هفتم با پرولین پر می‌شود. راهنمای کاربر Prosite[prositemanual] حاوی توضیح کاملی از پایگاه داده Prositeبه همراه ترکیب بیان‌های منظم Prosite است. سرور Expasy Prosite Web[prosite] احتمالات مختلف برای درخواست در پایگاه Prosite را ارائه می‌دهد. علاوه بر جستجوی کلیدواژه‌ها، فردی می‌تواند یک توالی را برای حضور موتیف‌های Prosite بررسی کند. به علاوه، با استفاده از الگوریتم ScanProsite، Prosite این امکان را فراهم می‌کند تا در Swissprot، TrEMBL و PDB به دنبال توالی‌های پروتئینی باشیم که حاوی الگوی تعریف شده کاربر است.



Entry: PS01159

General information about the entry

Entry name [info]	WW_DOMAIN_1
Accession [info]	PS01159
Entry type [info]	PATTERN
Date [info]	01-NOV-1995 CREATED; 01-DEC-2004 DATA UPDATE; 12-APR-2017 INFO UPDATE.
PROSITE Doc. [info]	PDOC50020

Name and characterization of the entry

Description [info]	WW/rsp5/WWP domain signature.
Pattern [info]	W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-{R}-{SA}-P.

Numerical results [info]

Numerical results for UniProtKB/Swiss-Prot release **2017_04** which contains **554'241** sequence entries.

Total number of hits	327 in 227 different sequences
Number of true positive hits	275 in 175 different sequences
Number of 'unknown' hits	0
Number of false positive hits	52 in 52 different sequences
Number of false negative sequences	56
Number of 'partial' sequences	0
Precision (true positives / (true positives + false positives))	84.10 %
Recall (true positives / (true positives + false negatives))	83.08 %

Comments [info]

Taxonomic range [info]	Eukaryotes
Maximum number of repetitions [info]	4

شکل ۸-۲) اطلاعات ثبت شده توالی PS01159 از پایگاه داده Prosite

۲-۳-۲ PRINTS

پایگاه داده PRINTS[prints] (Attwood et al. 2003) از اثر انگشت برای طبقه‌بندی توالی‌ها استفاده می‌کند. اثر انگشت از چند موتیف توالی تشکیل شده است، که در پایگاه داده PRINTS با هم ترازی‌های کوتاه، منطقه‌ای و بدون شکاف، ارائه می‌شوند (فصل ۳). پایگاه داده PRINTS این فایده را دارد که پروتئین‌ها معمولاً حاوی نواحی عملکردی هستند که منجر به

چند موتیف توالی در هر پروتئین می‌شود. با استفاده از اثر انگشت، حساسیت تحلیل‌ها افزایش می‌یابد، یعنی این امکان هست که نسبت پروتئین به خانواده پروتئین را حتی در غیاب یکی از موتیف‌های بررسی شده، ارزیابی کنیم. علاوه بر اطلاعات اینکه چگونه یک اثر انگشت را به دست آورده و کیفیت آن را قضاوت کنیم، PRINTS همچنین اجازه دسترسی به اطلاعات بیشتر در مورد خانواده پروتئین مورد نظر را می‌دهد. همانند پایگاه Prosite، PRINTS حاوی اطلاعاتی در مورد هر کدام از خانواده پروتئین و در صورت امکان، عمل زیستی هر موتیف در اثر انگشت می‌باشد. درخواست از پایگاه داده در سرور وب PRINTS[prints] از طریق جستجوی کلیدواژه انجام می‌شود. ولی جستجو برای اثر انگشت در توالی‌های پروتئین می‌تواند جالب‌تر باشد. مانند سِرور Prosite، سِرور PRINTS ابزاری را برای تحلیل توالی ارائه می‌دهد.

۳-۳-۲ Pfam

پایگاه داده Pfam[pfam] (Finn et al. 2016) خانواده‌های پروتئین را طبق پروفایل‌ها طبقه‌بندی می‌کند. یک پروفایل، الگویی است که احتمال حضور، اضافه شدن و یا حذف یک اسیدآمینه را در هر موقعیت از توالی پروتئین ارزیابی می‌کند.

موقعیت‌های محافظت‌شده سنگین‌تر از موقعیت‌های محافظت‌نشده هستند. Pfam بر اساس هم‌ترازی توالی است. هم‌ترازی‌های با کیفیت بالا، بصورت دستی بررسی شده به عنوان نقطه شروع برای ساخت خودکار مدل‌های مخفی مارکوف[1] (HMMs) هستند. سپس توالی‌های بیشتر به صورت خودکار به هم‌ترازی‌های فردی پایگاه داده SwissProt اضافه شدند. هم‌ترازی‌های حاصل باید ساختارهای جذاب از نظر عملکردی را نشان دهند و حاوی توالی‌های مرتبط از نظر تکاملی باشند. به دلیل ساختمان نیمه خودکار هم‌ترازی‌ها، این امکان وجود دارد که همترازی‌هایی در توالی رخ دهد که هیچ رابطه تکاملی با یکدیگر نداشته باشند. بنابراین، نتایج جستجو در برابر پایگاه داده Pfam باید به دقت مرور شود.

---

[1] Markov

۲-۳-۴ Interpro

منبع تلفیقی خانواده‌ها، دامین‌ها و مکان‌های پروتئینی[1] (Interpro) [interpro] (Mulder et al. 2007) پایگاه‌های داده ثانویه مهم را در یک پایگاه داده شناخته شده جامع تلفیق می‌کنند. Interpro پایگاه‌های داده Swissprot، TrEMBL، Prosite، Pfam، PRINTS، ProDom، Smart و TIGRFAMs[tigr] را تلفیق نموده و در نتیجه منجر به درخواست ساده و همزمان این پایگاه‌های داده می‌شود. صفحه حاصل خروجی درخواست‌های فردی را ترکیب می‌کند. این مورد باعث مقایسه سریع نتایج می‌شود در حالی که نقاط قوت و ضعف هر پایگاه داده را در نظر می‌گیرد. پایگاه Interpro تعدادی تسهیلات برای جستجوهای متن و توالی ارائه می‌دهد.

۲-۴ پایگاه‌های داده ژنوتیپ-فنوتیپ

برای ایجاد بیماری‌هایی که منتشر شده و پیشرفت می‌کنند، وجود چند ژن یا محصولات آن‌ها به فراوانی لازم هستند. بنابراین شناسایی ژن‌هایی که به بیماری مرتبط هستند، اهمیت حیاتی‌تری نسبت به تولید دارو دارد. امروزه تعدادی از پایگاه‌های داده ژنوتیپ- فنوتیپ تثبیت شده‌اند که روابط بین ژن‌ها و ویژگی‌های زیستی موجود را ثبت می‌کنند. پایگاه داده وراثت مندلی آنلاین در انسان[2] (OMIM) متعلق به سایت NCBI[omim] شناخته شده‌ترین پایگاه داده ژنوتیپ- فنوتیپ است. یک پایگاه داده جدید از این نوع، dbGaP[dbgap]، هم اخیراً در NCBI ثبت شده است. داده‌های این پایگاه با تحلیل اهمیت آماری رابطه ژنوتیپ-فنوتیپ مربوطه عمل می‌کند. پایگاه داده وراثت مندلی آنلاین در حیوانات[3] (OMIA) در NCBI[omia] هم حاوی روابط ژنوتیپ-فنوتیپ دو موجود مدل مهم، *D. melanogaster* و *C. elegans* است که به ترتیب در FlyBase[flybase] و WormBase[wormbase] ثبت شده‌اند. هر دو پایگاه داده حاوی اطلاعات بیشتری علاوه بر داده‌های ژنوتیپ-فنوتیپ هستند. شرح کامل از همه پایگاه‌های داده مزبور [nar] فراتر از بحث این کتاب است. آنچه در ادامه

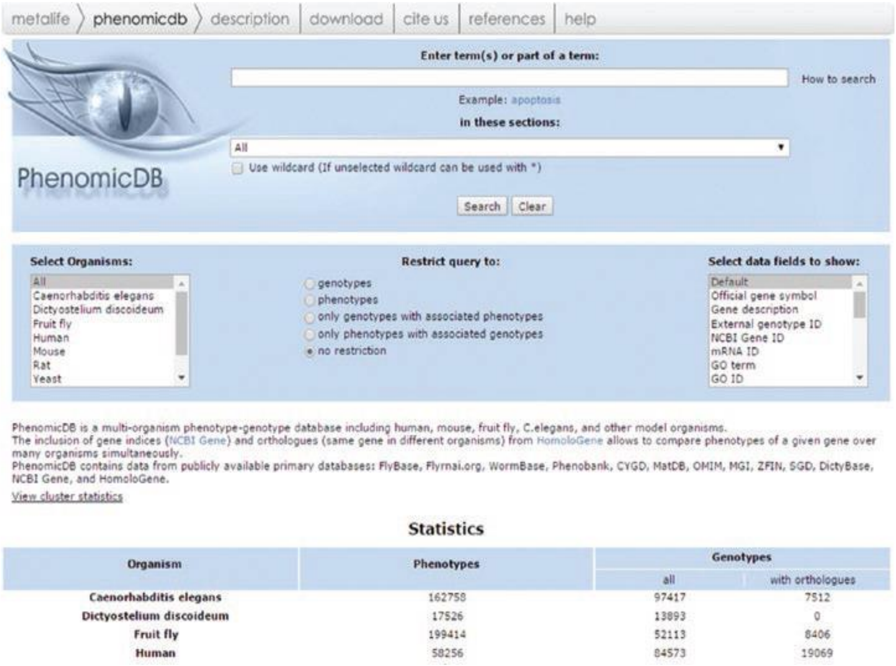
---

[1] Integrated Resource of Protein Families, Domains and Sites

[2] Online Mendelian Inheritancein Man

[3] Online MendelianInheritance in Animals

می‌آید، فقط یک پایگاه داده ژنوتیپ- فنوتیپ است که به صورت معنایی محتوای پایگاه‌های داده مذکور را با هم تلفیق می‌کند.



PhenomicDB is a multi-organism phenotype-genotype database including human, mouse, fruit fly, C.elegans, and other model organisms.
The inclusion of gene indices (NCBI Gene) and orthologues (same gene in different organisms) from HomoloGene allows to compare phenotypes of a given gene over many organisms simultaneously.
PhenomicDB contains data from publicly available primary databases: FlyBase, Flyrnai.org, WormBase, Phenobank, CYGD, MatDB, OMIM, MGI, ZFIN, SGD, DictyBase, NCBI Gene, and HomoloGene.

Statistics

Organism	Phenotypes	Genotypes	
		all	with orthologues
Caenorhabditis elegans	162758	97417	7512
Dictyostelium discoideum	17526	13893	0
Fruit fly	199414	52113	8406
Human	58256	84573	19069

شکل ۹-۲) صفحه نخست پایگاه داده PhenomicDB

PhenomicDB ۱-٤-۲

پایگاه داده PhenomicDB یک پایگاه داده ژنوتیپ–فنوتیپ از موجودات مختلف است که حاوی داده‌ها از انسان‌ها و سایر موجودات مهم مانند موش، ماهی زبرا (*Danio rerio*)، مگس میوه (*D. melanogaster*)، نماتد (*C. elegans*)، مخمر نان (*S. cerevisiae*) و گیاه آرابیدوپسیس (*Arabidopsis thaliana*) می‌باشد. PhenomicDB داده‌های پایگاه‌های داده مذکور و سایر پایگاه‌های داده ژنوتیپ–فنوتیپ اولیه را تلفیق می‌کند. فهرست کاملی از همه منابع داده تحت آن را می‌توان روی صفحه اصلی [phenomicdb] و در مقاله کاهرامان و همکاران (۲۰۰۵) یافت. یکی از ویژگی‌های PhenomicDB امکان مقایسه‌های بین روابط ژنوتیپ- فنوتیپ است. این مورد با ورود داده‌های ارتولوژیکی (مرتب شده) شاخص‌های ژنی از

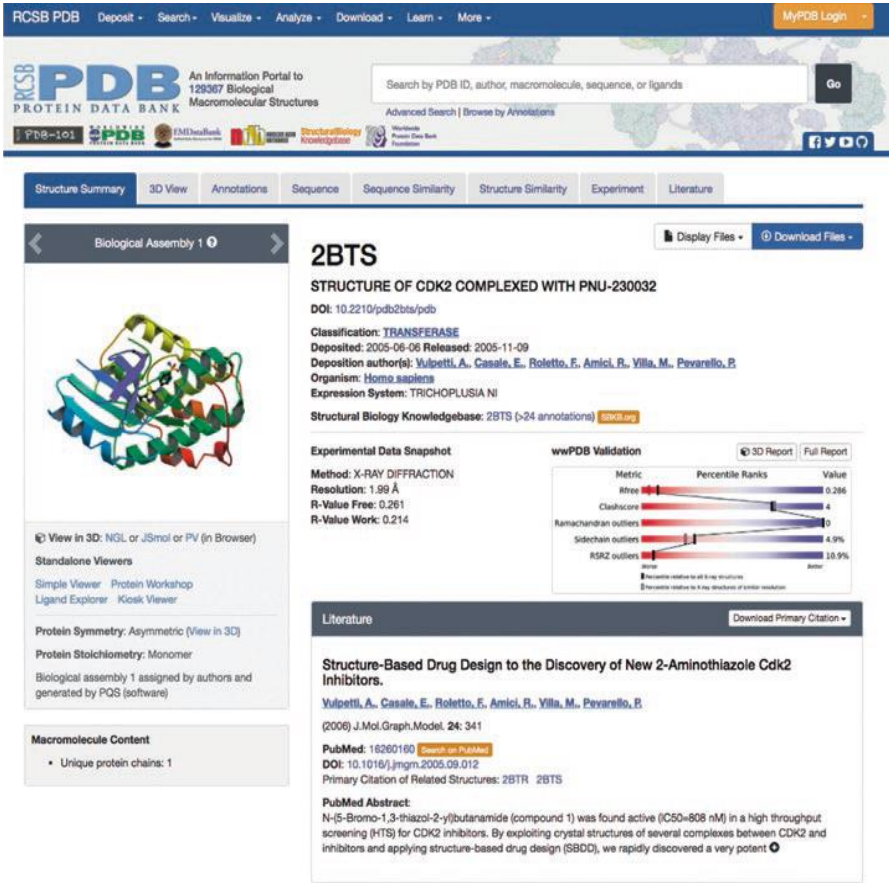
پایگاه‌های داده HomoloGene[homologene] در NCBI کامل می‌شود. برای مثال، علت بیماری پورفیری[1]، یک نقص آنزیمی وراثتی یا اکتسابی انسان‌ها، عدم عملکرد آنزیم آمینولوولینات دهیدراتاز[2] است. ژن مربوطه نماد ALAD را دارد. همانطور که در PhenomicDB آمده، نقص در ژن‌های ارتولوگ مخمر نان (نماد ژن: HEM2) به فنوتیپ مشابهی منجر می‌شود، که با کلیدواژه‌های اگزوتروفی[3]، نقص‌های مصرف کربن و نیتروژن، مصرف کربن و نقص تنفس مشخص می‌شود. البته، کسی نمی‌تواند انتظار داشته باشد که موجودات دور از هم مانند مخمر نان و انسان روابط ژنوتیپ-فنوتیپ یکسانی را در هر مورد نشان دهند. با این وجود، روابط مشابهی رخ می‌دهد که ممکن است فرضیه‌های جدیدی را ایجاد کند که بیماری‌زایی را در نظر بگیرد یا پیشرفت مدل بیماری را ایجاد کند، بنابراین از ایجاد داروهای جدید حمایت می‌کند.

---

[1] Porphyria

[2] δ-Aminolevulinate Dehydratase

[3] Auxotrophies



An Information Portal to 129367 Biological Macromolecular Structures

2BTS

STRUCTURE OF CDK2 COMPLEXED WITH PNU-230032

DOI: 10.2210/pdb2bts/pdb

Classification: TRANSFERASE
Deposited: 2005-06-06 **Released:** 2005-11-09
Deposition author(s): Vulpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Pevarello, P.
Organism: Homo sapiens
Expression System: TRICHOPLUSIA NI

Structural Biology Knowledgebase: 2BTS (>24 annotations)

View in 3D: NGL or JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Protein Symmetry: Asymmetric (View in 3D)

Protein Stoichiometry: Monomer

Biological assembly 1 assigned by authors and generated by PQS (software)

Macromolecule Content

- Unique protein chains: 1

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.99 Å
R-Value Free: 0.261
R-Value Work: 0.214

wwPDB Validation

Literature

Structure-Based Drug Design to the Discovery of New 2-Aminothiazole Cdk2 Inhibitors.

Vulpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Pevarello, P.

(2006) J.Mol.Graph.Model. **24:** 341

PubMed: 16260160
DOI: 10.1016/j.jmgm.2005.09.012
Primary Citation of Related Structures: 2BTR 2BTS

PubMed Abstract:
N-(5-Bromo-1,3-thiazol-2-yl)butanamide (compound 1) was found active (IC50=808 nM) in a high throughput screening (HTS) for CDK2 inhibitors. By exploiting crystal structures of several complexes between CDK2 and inhibitors and applying structure-based drug design (SBDD), we rapidly discovered a very potent

شکل ۲-۱۰) نمای کلی اطلاعات ژن 2BTS از پایگاه داده PDB

PhenomicDB از طریق رابط ساده جستجو درخواست می‌گیرد. واژه‌های جستجو به صورت خودکار یا دستی توسط جایگزین‌شونده‌ها تکمیل می‌شوند و به حوزه‌های پایگاه داده خاصی محدود می‌شوند. به علاوه، این امکان هست که جستجو را به موجودات انتخابی محدود کنیم. اگر ارتولوگ‌های یک ژن یافت شوند، صفحه حاصل لینکی را به ثبت پایگاه داده مربوطه ارجاع می‌دهد، که مقایسه سریع‌تری از روابط ژنوتیپ- فنوتیپ در بین موجودات قابل انجام است (شکل ۲-۹). به خاطر تلفیق معنایی پایگاه‌های داده اولیه، بعضی از اطلاعات از دست

می‌رود، ولی با پیوستگی داده‌های اولیه و وسعت اطلاعات وارد شده جبران می‌شود. بنابراین PhenomicDB می‌تواند به عنوان موتور فراجستجو برای اطلاعات فنوتیپی به کار رود.

۲-۵ پایگاه‌های داده ساختار مولکولی

۲-۵-۱ بانک اطلاعاتی پروتئین[1]

PDB یک پایگاه داده از ساختارهای کریستالی مشخص شده تجربی از درشت مولکول‌های زیستی است و با کنسرسیوم واقع در ایالات متحده، اروپا و ژاپن منطبق می‌شود [wwpdb] (Berman et al. 2000). احتمالاً بهترین صفحه PDB، بخش تحقیقاتی بیوانفورماتیک ساختاری[2] است [pdb]. PDB در آزمایشگاه ملی بروخاون[3] در سال ۱۹۷۱ تأسیس شد، که در استفاده مکرر از نام بانک داده پروتئینی بروخاون بازتاب یافته است. حدود ۱۲۱۰۰۰ ساختار درشت مولکول در پایگاه داده PDB ذخیره شده است (تا جولای ۲۰۱۶). این موارد بیشتر پروتئین‌ها هستند، ولی شامل ساختارهای DNA، RNA و کمپلکس‌های پروتئین- نوکلئیک اسید هم می‌شوند. ساختارهای سایر درشت مولکول‌ها، برای مثال، گلیکوپپتیدها و پلی‌ساکاریدها، فقط بخش کوچکی از کل ساختارها را تشکیل می‌دهند. از سال ۲۰۰۲، فقط آن ساختارهای کریستالی که به صورت تجربی حل شده‌اند در پایگاه داده PDB ذخیره شده‌اند، در حالی که داده‌های مدل‌های پروتئینی نظری در بخش خودشان نگهداری می‌شوند [pdb-models].

پایگاه داده PDB چند گزینه درخواست دارد. جستجوی بر پایه متن برای PDB ID یا یک کلیدواژه را می‌توان روی صفحه اصلی شروع کرد. به علاوه، تعدادی از گزینه‌های جستجو روی صفحه پایگاه داده جستجو وجود دارند، که شامل کلیدواژه‌های دقیق‌تر و درخواست‌های BLAST می‌شوند. یک ثبت پایگاه داده همه اطلاعات فایل را خلاصه می‌کند و چیزی در صفحات بعدی تشریح شده است. به علاوه، ساختار مولکولی را می‌توان به وسیله آپلت‌های مختلف مشاهده کرد.

---

[1] Protein Data Bank
[2] Research Collaboratory for Structural Bioinformatics
[3] Brookhaven

SCOP ۲-۵-۲

پروتئین‌هایی که عمل زیستی مشابهی را انجام می‌دهند و از نظر تکاملی به هم مرتبط‌اند، سازماندهی ساختاری مشابهی دارند. بنابراین باید این امکان باشد که عمل یک پروتئین ناشناخته را با مقایسه سازماندهی ساختاری با پروتئین شناخته شده پیش‌بینی کرد. دو پایگاه داده، SCOP و CATH، چنین پیش‌بینی را فراهم می‌کنند. SCOP (طبقه‌بندی ساختاری پروتئین‌ها[1]) [scop] (Murzin et al. 1995) پروتئین‌های با ساختار شناخته شده را به صورت سلسله مراتبی طبقه‌بندی کرده است. سه طبقه‌بندی اصلی به ترتیب شامل: خانواده‌ها، فوق‌خانواده‌ها و تاخوردگی‌ها می‌باشند. خانواده‌ها، پروتئین‌های با رابطه تکاملی مشخص را نسبت به یکدیگر تشریح می‌کند و باید حداقل بیشتر از ۳۰ درصد کل طول پروتئین همسانی توالی داشته باشند. با این وجود، پروتئین‌هایی که زیر این محدوده هستند هم می‌توانند در یک خانواده قرار گیرند به این شرط که ارتباط را به دلیل ساختارها و اعمال مشابه تأیید شده نشان دهند. پروتئین‌هایی با همسانی توالی بسیار اندک، به دلیل ویژگی‌های ساختاری و عملکردی، در فوق خانواده‌ها قرار می‌گیرند، ولی پروتئین‌هایی که آرایش یکسانی از عناصر ساختار ثانویه با توپولوژی یکسان دارند در تاخوردگی‌ها طبقه‌بندی می‌شوند. البته، این امر مهم نیست که پروتئین‌ها رابطه عملکردی دارند یا شباهت تاخوردگی‌ها بر اساس اصول فیزیکوشیمیایی باشد. اخیراً، نسخه جدید، پایگاه داده SCOP2[scop2] (Andreeva et al. 2014) ایجاد شده است.

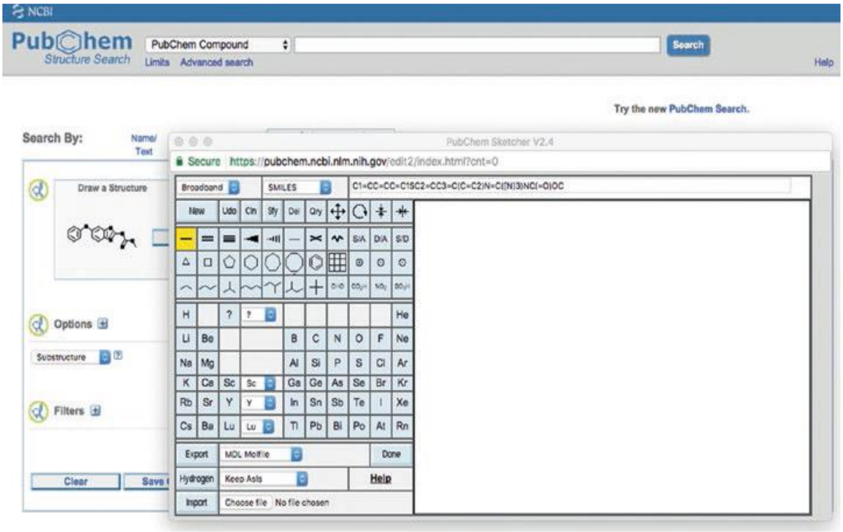
CATH ۲-۵-۳

پایگاه داده CATH[cath] (Greene et al. 2007) ساختارهای پروتئینی را به صورت سلسله مراتبی در چهار دسته قرار می‌دهد: رده (C)، معماری (A)، توپولوژی (T) و فوق خانواده همولوگ (H). طبقه‌بندی پروتئین‌ها در دسته رده اصولاً خودکار است، ولی در صورت لزوم با تفسیر دستی تکمیل می‌شود. در دسته رده، بخش عناصر ساختاری ثانویه بدون در نظر گرفتن آرایش آن‌ها یا اتصالات حساب می‌شود. چهار رده مجزا برای پروتئین‌ها مطرح شده است: پروتئین‌ها اصولاً از مارپیچ (معمولا آلفا)، ورقه (معمولا بتا)، هم مارپیچ و هم ورقه‌ها (آلفا-

---

[1] Structural Classification Of Proteins

بتا) و پروتئین‌هایی با عناصر ساختار ثانویه بسیار کم، تشکیل شده‌اند. دسته معماری، آرایش عناصر ساختار ثانویه را نسبت به یکدیگر تشریح می‌کنند و به صورت دستی سرپرستی می‌شوند. طبقه‌بندی آن از طریق توصیف گره‌های ساده مانند: استوانه‌ای، ساندویچی و پیچ بتا انجام می‌شود. در دسته توپولوژی، شکل پروتئین و اتصالات بیرونی عناصر ساختار ثانویه تشریح می‌شوند. دسته‌بندی آن بر اساس الگوریتمی است که از پارامترهای شهودی برای طبقه‌بندی دامین استفاده می‌کند. دسته فوق‌خانواده همولوگ از دامین‌های پروتئینی همولوگ تشکیل شده‌اند، یعنی دامین‌هایی با ناحیه مشترک. تشابه توالی‌ها با مقایسه توالی تعیین می‌شود که بعد از آن مقایسه ساختاری طبق طبقه‌بندی دسته توپولوژی صورت می‌گیرد. علاوه بر این چهار دسته (که حرف اول هر کدام نام پایگاه داده را تشکیل می‌دهد)، دسته پنجم هم با عنوان خانواده‌های توالی، تعریف شده است. در اینجا، دامین‌ها بر اساس همسانی زیاد توالی، طبقه‌بندی می‌شوند (حداقل ۳۵ درصد همسانی در ۶۰ درصد طول دامین بزرگتر) و بنابراین عملکرد یکسانی خواهند داشت.



https://pubchem.ncbi.nlm.nih.gov/edit2/index.html?cnt=0

SMILES: C1=CC=CC=C1SC2=CC3=C(C=C2)N=C([N]3)NC(=O)OC

شکل ۱۱-۲) ساختار دو بعدی مولکول در پایگاه داده PubChem

٤-٥-٢ PubChem

پایگاه داده PubChem در NCBI[pubchem] مولکول‌های شیمیایی کوچک و اطلاعات در مورد فعالیت‌های زیستی‌شان را ذخیره می‌کند. از سه قسمت تشکیل شده است، ترکیب PubChem، ماده PubChem و سنجش زیستی PubChem. ترکیب PubChem حاوی حدود ۹۱ میلیون مولکول (جولای ۲۰۱۶) با ساختارهای مولکولی دو بعدی (2D) است. درخواست به صورت گرافیکی از طریق ویرایشگر ساختار مولکولی انجام می‌شود که منجر به کشیدن ساختار مطلوب (نسبی) می‌شود (شکل ۲-۱۱). به علاوه، ترکیب PubChem امکان جستجو برای مولکول‌هایی را فراهم می‌کند که پارامترهای فیزیکوشیمیایی معینی دارند، برای مثال: یک دامنه وزن مولکولی خاص، تعداد مشخصی از پذیرنده یا دهنده برای پیوندهای هیدروژنی و یک دامنه logP مشخص دارند. پایگاه PubChem اجازه جستجو برای پیدا کردن یک ماده تولید شده توسط سازندگان شرکت‌های مختلف، نمونه‌های ترکیب ناشناخته و مواد طبیعی با ساختار مولکولی دو بعدی ناشناخته را می‌دهد. ثبت‌های هر دو پایگاه داده به هم متصل شده‌اند و در صورتی که داده‌های مربوطه موجود باشند، لینکی برای پایگاه سوم، PubChem BioAssay اعمال می‌شود. اطلاعات سنجش‌ها و مولکول‌های زیستی که در این سامانه‌ها بررسی شده‌اند در PubChem BioAssay ثبت شده‌اند، و این پایگاه داده را می‌توان با جستجوی متن در سامانه Entrez مورد بررسی قرار داد.

پایگاه‌های داده PubChem چند کاربرد به دلیل اتصال به پایگاه‌های داده داخلی و خارجی شامل PubMed دارند. برای مثال، با دانستن یک ممانعت‌کننده آنزیم مشخص، امکان یافتن سایر ممانعت‌کننده‌های بالقوه مشابه حاصل می‌گردد. به علاوه، مولکول‌های شیمیایی کوچکی را می‌توان شناسایی کرد که ساختارهای مختلفی دارند ولی هنوز نشان داده می‌شود که اثرات مشابهی در سامانه بررسی زیستی دارند.

سایت‌های مفید

bankit. http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank
cath. http://www.cathdb.info/
dbgap. http://www.ncbi.nlm.nih.gov/gap
ddbj. http://www.ddbj.nig.ac.jp/
ebi. http://www.ebi.ac.uk/
ebi-manual. http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

ena. http://www.ebi.ac.uk/ena/
entrez. http://www.ncbi.nlm.nih.gov/nucleotide
entrez-help. http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html
expasy. http://www.expasy.org/
flybase. http://www.flybase.org/
gb-sample. http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
genbank. http://www.ncbi.nlm.nih.gov/Genbank/
homologene. http://www.ncbi.nlm.nih.gov/homologene
interpro. http://www.ebi.ac.uk/interpro/
mgd. http://www.informatics.jax.org/
nar. http://nar.oxfordjournals.org/
ncbi. http://www.ncbi.nlm.nih.gov/
nig.https://www.nig.ac.jp/nig/
omia. http://omia.angis.org.au/home/
omim. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
pdb. http://www.rcsb.org/pdb/home/home.do
pdb-models. http://www.rcsb.org/pdb/search/searchModels.do
pfam. http://pfam.xfam.org/
phenomicdb. http://www.phenomicdb.de/
pir. http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml
prints. http://bioinf.man.ac.uk/dbbrowser/PRINTS/
prosite. http://prosite.expasy.org/
prosite-manual. http://prosite.expasy.org/prosuser.html
pubchem. http://pubchem.ncbi.nlm.nih.gov/
scop. http://scop.mrc-lmb.cam.ac.uk/scop/
scop2. http://scop2.mrc-lmb.cam.ac.uk/
sequin. http://www.ncbi.nlm.nih.gov/Sequin/
swissprot. http://www.expasy.org/sprot/
tigr. http://maize.jcvi.org/
uniprot. http://www.uniprot.org/
wormbase. http://www.wormbase.org/
wwpdb. http://www.wwpdb.org/

منابع

1. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42(Databaseissue):D310–D314.
2. Attwood TK, Bradley P, Flower DR, Gaulton A et al (2003) PRINTS and its automatic supplement, pre- PRINTS. Nucleic Acids Res 31:400–402.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242.

4. Finn RD, Coggill P, Eberhardt RY, Eddy SR et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285.
5. Greene LH, Lewis TE, Addou S, Cuff A et al (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297.
6. Kahraman A, Avramov A, Nashev L, Popov D et al (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 21:418–420.
7. Kim KS, Lilburn TG, Renner MJ, Breznak JA (1998) arfI and arfII, two genes of encoding alpha-L arabinofuranosidases in Cytophaga xylanolytica. Appl Environ Microbiol 64:1919–1923.
8. Mulder NJ, Apweiler R, Attwood TK, Bairoch A et al (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–D228.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
10. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 45:D12–D17.
11. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347.
12. The UniProt Consortium (2016) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169.

فصل سوم

جستجوی پایگاه‌ها و مقایسه‌ی توالی‌ها

۱-۳ مقایسه‌های دوتایی و چندگانه توالی

مقایسه توالی‌های پروتئین و DNA یک روش تحلیلی مهم در بیوانفورماتیک کاربردی بشمار می‌رود. تفسیر توالی‌های جدید نوکلئوتیدی و پروتئینی، ساختن ساختارهای مدل برای پروتئین‌ها، طراحی و تحلیل مطالعات بیانی و انواعی از سایر بررسی‌های بیوانفورماتیکی و زیستی، همگی بر اساس این تحلیل‌ها صورت می‌پذیرند. اصولا طبیعت به صورت محافظه‌کارانه عمل می‌کند بدین معنی که، نوع جدیدی از خواص زیست‌شناختی را برای هر نوعاز جانداران ایجاد نمی‌کند بلکه به صورت مداوم و آهسته تغییر نموده و با شرایط جدید سازگار می‌شود. فاکتورهای جدید،خود به خود پدیده نمی‌آید زیرا هیچ ژن جدیدی ناگهانی ظهور نمی‌یابدبلکه در طول دوران تکامل ایجاد شده و تغییر می‌کند. بنابراین، با توجه به این مفهوم، اگر فردی هر میزان از همسانی یک نوع پروتئین را با پروتئین دیگری داشته باشد، ممکن است اطلاعات عملکردی را از پروتئینی به دیگری انتقال دهد. ولی این فرایند باید به صورت کاملا تدریجی صورت پذیرد زیرا پروتئین‌های مشابه ممکن است اعمال متفاوتی داشته باشند. تشابه دو پروتئین می‌تواند بر اساس تکامل از یک جد مشترک (تکامل همگرا)[1] یا مستقل از همدیگر از پروتئین‌های اجدادی مختلف (تکامل واگرا)[2] باشد.

قبل از شروع تحلیل باید دقت نمود کهبین توالی‌ها نوعی ارتباط احتمالی وجود داشته باشند، برای حل این موضوعلازم است که برخی از واژه‌ها را تعریف نمائیم. توالی‌های مرتبط به گونه‌ای طراحی شده‌اند که همولوگ هستند؛ ولی واژه همولوژی اغلب باعث سردرگمی می‌شود. همولوژی[3] مقدار تشابه نیست، بلکه تأکید می‌کند که توالی‌ها تاریخچه تکاملی مشترکی دارند و بنابراین توالی اجدادی مشترکی دارند (Tatusov et al. 1997). با این وجود، تعاریف واژه‌های ارتولوگ[4] و پارالوگ[5] در ترکیب با عمل پروتئین موضوع بحث نیست (Jensen 2001; Gerlt and Babbitt 2001). در کل، زیست‌شناسان این واژه‌ها را به صورتی که در ادامه می‌آید تعریف کرده‌اند. پروتئین‌های همولوگی که از گونه‌های مختلفی باشندو دارای عملکرد یکسانی هستند (مانند کینازهای مربوط به مسیر انتقال پیام در انسان‌ها و موش) ارتولوگ می‌نامند. در

---

[1] Convergent Evolution
[2] Divergent Evolution
[3] Homology
[4] Ortholog
[5] Paralog

مقابل، پروتئین‌های همولوگی که اعمال مختلفی را کنترل نموده و در یک گونه قرار دارند (مانند دو کیناز در مسیرهای انتقال پیام انسان‌ها) پارالوگ گویند.

حتی اگر دو توالی همولوگ باشند یا نباشند اما میزانهمولوژی قابل اندازه‌گیری نیست. ولی همسانی یا تشابه[1] دو توالی قابل اندازه‌گیری است. همسانی با محاسبه تعداد آمینواسید یا نوکلئوتید یکسان در توالی نسبت به کل آمینواسیدها یا نوکلئوتیدها اندازه‌گیری می‌شود. بر خلاف همسانی، شباهت به آسانی قابل محاسبه نیست. قبل از اینکه شباهت تعیین شود، ابتدا درجه تشابه بلوک‌های ساختمانی توالی‌ها نسبت به همدیگر باید تعیین شود. این کار با استفاده از ماتریس تشابه[2] انجام می‌شود که به ماتریس جانشینی یا نمره‌دهی نیز شناخته می‌شود. ماتریس‌های تشابه احتمال اینکه یک توالی به توالی دیگری در طول زمان تبدیل شود را مشخص می‌کنند. البته، این مورد به زمان انقضا و میزان جهش نوکلئوتیدها بستگی دارد. همسانی یک مقدار مشخص است که در مقابل تشابه می‌باشد که بر اساس یک مدل تعریف شده خاص یا ماتریس تشابه نیست. توالی‌های دو پروتئین همولوگ تشابه ۶۰ درصد و همسانی ۴۰ درصد دارند. آن‌ها همولوژی ۶۰ یا ۴۰ درصد را نشان نمی‌دهند، مطلبی که گاهی حتی در منابع استاندارد بیوانفورماتیکی نیز هم به اشتباه بکار برده می‌شود. باید اشاره شود که شباهت را فقط می‌توان برای توالی‌های آمینواسید یافت نه توالی‌های نوکلئوتیدی.

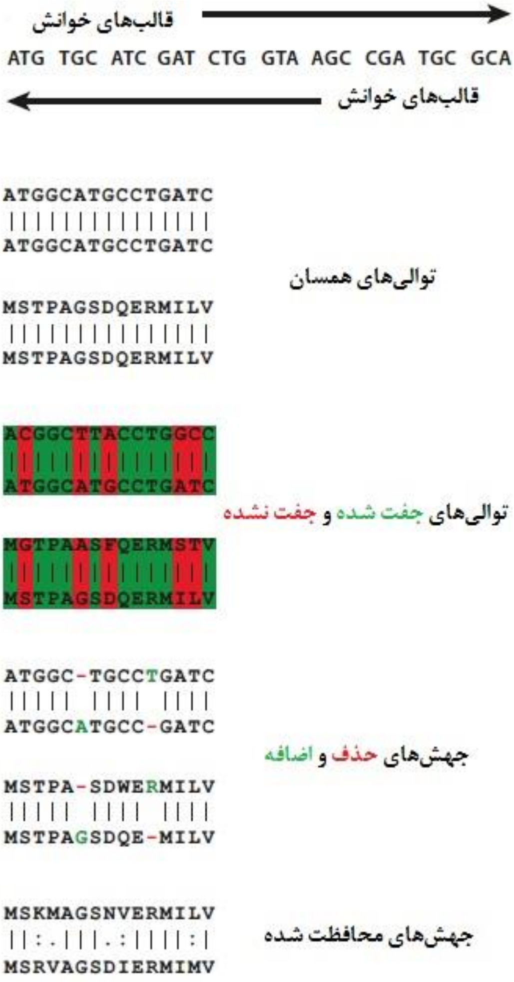
قبل از تصمیم‌گیری در مورد همسانی یا تشابه دو توالی نوکلئوتیدی یا آمینواسیدای، ابتدا باید همردیفی[3] محاسبه شود. اصول بررسی هر همردیفی نسبتاً ساده است (شکل ۳-۱). دو توالی به صورت اختیاری در کنار یکدیگر قرار می‌گیرند و همردیفی طبق مقدار کیفیت قضاوت می‌شود (مانند ماتریس تشابه). سپس دو توالی نسبت به یکدیگر جا به جا می‌شوند، و برای هر موقعیت یک امتیاز محاسبه می‌شود. این فرایند تا زمانی تکرار می‌شود که بهترین همردیفی یافت شود.

---

[1] Identity or Similarity

[2] Similarity Matrices

[3] Alignment



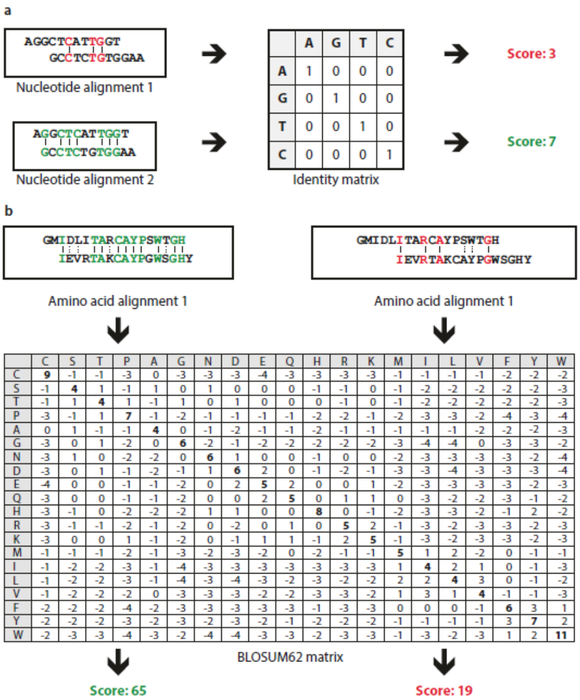
شکل ۳-۱) همردیفی توالی‌های نوکلئوتیدی و آمینواسیدی

یکی از چالش‌های اصلی این فرایند، تعیین میزان کیفیت است. برای توالی‌های نوکلئوتیدی، ساده‌ترین راه، یک ماتریس همسانی[1] است (شکل ۳-۲a). در اینجا، فرض می‌شود که چهار نوکلئوتید هیچ شباهتی نسبت به یکدیگر نشان نمی‌دهند، و بنابراین، فقط

---

[1] Identity Matrix

نوکلئوتیدهای یکسان در نمره‌دهی شباهت وارد فاکتور می‌شوند. آن‌ها به عنوان علائم یکسان (جفت)[1] یا متفاوت (عدم جفت شدن)[2] در نظر گرفته می‌شوند. برای نمره‌دهی نهایی فقط نوکلئوتیدهای یکسان اضافه می‌شوند.



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

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-3	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

شکل ۳-۲) ماتریس نمره‌دهی منجر به محاسبه همردیفی‌های بهینه می‌شود

a) استفاده از ماتریس همسانی برای ساخت یک همردیفی نوکلئوتیدی. b) استفاده از ماتریس BLOSUM62 برای ساخت یک همردیفی آمینواسیدای. دو همردیفی بالقوه برای هر کدام ارائه می‌شود؛ همردیفی بهینه به رنگ سبز نشان داده می‌شود.

برای توالی‌های پروتئینی، تنها استفاده از یک ماتریس همسانی برای فرایندهای زیستی و تکاملی مناسب نیست. آمینواسیدها با همان احتمال قابل تبدیل نیستند زیرا ممکن است به لحاظ نظری اینطور تصور شود. به عنوان مثال تبدیل آسپارژیک‌اسید به گلوتامیک‌اسید اغلب در طبیعت مشاهده می‌شود؛ ولی تبدیل آسپارژیک‌اسید به تریپتوفان به ندرت مشاهده شده است. دلیل آن وجود کدهای ژنتیکی سه‌تایی آن‌هامی‌باشد (فصل ۷-۱). برای تبدیل آسپارژیک‌اسید

---

[1] Match
[2] Mismatch

به گلوتامیک‌اسید فقط یک جهش نوکلئوتید در کدون سوم لازم است (GAT/GAC به GAA/GAG). در مقابل، جهش کامل سه‌تایی باید برای آسپارژیک‌اسید رخ دهد تا به تریپتوفان تبدیل شود (GAT/GAC به TGG). البته چنین جانشینی جهش کاملی با احتمال کمتری رخ می دهد و به مدت زمانی بیشتری نیاز دارد. دلیل دوم جهش آسپارژیک‌اسید به گلوتامیک‌اسید این است که، هر دو ویژگی‌های مشابهی دارند (فصل ۷-۱). اما، آسپارژیک‌اسید و تریپتوفان از نظر شیمیایی هم متفاوت هستند (تریپتوفان آب‌گریز، اغلب در مرکز پروتئین یافت می‌شود، در حالی که آسپارژیک‌اسید آب‌دوست، بیشتر در سطح پروتئین دیده می‌شود). تبدیل آسپارژیک‌اسید به تریپتوفان، می‌تواند به شدت ساختار سوم پروتئین و در نتیجه، عمل آن را تغییر دهد. چنین تبدیل بزرگ آمینواسید که با از دست رفتن عمل همراه است به ندرت رخ می‌دهد.

بنابراین، بیشتر الگوریتم‌ها از ماتریس جایگزینی[1] استفاده می‌کنند تا توالی‌های پروتئین را همردیف کنند. این ماتریس‌های جایگزینی آمینواسید این احتمال را تشریح می‌کند که آمینواسیدها با تکامل تبدیل می‌شوند. این ماتریس‌ها یک رابطه لگاریتمی‌برای تعیین احتمال دو روابط می‌باشدبه این صورت که یک جفت آمینواسید یا نوکلئوتید در یک همردیفی ظاهر می‌شوند، یعنی احتمال رخداد تصادفی و احتمال رخداد تکاملی محاسبه می‌گردند. مقادیر منفی در ماتریس به این معنی است که رخداد ممکن است تصادفی باشد در حالی که مقادیر مثبت بیانگر رخداد تکاملی است. از آنجا که مقادیر ماتریس لگاریتم‌های روابط است، اضافه کردن تعداد به همردیفی کامل منجر می‌شود. عمومی‌ترین ماتریس‌های نمره‌دهی مورد استفاده،موقعیت جهش‌پذیر[2] (PAM) (Dayhoff et al. 1978) و ماتریکس بلوک‌های جایگزین[3] (BLOSUM) (Henikoff and Henikoff 1992) می‌باشد (شکل ۳-B۲).

همردیفی، هم به صورت کلی[4] و هم منطقه‌ای[5] رخ می‌دهد (شکل ۳-۳). در همردیفی کلی توالی‌های کامل نوکلئوتیدی یا پروتئینی با یکدیگر در طول کامل توالی، مقایسه می‌شوند. شکل ۳-۴ محاسبه همردیفی کلی را نشان می‌دهد. ولی حتی توالی‌های بسیار مشابه نیز

---

[1] Substitution Matrices
[2] Position Accepted Mutation
[3] Blocks Substitution Matrix
[4] Globally
[5] Locally

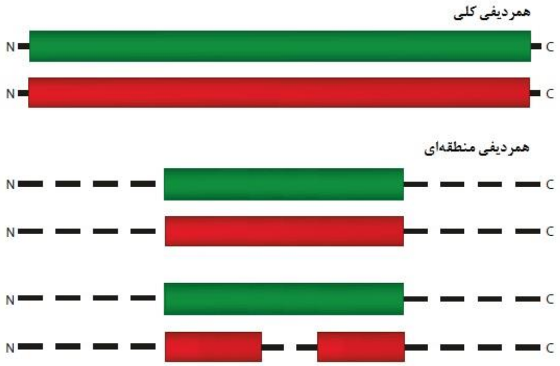
حذف[1] یا اضافه شدن[2]، و در نتیجه تعداد متفاوتی از آمینواسید یا نوکلئوتید را دارند. برای ارائه مناسب این همردیفی‌ها، شکاف‌هایی[3] باید درون این توالی‌ها وارد شود. از لحاظ تئوری همه توالی‌های ممکن با ورود شکاف‌ها همردیف می‌شوند. برای ممانعت از این امر، مقادیر خطای تثبیت شده (خطاهای امتیازدهی) و خطای طویل شده (طویل شدن شکاف) ارائه می‌شوند. سپس این خطاها از امتیاز همردیفی کم می‌شوند تا امتیاز کل به دست آید. در نهایت همردیفی با بالاترین امتیاز،میزان بهینه را در نظر می‌گیرد. این روش بر اساس الگوریتم نیدلمن و وونش (۱۹۷۰) است.

گاهی اوقات ممکن است تمرکز فقط بر همردیفی طول‌هایی با بیشترین شباهت باشد که همردیفی منطقه‌ای نامیده می‌شود (شکل ۳-۳). این رویکرد این امکان را برای شناسایی دامین‌ها و موتیف‌های پروتئینی فراهم می‌کند (مانند نواحی اتصال ATP، دامین‌های اتصال DNA، نواحی N-گلیکوزیلاسیون). در اصل، همردیفی منطقه‌ای مانند همردیفی کلی با استفاده از ماتریس جایگزینی و ورود و طویل شدن شکاف‌ها، محاسبه می‌شود. تفاوت این دو در نحوه محاسبه امتیاز و مسیر ماتریس است. یک امتیاز منفی با صفر جایگزین می‌شود، و بنابراین مسیر ماتریس از راست پایین‌تر به چپ بالاتر حرکت نمی‌کند ولی در مکان‌های تصادفی شروع و پایان می‌پذیرد (شکل ۳-۴). همردیفی منطقه‌ای که در ماتریس با بالاترین امتیاز معرفی شود نقطه بهینه و شروع در نظر گرفته می‌شود. همردیفی زمانی پایان می‌پذیرد که به ورودی صفر برسد. این روش بر اساس الگوریتم اسمیت و واترمن است (۱۹۸۱). برای مقایسه بیش از دو توالی نوکلئوتیدی یا پروتئینی، می‌توان همه توالی‌ها را به صورت جفتی مقایسه کرد و سپس این همردیفی‌ها را بررسی نمود. ولی انجام همردیفی چندتایی (شکل ۳-۵) و تحلیل همردیفی کل سریعتر است. امروزه برنامه[clustalomega] Clustal Omega (Sievers et al. 2011) جایگزین یک برنامه مفید به نام ClustalW (Thompson et al. 1994) شده است، که این برنامه‌ها از این واقعیت استفاده می‌کنند که توالی‌های مشابه، معمولاً هومولوگ هستند. اساس بررسی همردیفی توالی‌های چندتایی، همردیفی جفتی همه توالی‌ها است. به این صورت که، یک درخت فیلوژنی که رابطه تکاملی بین توالی‌ها را در ساختار درختی ارائه می‌دهد، ساخته می‌شود. فواصل تکاملی با طول شاخه‌های افقی ارتباط مستقیم دارند (شکل ۳-۶). در این

---

[1] Deletion
[2] Insertion
[3] Gaps

روش، همردیفی چندتایی توالی نهایی با دو توالی بسیار مشابه شروع می‌شود. در مرحله بعد گام به گام مشابه‌ترین توالی بعدی اضافه می‌شود تا زمانی که همردیفی چندتایی نهایی حاصل گردد.



شکل ۳-۳) همردیفی توالی کلی و منطقه‌ای
بیشتر شکاف‌ها در همردیفی کلی ظاهر می‌شوند ولی در همردیفی منطقه‌ای کمتر دیده می‌شود

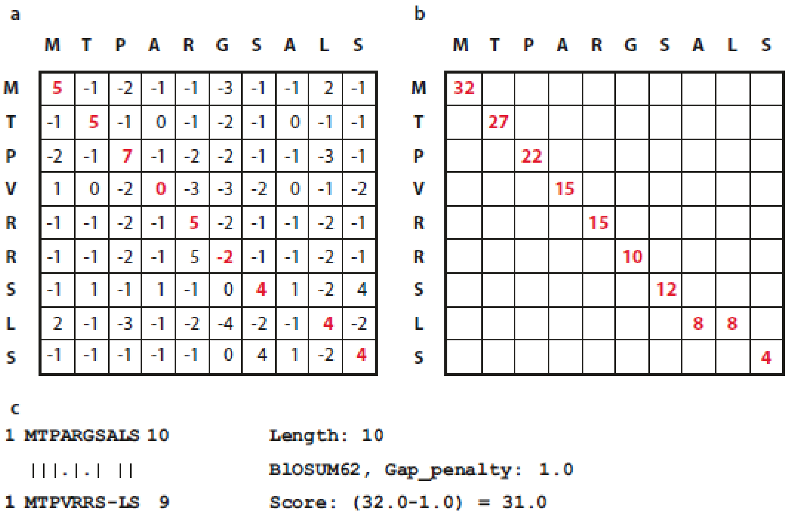
۳-۲ جستجوهای پایگاه داده با توالی‌های نوکلئوتید و پروتئین

یک از فراوان‌ترین کاربرد همردیفی‌های دوتایی، جستجو برای توالی‌های پروتئینی یا نوکلئوتیدی در پایگاه‌های داده توالی است. اجرای الگوریتم‌های همردیفی پویای قدیمی‌تر مانند مواردی که توسط اسمیت و واترمان (۱۹۸۱) یا نیدلمن و وونش (۱۹۷۰) طراحی شده‌اند، حتی در رایانه‌های امروزی، بسیار کند عمل می‌کنند. در عوض، امروزه الگوریتم‌هایی مانند ابزار جستجوی همردیفی منطقه‌ای پایه[1] (BLAST) (Altschul et al. 1990; Boratyn et al. 2013) به کار گرفته می‌شوند (شکل ۳-۷ و ۳-۸). روش‌های امروزی، به دست آوردن نتایج دقیق و استفاده از توالی و همردیفی را برای جستجو در پایگاه‌های داده بزرگ ممکن می‌سازند. این روش‌ها همردیفی بهینه را تضمین نمی‌کنند، ولی منجر به دسترسی به جستجوهای حساس و سریع می‌شوند. پایگاهBLAST به صورت خدمات وب در سایت‌های NCBI[ncbi-blast]، EMBL-EBI[embl-blast] و DDBJ[ddbj-blast]قابل استفاده است. BLAST

---

[1] Basic Local Alignment Search Tool

ابتدا در پایگاه‌های داده بدون افزونه[1] ثبت می‌شوند که وظیفه گردآوری ورودی‌ها از پایگاه‌های داده مختلف را انجام می‌دهند. در یک پایگاه داده بدون افزونه، ورودهای چندگانه حذف می‌شود بنابراین هر ثبت فقط یک بار در دسترس است. این پایگاه‌های داده هم برای توالی‌های نوکلئوتیدی و هم پروتئینی قابل استفاده‌اند.



a

	M	T	P	A	R	G	S	A	L	S
M	5	-1	-2	-1	-1	-3	-1	-1	2	-1
T	-1	5	-1	0	-1	-2	-1	0	-1	-1
P	-2	-1	7	-1	-2	-2	-1	-1	-3	-1
V	1	0	-2	0	-3	-3	-2	0	-1	-2
R	-1	-1	-2	-1	5	-2	-1	-1	-2	-1
R	-1	-1	-2	-1	5	-2	-1	-1	-2	-1
S	-1	1	-1	1	-1	0	4	1	-2	4
L	2	-1	-3	-1	-2	-4	-2	-1	4	-2
S	-1	-1	-1	-1	-1	0	4	1	-2	4

b

	M	T	P	A	R	G	S	A	L	S
M	32									
T		27								
P			22							
V				15						
R					15					
R						10				
S							12			
L								8	8	
S										4

c

```
1 MTPARGSALS 10          Length: 10
  |||.|.| ||             BlOSUM62, Gap_penalty: 1.0
1 MTPVRRS-LS  9          Score: (32.0-1.0) = 31.0
```

شکل ۳-۴) محاسبه همردیفی کلی دو توالی پروتئینی مشابه

a) هر دو توالی در ماتریس دو بعدی مقایسه شده‌اند، و شباهت آمینواسیدها با استفاده از ماتریس‌های شباهت تعیین شده‌اند. هر همردیفی را می‌توان به صورت مسیری از طریق ماتریس دو بعدی تشریح کرد، که با بالاترین نمره‌دهی جفت آمینواسید در انتهای N شروع می‌شود. b) اضافه کردن مقادیر نمره‌های مربوطه برای مسیرهای مختلف تولید می‌شود. همردیفی با بالاترین نمره بهینه در نظر گرفته می‌شود (قرمز). c) همردیفی بهینه با ورود شکاف به دست می‌آید و حاوی ۱۰ آمینواسید است که ۷ تا یکسان هستند. با استفاده از ماتریکس شباهت BLOSUM62 و خطای شکاف 1.0 نمره 31.0 به دست می‌آید.

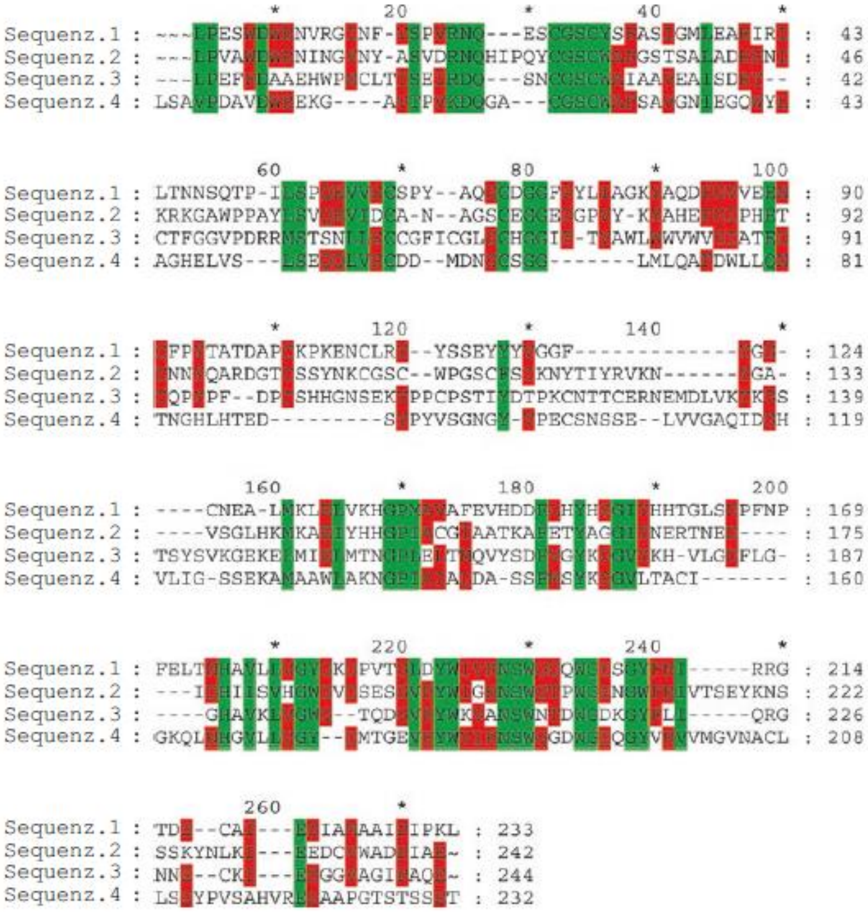
برای اجرای یک جستجوی معنادار در یک پایگاه داده نوکلئوتیدی یا پروتئینی، الگوریتم مربوطه باید از گروه BLAST انتخاب شود، که به توالی هدف (نوکلئوتید یا پروتئین) بستگی دارد (جدول ۳-۱). برای مثال، برای درخواست یک پایگاه داده نوکلئوتیدی با توالی پروتئینی، هر توالی نوکلئوتیدی پایگاه داده، باید به همه شش توالی پروتئین احتمالی ترجمه شود که

---

[1] Nonredundant Database

برای خوانش و نقطه شروع سه‌تایی ناشناخته است (شکل ۳-۱). فقط بعد از آن می‌توان توالی درخواست را با پایگاه داده مقایسه کرد. این فرایند پیچیده به صورت خودکار با الگوریتم tblastn انجام می‌شود. بسته به نوع درخواست و پایگاه‌های داده مورد استفاده، همه پنج الگوریتم ممکن است بکار گرفته شوند (جدول ۳-۱).

در مواردیکه همسانی توالی بسیار اندکیمشاهده می‌شود، ممکن است همردیفی ظاهراً معنی‌دار ولی تصادفی یافت می‌شود که بر اساس توالی جد مشترک نیستند. بنابراین،می‌توان از شاخصE-value (مقدار خطا) که بخشی از نتایجBLAST است برای ارزیابی اهمیت همردیفی استفاده شود. همردیفی‌های دوتایی با $E\text{-value} < 0.02$ بر اساس توالی‌های همولوگ، معنی‌دار محسوب می‌شوند. همردیفی‌های تصادفی $E\text{-values} > 1$ را نشان می‌دهند.



شکل ۳-۵) همردیفی چندگانه توالی چهار پروتئین مرتبط

آمینواسیدهای محافظت شده در همه چهار توالی به رنگ سبز درآمده‌اند؛ آن‌هایی که در سه تا از چهار توالی محافظت شده هستند با قرمز مشخص شده‌اند

درون خانواده BLAST الگوریتم‌های مختلفی مانند: BLAST جایگاه اختصاصی تکراری (PSIBLAST)[۱] (Altschul et al. 1997)، BLAST شروع شده با الگو (PHI-BLAST)[۲] (Zhang et al. 1998)، و bl2seq (بلاست دو توالی)[۱] (Tatusova and

---

[1] Position-Specific Iterated BLAST

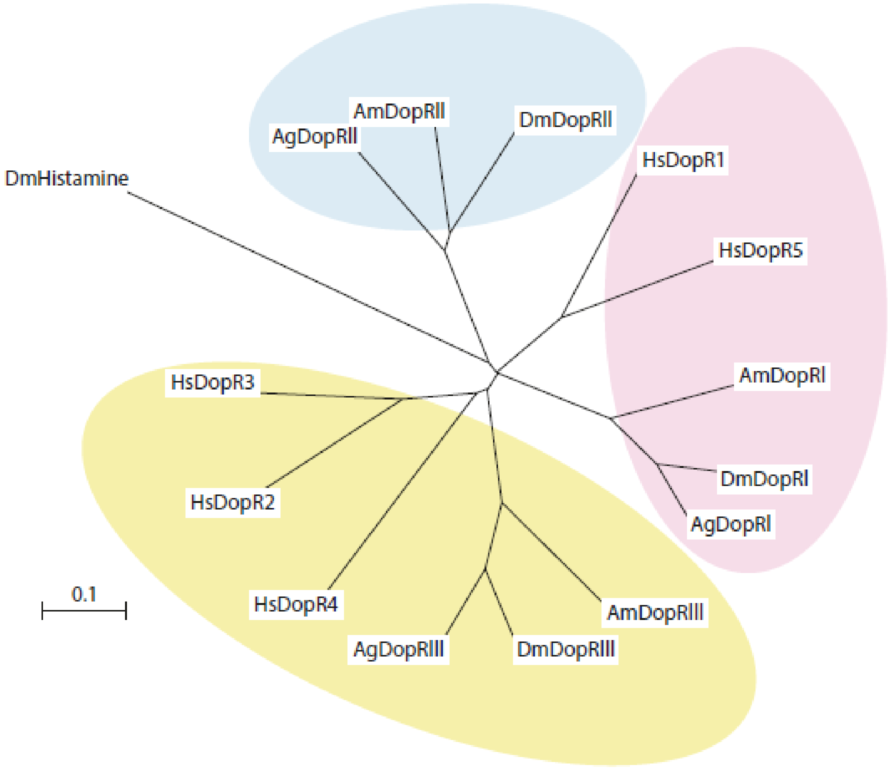
[2] Pattern-Hit Initiated BLAST

Madden 1999) هم به صورت اختصاصی مورد استفاده قرار می‌گیرند. الگوریتم bl2seq یک همردیفی منطقه‌ای از دو توالی را اجرا می‌کند. PHI-BLAST به جستجوی پروتئین در یک پایگاه داده با موتیف‌های مشابه موارد درخواست می‌پردازد.PSI-BLAST مخلوطی از همردیفی دوتایی و چندتایی است. ابتدا، یک جستجوی BLAST طبیعی انجام می‌شود. به همراه همردیفی چندگانه حاصل، یک پروفایل توالی ساخته می‌شود که تا جستجو برای توالی‌های جدید ادامه یابد تا زمانی که مورد دیگری یافت نشود. تفسیر نتایج معمولاً بسیار سخت و گمراه‌کننده است زیرا توالی‌هایی که مستقیماً به هم مرتبط نیستند نیزبررسی می‌شوند. بنابراین، نتایج PSI-BLAST به بررسی دقیق نیاز دارد. مدل‌های مخفی مارکوف (HMMs)[۲] (Eddy 2004) نیز با حالت مشابهی عمل می‌کند، ولی آهسته‌تر و با حساسیت بیشتر. همچنین، نتایج حاصل از HMMsنیز باید منتقدانه بررسی شوند. جستجوی دامین‌های محافظت شده توسط پایگاهNCBI، دامین‌های محافظت شده را درون توالی‌های مورد تحلیل شناسایی می‌کند (Marchler- Bauer et al. 2015). همچنین در BLASTکاربردهای اختصاصی ژنوم برای ژنوم‌های انسانی، میکروبی و سایرین وجود دارد. این موارد در صفحه وب NCBI-BLAST موجود است [ncbi-blast].

---

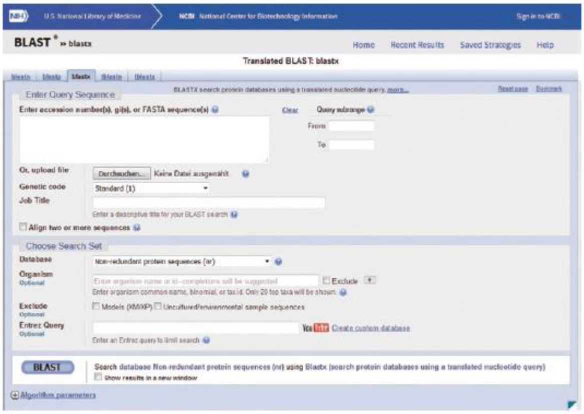
[1] BLAST Two Sequences

[2] Hidden Markov models



شکل ۳-۶) درخت فیلوژنی توالی‌های گیرنده دوپامین

رابطه تکاملی بین توالی‌ها، با طول شاخه‌ها مشخص می‌شود. توالی‌های گیرنده دوپامین بی مهره‌گان (*Dm*، *Drosophila melanogaster*؛ *Ag*، *Anopheles gambiae*؛ *Am*، *Apismellifera*) با انسان مقایسه شده‌اند (*Hs*، *Homo sapiens*). سه خوشه مشخص تشکیل شده است. توالی دور فیلوژنتیکی گیرنده هیستامین *Dm* به عنوان نمونه شاهد است که در هیچ خوشه‌ای یافت نمی‌شود



Genetic code: Standard (1)

Database: Non-redundant protein sequences (nr)

Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)

شکل ۳-۷) صفحه شروع BLAST در پایگاه NCBI.
الگوریتم blsatx برای مقایسه توالی نوکلئوتید با پایگاه داده توالی پروتئین استفاده شده است

۳-۲-۱ الگوریتم‌های مهم برای جستجوی پایگاه داده

نیدلمن و وونش (۱۹۷۰)، روش همردیفی کلی بدون شکاف را ایجاد کردند. این روش در مقایسه با محاسبه همه همردیفی‌های ممکن، مناسب‌تر و سریعتر است. این روش به دلیل مکانیسم زمان‌بر در تحلیل داده‌ها هنوز برای تحلیل پایگاه داده‌های بزرگ مناسب نیست.

اسمیت و واترمن (۱۹۸۱)، همردیفی منطقه‌ای، براساسعملکرد بدون شکاف را ایجاد نمودند. این روش بسیار مشابه روش نیدلمن و وونش زمان‌بر است.

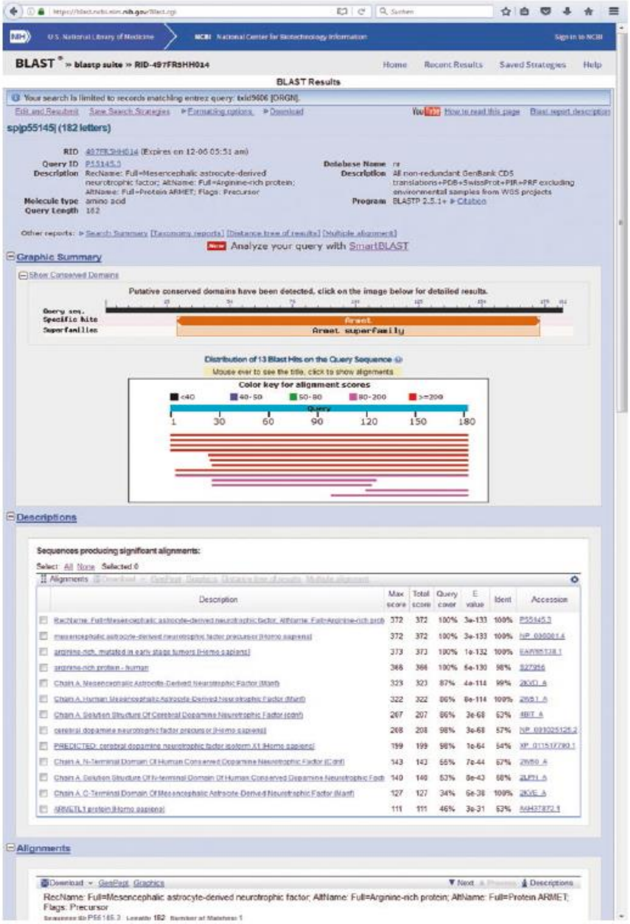
FastA (پیرسون و لیپمن ۱۹۸۸)، همردیفی منطقه‌ای را با استفاده از روش ذهنی[1] بسیار سریع ارزیابی نمودند. این روش نواحی کوتاه را شناسایی می‌کند تا یک همردیفی با شکاف را به دست آورد.

BLAST بدون شکاف (آلتشول و همکاران، ۱۹۹۰)، همردیفی منطقه‌ای را با استفاده از روش ذهنی انجام می‌دهد. قطعات سپس تا جایی که پارامترهای آستانه به دست آیند طویل می‌شوند. روشBLAST تا حدود ۱۰۰ برابر سریع‌تر و کارآمدتر از الگوریتم اسمیت و واترمن است.

---

[1] Heuristic Method

BLAST شکاف‌دار (آلتشول و همکاران، ۱۹۹۷)،همردیفی منطقه‌ای، با شکاف‌های هر دو جهت طویل می‌شود. الگوریتم BLAST شکاف‌دار سه برابر سریع‌تر از الگوریتم BLAST بدون شکاف است.



BLAST Results

Your search is limited to records matching entrez query: txid9606 [ORGN].

sp|p55145| (182 letters)

RID 497FR5HH014 (Expires on 12-06 05:51 am)
Query ID P55145.3
Description RecName: Full=Mesencephalic astrocyte-derived neurotrophic factor; AltName: Full=Arginine-rich protein; AltName: Full=Protein ARMET; Flags: Precursor
Molecule type amino acid
Query Length 182
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.5.1+ ▶ Citation

Other reports: ▶ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

New Analyze your query with SmartBLAST

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 13 Blast Hits on the Query Sequence

Mouse over to see the title, click to show alignments

Sequences producing significant alignments:

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	RecName: Full=Mesencephalic astrocyte-derived neurotrophic factor; AltName: Full=Arginine-rich prot	372	372	100%	3e-133	100%	P55145.3
☐	mesencephalic astrocyte-derived neurotrophic factor precursor [Homo sapiens]	372	372	100%	3e-133	100%	NP_006001.4
☐	arginine-rich, mutated in early stage tumors [Homo sapiens]	373	373	100%	1e-132	100%	EAW65138.1
☐	arginine-rich protein - human	366	366	100%	6e-130	98%	S27956
☐	Chain A, Mesencephalic Astrocyte-Derived Neurotrophic Factor (Manf)	323	323	87%	4e-114	99%	2KVD_A
☐	Chain A, Human Mesencephalic Astrocyte-Derived Neurotrophic Factor (Manf)	322	322	86%	8e-114	100%	2W51_A
☐	Chain A, Solution Structure Of Cerebral Dopamine Neurotrophic Factor (cdnf)	207	207	86%	3e-68	63%	4BIT_A
☐	cerebral dopamine neurotrophic factor precursor [Homo sapiens]	208	208	98%	3e-68	57%	NP_001025125.2
☐	PREDICTED: cerebral dopamine neurotrophic factor isoform X1 [Homo sapiens]	199	199	98%	1e-64	54%	XP_011517790.1
☐	Chain A, N-Terminal Domain Of Human Conserved Dopamine Neurotrophic Factor (Cdnf)	143	143	55%	7e-44	67%	2W50_A
☐	Chain A, Solution Structure Of N-terminal Domain Of Human Conserved Dopamine Neurotrophic Fact	140	140	53%	8e-43	68%	2LPN_A
☐	Chain A, C-Terminal Domain Of Mesencephalic Astrocyte-Derived Neurotrophic Factor (Manf)	127	127	34%	6e-38	100%	2KVE_A
☐	ARMETL1 protein [Homo sapiens]	111	111	46%	3e-31	63%	AAH37872.1

RecName: Full=Mesencephalic astrocyte-derived neurotrophic factor; AltName: Full=Arginine-rich protein; AltName: Full=Protein ARMET; Flags: Precursor
Sequence ID: P55145.3 Length: 182 Number of Matches: 1

شکل ۳-۸) نمایش گرافیکی نتیجه BLAST

گراف تعداد و طول نتایج را با توجه به توالی درخواست خلاصه می‌کند. کیفیت (نمره همردیفی) نتایج با کدگذاری رنگ نمایش داده می‌شود

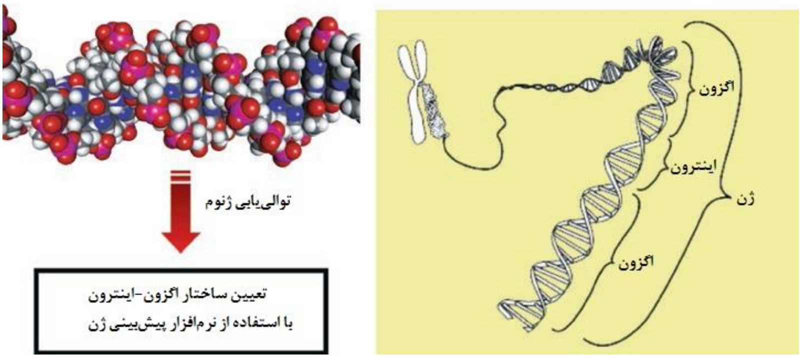
۳-۳ نرم‌افزار برای تحلیل توالی

علاوه بر توالی‌های ژن و پروتئین، پایگاه‌هایNCBI، EBI و دیگر سرورهای قابل دسترس عمومی نیز توالی‌های ژنومی‌را فراهم می‌کنند. چنین توالی‌هایی معمولاً اطلاعات خام هستند زیرا آن‌ها مستقیماً توسط واحدهای توالی‌یابی مانند: مؤسسه سنگر به چاپ رسیده‌اند [sanger]. فایده داده‌های خام توالی این است که ژن‌های پیش‌بینی شده مستقیماً شناسایی می‌شوند (شکل ۳-۹). تعدادی از نرم‌افزارها برای پیش‌بینی‌های ژن در وب ارائه شده‌اند. سرور Gensscanتحت کنترل مؤسسه تکنولوژی ماساچوست [genscan]می‌باشد و بر اساس شبکه‌های عصبی فعالیت می‌کنند تا ساختار اگزون-اینترون را از ژن‌های یوکاریوتی در توالی‌های ژنومی استخراج کنند. نتیجه عمومی‌تحلیل Genscan در شکل ۳-۱۰ نشان داده شده است. نرم‌افزار دیگر موجود برای پیش‌بینی ژن در توالی‌های پروکاریوتی،Glimmer از مؤسسات TIGR (بخش جدیدی از مؤسسه جی. کریگ ونتر) است، که نسخه سوم این نرم‌افزار در مرکز زیست‌شناسی دانشگاه جان‌هاپکینز موجود می‌باشد[glimmer].

جدول ۳-۱) الگوریتم‌های بسیار مهم BLAST و کاربردهای آن‌ها

الگوریتم	**توالی هدف**	**نتیجه درخواست**	**توضیحات**
blastp	پروتئین	پروتئین	توالی یک پروتئین را با پایگاه اطلاعاتی پروتئینی مقایسه می‌کند
blastx	نوکلئوتید	پروتئین	توالی نوکلئوتیدی هدف به صورت شش قالب خواندنی ترجمه شده و بصورت پروتئینی در پایگاه‌های اطلاعاتی پروتئینی جستجو می‌شود
blastn	نوکلئوتید	نوکلئوتید	توالی نوکلئوتیدی را با یک پایگاه اطلاعاتی نوکلئوتیدی مقایسه می‌کند
tblastn	پروتئین	نوکلئوتید	توالی مورد هدف پروتئینی است که بصورت شش قالب خواندنی ترجمه شده، و بصورت توالی پروتئین با پایگاه اطلاعاتی نوکلئوتیدی مقایسه می‌شود
tblastx	نوکلئوتید	نوکلئوتید	توالی هدف نوکلئوتیدی و توالی درخواست نوکلئوتیدی هردو در شش قالب خواندنی ترجمه شده و سپس ترجمه‌ها در سطح اسیدآمینه باهم مقایسه می‌شوند

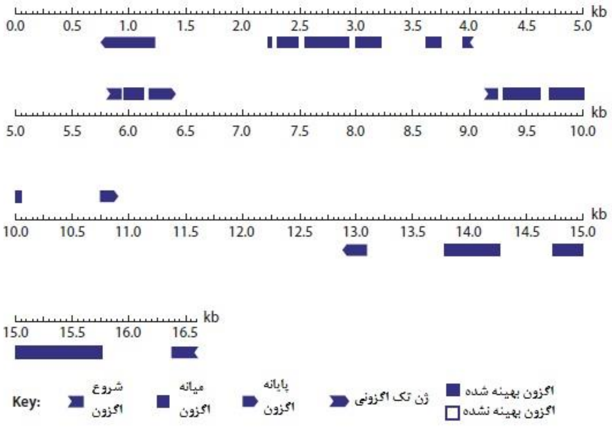
امروزه پیشرفت جالب در حوزه تحلیل توالی، تلفیق نرم‌افزار باز زیست شناسی مولکولی اروپا[1] (EMBOSS) است (Rice et al. 2000) [emboss]. EMBOSS یک منبعقابل دسترس برای سامانه‌های اجرایی UNIX و Linux است. دامنه عملکردی بسته نرم‌افزار به صورت ثابت رشد می‌کند و با بسته‌های تجاری مانند Wisconsin GCG (Biova)، DNA-Star (DNASTAR Inc.)، یا نرم‌افزار Vector NTI (Thermo Fischer Scientific Inc.) قابل مقایسه است. همچنین به پایگاه‌های Expasy[expasy] و EMBnet[embnet]نیز باید اشاره شود. علاوه بر پایگاه داده‌ها، Expasy تعدادی از زیرنویس‌ها را برای نرم افزارهای بیوانفورماتیکی ارائه می‌دهد. پایگاه EMBnet یک رابطه بین پایگاه‌ها و مؤسسات جستجوی مختلف و چند نرم افزار آزاد برای تحلیل توالی ارائه می‌دهد.



شکل ۳-۹) شناسایی ژن‌ها و پروتئین‌های جدید با استفاده از توالی‌یابی ژنوم

---

[1] European Molecular Biology Open Software Suite



شکل ۳-۱۰) نتیجه شماتیک از آنالیز نرم‌افزار Genscan

سایت‌های مفید

bioedit. http://www.mbio.ncsu.edu/bioedit/bioedit.html
blast. https://blast.ncbi.nlm.nih.gov
clustalomega. http://www.ebi.ac.uk/Tools/msa/clustalo/
ddbj-blast. http://ddbj.nig.ac.jp/blast/blastn?lang=en
embnet. http://www.embnet.org/
embl-blast. https://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html
emboss. http://emboss.sourceforge.net/
expasy. https://www.expasy.org/
genscan. http://genes.mit.edu/GENSCAN.html
glimmer. http://ccb.jhu.edu/software/glimmer/index.shtml
ncbi. http://www.ncbi.nlm.nih.gov/
ncbi-blast. http://www.ncbi.nlm.nih.gov/blast/
sanger. http://www.sanger.ac.uk/
seaview. http://doua.prabi.fr/software/seaview
treeview. http://etetoolkit.org/treeview/

منابع

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol, 215: 403-410.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

3. Boratyn GM, Camacho C, Cooper PS et al (2013) BLAST: a more efficient report with usability improvements. Nucleic Acids Res 41:W29-W33.
4. Dayhoff MO, Schwartz RM, Orcutt BC (1978) In: Dayhoff MO (ed) Atlas of protein sequence and structure,Vol. 5, Suppl. 3. NBRF, Washington, DC, p 345.
5. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 10: 1315–1316.
6. Gerlt J, Babbitt P (2001) Respond: Orthologs and paralogs – we need to get it right. Genome Biol, 2(8): 1002.1-1002.3.
7. Henikoff SB, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919.
8. Jensen RA (2001) Orthologs and paralogs – we need to get it right. Genome Biol 2(8):INTERACTIONS1002.
9. Marchler-Bauer A, Derbyshire MK, Gonzales NR et al (2015) CDD: NCBI’s conserved domain database. Nucleic Acids Res 43:D222–D226.
10. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453.
11. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 4:2444–2448.
12. Ramsden J (2015) Bioinformatics: An Introduction. Springer, New York City
13. Rice P, Longden I, Bleasby A (2000) EMBOSS: The european molecular biology open software suite. Trends Genet 16:276–277.
14. Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.
15. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.
16. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 287: 631-637.
17. Tatusova TA, Madden TL (1999) Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247–250.
18. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.
19. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26:3986–3990

فصل چهارم

رمزگشایی ژنوم یوکاریوتی

٤-۱توالی‌یابی کامل ژنوم

یک دوره جدید در تحقیق ژنوم در سال ۱۹۹۵ با انتشار اولینتوالیکامل ژنوم باکتریاییاز بیمارگر انسانی با نام*Haemophilus influenzae* کشف شد. برای اولین بار می‌توان یک ژنوم کامل را تحلیل کرد که شامل ژن‌ها و مناطق تنظیمی آن‌ها می‌شود. سه سال بعد، توالی‌یابی اولین ژنوم یوکاریوتی چند سلولی، نماتد *Caenorhabditis elegans*، نیز تکمیل شد. ژنوم یوکاریوتی، بسیار بزرگتر و پیچیده‌تر از ژنوم باکتریایی است (۷-۷). مقایسه ژنوم‌های یوکاریوتی و پروکاریوتی نشان داد که ژن‌هایی که پروتئین‌ها را کد می‌کنند، جزء بسیار کوچکی از ژنوم یوکاریوتی را تشکیل می‌دهند. بنابراین در انسان‌ها و موش‌ها فقط ۱/۴ درصد از ژنوم در واقع ژن‌های کدکننده هستند و تنها ۵ درصد از هر دو ژنوم بسیار محافظت‌شده[1] هستند، هرچند هر دو تقریباً ۸۰ درصد از ارتولوژی ژن[2]را دارند. علاوه بر ژن‌های کدکننده پروتئین، مناطق حفاظت‌شده حاوی عناصر مهم تنظیمی، ژن‌های غیرکدکننده پروتئین[3] و مناطق مهم برای ساختار کروموزوم هستند. با این حال، برای اکثر نواحی ژنوم، اطلاعات کمی در مورد نحوه عملکرد وجود دارد (کنسرسیوم توالی‌یابی ژنوم موش، ۲۰۰۲).

در مراحل اولیه توالی‌یابی، تعداد نسبتاً کم ژن‌های شناسایی شده در ژنوم انسان شگفت‌انگیز بود. در ابتدای پروژه توالی‌یابی ژنوم انسان، برآورد شده بود که تعداد ژن‌ها در حدود ۱۰۰۰۰۰ تا ۱۵۰۰۰۰ ژن باشند. با این حال تا به امروز، تنها ۱۹۰۰۰-۲۰۰۰۰ ژن مطرح شده است (ازکردیا و همکاران، ۲۰۱۴). تعداد ژن مشابهی نیز برای ژنوم موش پیش‌بینی شده است. جالب توجه است که تعداد ژن‌های انسان تنها حدود ۳۰۰۰ ژن بیش از نماتد *C. elegans* می‌باشد. با توجه به این واقعیت که بدن انسان حاوی چند میلیارد سلول است، در حالی که *C. elegans* تنها دارای ۹۵۹ سلول سوماتیکی است، این تفاوت کوچک در تعداد ژن‌ها بسیار قابل توجه است.

---

[1] Highly Conserved

[2] Gene Orthology

[3] non-Protein-Encoding Genes

۲-٤ مشخص کردن ژنوم‌ها با استفاده از توالی‌هایSTS و EST

۱-۲-٤نواحی برچسب شده توالی نشانه‌هایی در ژنوم انسانی هستند

موفقیت در توالی‌یابی کل ژنوم انسانی دستاوردی بسیار بزرگ بود. بیش از سه میلیارد نوکلئوتید باید توالی‌یابی و در ترتیب صحیح تکثیر می‌شدند. به یک معنا، پروژه را می‌توان در مقایسه با ساختن یک پازل پیچیده مقایسه کرد. برای اولین بار لازم بود که نشانه‌ها را در ژنوم ایجاد نمایندکه باعث موقعیت صحیح مناطق توالی شود. مهمترین نشانه‌هایی که در ژنوم وجود دارد، نواحی برچسب شده توالی (STS)[1] هستند، توالی‌های کوتاه DNAبه طول ۲۰۰-۵۰۰ نوکلئوتید که تنها در ژنوم یک موجود زنده وجود دارد. STS توسط واکنش زنجیره‌ای پلیمراز (PCR)[2]تولید می‌شود، روشی که برای تکثیر توالی‌های اختصاصی استفاده می‌شود. از آنجا که STSها منحصر به فرد هستند، آن‌ها همیشه می‌توانند به وسیلهPCR از DNA ژنومی تکثیر شوند.

کلون‌های DNA با کمک جستجو در پایگاه داده‌ها جهت وجود انطباق در مناطق STS مورد بررسی قرار می‌گیرند و سپس در کروموزوم‌ها یا ژنوم‌ها قرار داده می‌شوند. با استفاده از این روش، یک نقشه دقیق فیزیکی ژنوم انسان تولید می‌شود.

یک پایگاه اختصاصی داده برای توالی‌های STS از سال ۱۹۹۴ پدید آمده است؛ این پایگاه dbSTS می‌باشد که در سال ۲۰۱۳ به یک بخش GenBank منتقل شد. در این پایگاه می‌توان تمام اطلاعات موجود برای هر توالی STS را پیدا نمود، که شامل: نام STS، توالی الیگونوکلئوتید مورد نیاز برای تکثیر PCR، اندازه محصول PCR، شرایط مناسب برای انجام PCR و همچنین یافتن توالی نوکلئوتیدی STS.

بلافاصله پس از انتشار مفهوم نقشه‌برداری مبتنی بر STS در سال ۱۹۸۹، مشخص شد که STSها همچنین می‌توانند از کلون‌های DNA مکمل (cDNA)[3] نیز تولید شوند. چنین کلون‌های cDNA از mRNA سلولی حاصل می‌شوند و بنابراین به ژن‌های بیان شده یک سلول مربوط می‌شوند. علاوه بر نقشه‌برداری ژنوم، STSهای حاصل از cDNA نیز می‌توانند

---

[1] Sequence-Tagged Sites (STSs)
[2] Polymerase Chain Reaction (PCR)
[3] complementary DNA (cDNA)

برای تعیین ژن‌های درون ژنوم استفاده شوند. در واقع، تا سال ۱۹۹۶، تنها یک نقشه ژنتیکی از ژنوم انسان تکثیر شده بود.

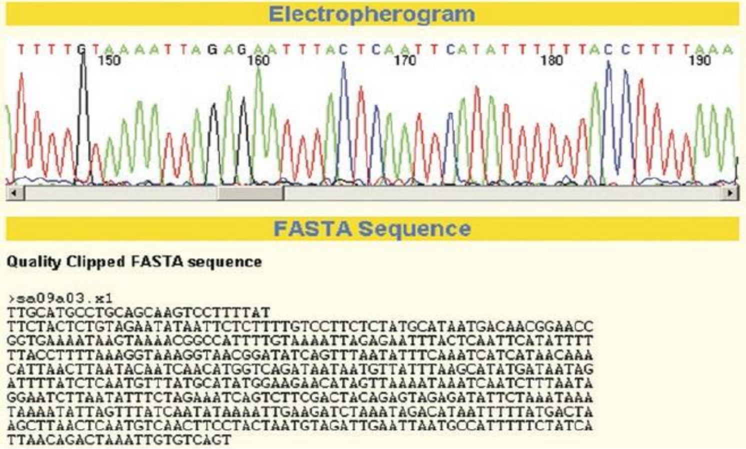
۲-۲-٤برچسب‌های توالی بیان‌شده

متخصصین به سرعت متوجه شدند که می‌توان از توالی‌های ناقص کلون‌هایcDNA در کشف ژن‌های جدید استفاده کنند (آدامز و همکاران، ۱۹۹۱). از آنجا که کلون‌هایcDNA از ژن‌های بیان شده مشتق می‌شوند، این توالی‌ها را، برچسب‌های توالی بیان‌شده (ESTs)[1]نامیدند. ESTها با توالی‌یابی انتهای cDNA تولید می‌شوند (شکل ۴-۱). ESTها با قیمت مناسب به آسانی قابل تولید هستند، و بسیاری از پروژه‌هایEST منجر به شناسایی ژن‌های جدید شده است. با این حال، مفهوم توالی‌یابی EST نیز با مخالفت مواجه شد. منتقدان خاطر نشان کردند که توالی‌یابی cDNA به تنهایی، منطقه‌های تنظیم‌کننده ژنی مهم و بیان نشده را از دست خواهد داد. دوم اینکه، برخی از ESTها بسیار کوتاه‌تر از آن هستند که بتوانند عمل ژن را تعیین کنند و در نهایت،ESTها، که به صورت خودکار تولید می‌شوند، کیفیت توالی ضعیفی دارند. در اغلبESTها، نه تنها تغییرات نوکلئوتیدی رخ می‌دهد، بلکه حذف و اضافه بازها منجر به خطاهایی درچارچوب[2]می‌شود. به سادگی این ترس وجود دارد که بسیاری از پایگاه‌های اطلاعاتی عمومیEST، کیفیت پایینی دارند.

---

[1] Expressed Sequence Tags (ESTs)

[2] Frameshift Errors



```
>sa09a03.x1
TTGCATGCCTGCAGCAAGTCCTTTTAT
TTCTACTCTGTAGAATATAATTCTCTTTTGTCCTTCTCTATGCATAATGACAACGGAACC
GGTGAAAATAAGTAAAACGGCCATTTTGTAAAATTAGAGAATTTACTCAATTCATATTTT
TTACCTTTTAAAGGTAAAGGTAACGGATATCAGTTTAATATTTCAAATCATCATAACAAA
CATTAACTTAATACAATCAACATGGTCAGATAATAATGTTATTTAAGCATATGATAATAG
ATTTTATCTCAATGTTTATGCATATGGAAGAACATAGTTAAAATAAATCAATCTTTAATA
GGAATCTTAATATTTCTAGAAATCAGTCTTCGACTACAGAGTAGAGATATTCTAAATAAA
TAAAATATTAGTTTATCAATATAAAATTGAAGATCTAAATAGACATAATTTTTATGACTA
AGCTTAACTCAATGTCAACTTCCTACTAATGTAGATTGAATTAATGCCATTTTTCTATCA
TTAACAGACTAAATTGTGTCAGT
```

شکل ۴-۱) بخشی از الکتروفروگرام، واکنش توالی‌یابی دی‌داکسی (dideoxy)با توالی‌های نوکلئوتیدی برچسب‌های توالی بیان‌شده (EST) را نشان می‌دهد

با وجود این انتقادات، پروژه‌هایEST به طور گسترده پذیرفته شده‌اند. به طور خاص، سرعتی که ESTها می‌توانند در یک مقیاس با توان بالا[1] تولید شوند (به علت خودکارسازی تکنولوژی توالی‌یابیDNA و تولیدDNAپلاسمید)، موجب رشد واقعی پروژه‌هایEST شده است. برای مثال، پروژه‌های مهم EST در دانشگاه واشنگتن آغاز شده است. این دانشگاه در همکاری با شرکت داروسازی مرک در شهر کنیلورز، ایالت نیوجرسی ایالات متحده، ۵۸۰۰۰۰ EST انسانی را بین سال‌های ۱۹۹۵ تا ۱۹۹۷ توالی‌یابی کردند. اینESTها از کتابخانه‌های cDNA[2]ساخته شده‌اندکه توسط کنسرسیوم تحلیل تلفیق‌شده مولکولی ژنوم و بیان آن‌ها (IMAGE)[3]ایجاد شده بودند، این کنسرسیوم از ادغام چندین گروه تحقیقاتی علمی پدیدار شده است که کتابخانه‌های cDNA با کیفیت بالا را تولید و آن‌ها را برای تحقیقات دیگر مانند پروژه‌های EST در دسترس قرار می‌دهند. کنسرسیوم IMAGE دارای بزرگترین مجموعه از کتابخانه‌های عمومی cDNA در سراسر جهان است.

به دلیل افزایش حجم داده‌های EST، پایگاه dbEST در سایت NCBI برای جمع‌آوری تمام ESTهای قابل دسترس عمومی، تثبیت شده است. در سال ۱۹۹۳، کمتر از ۵۰۰۰۰ توالی در dbEST ذخیره شده بود؛ اما امروزه، بیش از ۷۴ میلیونEST از بیش از ۲۴۰۰ موجود زنده

---

[1] High-Throughput Scale

[2] cDNA Libraries

[3] Molecular Analysis of Genomes and their Expression (IMAGE) consortium

در این پایگاه داده ذخیره شده است (dbESTتائید شده به شماره ۱۳۰۱۰۱ در ژانویه ۲۰۱۷). یکی از نقص‌هایdbEST این است که دارای EST اضافی است، مخصوصاً برای ژن‌های بیان شده بسیار قوی مانند اکتین. به همین دلیل، پایگاه داده UniGene پدید آمد که در آن تمامی داده‌های cDNA و EST که از یک ژن یکسان منشأ می‌گیرند، در یک گروه یا خوشه ترکیب می‌شوند. نتیجه این کار، کاهش تعداد ورودی‌ها به تعداد واقعی پروتئین تولید شده در یک موجود است. به علت وجود عدم تکرار در ESTها، UniGene یک پایگاه مفید برای دیگر بانک‌های اطلاعاتی مانند ProtEST و HomoloGene است. پایگاه ProtEST در UniGene تلفیق شده و اطلاعاتی در مورد این موضوع بیان می‌کند که، آیا cDNAها و ESTهایی که به یک خوشه‌ی UniGene اختصاص داده می‌شوند، مشابه نتایج توالی پروتئین شناخته شده‌ی بعد از ترجمه هستند یا خیر. در مقابل، پایگاه داده مستقل HomoloGene اطلاعاتی در مورد اینکه آیا خوشه‌های UniGene انسانی در سایر گونه‌ها همولوگ دارند یا خیر، فراهم می‌آورد.

یکی دیگر از پایگاه‌های داده NCBI، dbGSS است که توالی‌های بررسی ژنوم (GSSs)[۱]را ذخیره می‌کند. ماندESTها، GSSها توالی‌های نوکلئوتیدی ناقص با طول تا ۱۰۰۰ باز هستند و توسط کلون‌هایانفرادی توالی‌یابی انتهایی تولیدمی‌شوند. تفاوت GSSها و ESTها درمواد حاصل ازمنبع اسیدنوکلئیک است: GSSها از کتابخانه‌های ژنومی تهیه می‌شوند، در حالی که از کتابخانه‌های cDNA برای EST استفاده می‌شوند. بنابراین، GSSها از ESTها از نظر قطعات DNA که در خارج از ژن‌های کدکننده قرار دارند، متفاوت هستند. بیش از ۳۵ میلیون توالی از بیش از ۱۰۰۰ موجود زنده در dbGSS ذخیره شده‌اند (نسخه dbGSS۱۳۰۱۰۱، ژانویه ۲۰۱۷).

اگرچه اهمیت پروژه‌هایEST در طول سال‌ها کاهش یافته است، اما یک مرور کلی از نحوه انجام یک پروژه EST ارائه خواهیم داد زیرا اصول پایه‌ای، مشابه روش توالی‌یابی‌های پیشرفته با عملکرد بالا (بخش ۴-۶-۳)می‌باشد. اگر ما پروژه EST را باتوالی‌یابی کامل نسخه‌برداری شاتگان[۲]مقایسه کنیم،که با نام RNA-Seq هم شناخته می‌شود، شباهت در این دو رویکرد به آسانی قابل تشخیص است (وانگ و همکاران ۲۰۰۹). هر دو روش با ایجاد یک کتابخانه cDNA آغاز می‌شود. علاوه بر این، اگر به یاد داشته باشیم که چگونه یک پروژه ESTانجام می‌شود مراحل بعدی برای توالی‌یابی با عملکرد بالا را راحت‌تر درک می‌کنیم.

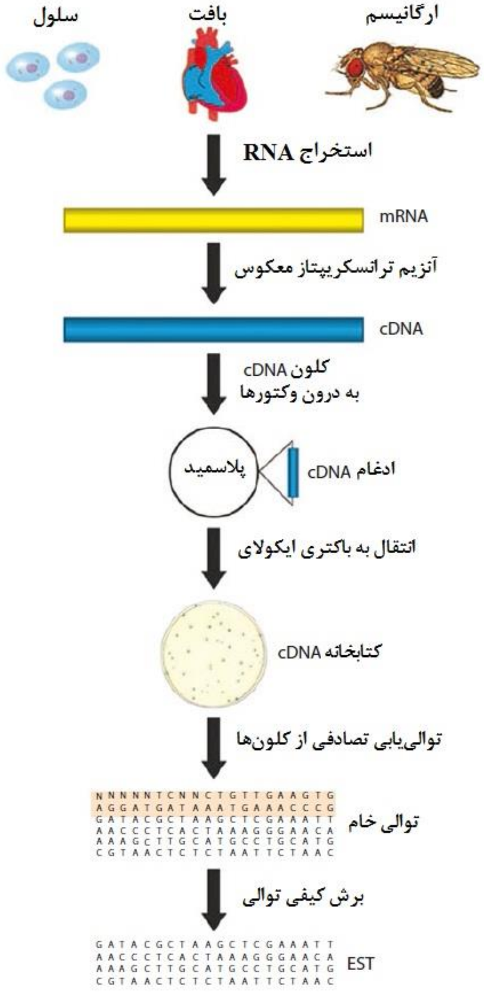
---

[1] Genome Survey Sequences (GSSs)

[2] Whole Transcriptome Shotgun Sequencing

۳-٤ پیاده‌سازی پروژه EST

در مراحل اولیه شروع پروژه EST ماده اولیه برای ساخت کتابخانهcDNA انتخاب می‌شود. اینکتابخانه می‌تواند سلول‌ها، بافت‌های خاص یا حتیکل موجودات زنده باشد (شکل ۴-۲). از این مواد، کل RNA جدا می‌شود که عمدتاً شامل RNA ریبوزومی(rRNA)، RNA ناقل (tRNA) و RNAپیامبر (mRNA) است.



شکل ۴-۲) نمودار مربوط به ساخت کتابخانه cDNA و تولید توالی‌های EST است

مهمترین بخش در ساخت یک کتابخانه cDNA، وجود mRNAاست زیرا تمام ژن‌های فعال از یک سلول یا بافت خاص را نشان می‌دهد. این امر تنها در مقادیر بسیار کمی(حدود ۳ درصد از کل RNA) موجود است. mRNA ناپایدار به cDNAپایدار،بوسیله آنزیمترانسکریپتاز معکوس ویروسی[۱]، رونویسی می‌شود. سپس cDNA درون پلاسمیدهایی که به عنوان ناقل عمل می‌کنند، کلون می‌شود. معمولاً cDNAها به صورت مستقیم کلون می‌شوند، یعنی در یکی از پایانه‌های ′۵ و ′۳ ناقل cDNA جاگذاری می‌شوند. پلاسمیدها با انتقال به باکتری *Escherichia coli* تکثیر می‌شوند،که کتابخانه cDNA مورد نظر حاصل می‌شود و می‌تواند مبنای تولید توالی‌های EST باشد. باکتری‌های ترانسفرم شده بر روی محیط کشت قرار گرفته و رشد می‌کنند و DNA پلاسمید از کلون‌های تکی انتخاب شده جدا می‌شوند. سپس cDNA کلون شده از انتهای'۵ یا'۳ یا از دو طرف به طور همزمان توالی‌یابی می‌شود. نهایتاً توالی نوکلئوتیدی شناسایی شده به یک کامپیوتر منتقل می‌شود و داده‌های خام از نظر بیوانفورماتیکی پردازش می‌گردند.

کیفیت داده‌ها برای اولین بار در یک فرایندی به نام پیرایش کیفی[۲]بررسی می‌شوند. به عنوان مثال، پیرایش کیفی، حداقل طولی را تعریف می‌کند که EST باید داشته باشد و تعداد نوکلئوتیدی‌های مبهم (متغیرN) را نسبت به نوکلئوتیدهای غیرمبهم (A/T/G/C) می‌سنجد. توالی‌یاب‌های مدرن اجازه محاسبه نمرات کیفیت را می‌دهند که بیانگر میزان کیفیت توالی‌یابی هر نوکلئوتید تکی است. با استفاده از این مقادیر، مناطق توالی با کیفیت پایین، به عنوان مثال، انتهای توالی‌ها، حذف می‌شوند. در نهایت، هرگونه آلودگی از توالی ناقل و باکتری نیز، حذف می‌شود.

EST حاصل، شامل مجموعه‌ای از توالی‌های cDNA تصادفی با طول‌های متفاوت است که بسیاری از آن‌ها از رونوشت‌های یکسان به دست می‌آیند. ESTهای بسیاری مخصوصاً برای ژن‌هایی با بیان بالا، پیدا خواهند شد. بنابراین برای حذف تکرارها، از همردیفی اینESTها برای ایجاد توالی‌های همپوشانی استفاده می‌شود که تا حد ممکن بلند باشند (شکل ۴-۳).

این توالی‌های یکسان دوباره با سایرESTها مقایسه می‌شود تا ESTهای مشابه بیشتر، در همردیفی بکار گرفته شوند. این فرایند تکراری به عنوان سرهم‌بندی توالی[۳]معرفی می‌شود.

---

[1] Viral Enzyme Reverse Transcriptase

[2] Quality Trimming

[3] Sequence Assembly

اغلب از برنامه‌های مانند:CAP3 و Phrap جهت سرهم‌بندی توالی استفاده به عمل می‌آید. سرهم‌بندی‌های توالی یا کانتیگ‌هایی[۱] هستند که با همردیفی توالی‌های توافقی[۲] مطابقت دارند یا سینگلتون‌هایی[۳]هستند که مشابه با دیگرESTها نیست و نمی‌توانند درکانتیگ‌ها گروه‌بندی شوند.

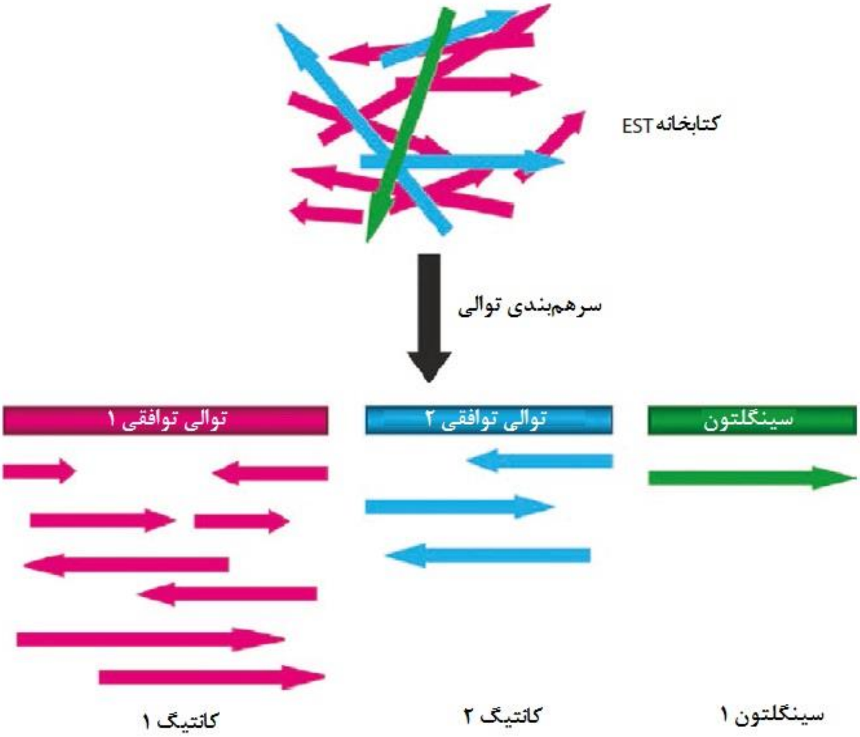
برای مجموعه داده‌های بزرگ EST، می‌توان ابتدا ESTها را به خوشه‌ها یا گروه‌ها تقسیم نمود. خوشه‌هایی که نوکلئوتیدهای یکسان برای یک منطقه مشخص را نشان می‌دهند، در گروه‌ها خلاصه می‌شوند. در نهایت، در داخل این گروه‌ها یک سرهم‌بندی توالی دقیق‌تر برای تولید توالی‌های یکسان انجام می‌شود. به این ترتیب،ESTهایی که از اشکال متناوب حاصل می‌شوند، در خوشه‌های یکسان، اما کانتیگ‌های متفاوتی قرار داده می‌شوند، که این کار روابط EST را بهتر نشان می‌دهد. یک برنامه مفید جهت خوشه‌بندی توالی استفاده از نرم‌افزار stackPACK است.

٤-٤ شناسایی ژن‌های ناشناخته

هنگامی‌که ESTها درون کانتیگ‌ها قرار می‌گیرند، توالی‌های یکسان مربوطه برای شناسایی ژن‌های ناشناخته استفاده می‌شوند. برای این منظور، تفسیر و جستجوی توالی در برابر پایگاه‌های مختلف داده انجام می‌شود.

---

[1] Contigs
[2] Consensus Sequences
[3] Singletons



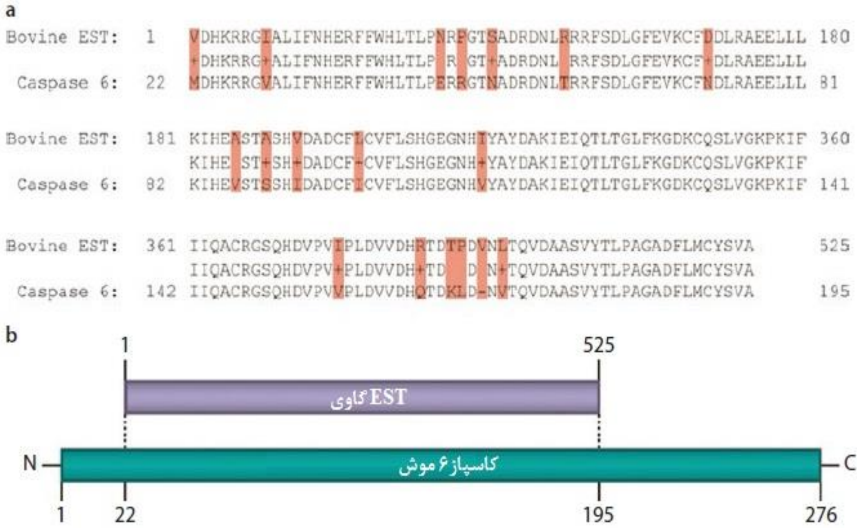
شکل ۴-۳) طبقه‌بندی ESTها درون کانتیگ‌ها و تشکیل توالی توافقی

ESTها معمولاً ابتدا تفسیر می‌شوند، یعنی یک عمل بالقوه برای هر دو سطح از ESTهای تکی و کانتیگ‌های سرهم‌بندی شده، بوسیله مقایسه با پروتئین‌هایی با عملکرد شناخته شده، اختصاص داده می‌شود. معمولاً از الگوریتم BLASTx با استفاده از نوکلئوتیدهای EST استفاده می‌شود. توالی‌ها برای اولین بار به تمام شش قالب خوانش[1]، ترجمه می‌شوند. این فرآیند در شکل ۴-۴ نشان داده شده است که با استفاده از توالیEST از روده گاو، گرفته شده است. در این تصویرEST توسط BLASTx در برابر یک پایگاه داده پروتئین بدون تکرار، تفسیر شد و شباهت زیادی با بخشی از آنزیم کاسپاز۶ موش نشان می‌دهد. کاسپازها، پروتئازهایی هستند که در طول برنامه‌ریزی مرگ سلولی (آپوپتوز)[2] عمل می‌کند. با توجه به

---

[1] Reading Frames
[2] Apoptosis

شباهت به کاسپاز، می‌توان نتیجه گرفت که رونوشت، ژنی که از EST استخراج می‌شود، یک کاسپاز واقعی یا یک پروتئین را کد می‌کند که حاوی یک دامین کاسپاز است. مهم است که بگوییم ESTها معمولاً توالی‌های ناقص ژن هستند و بنابراین همردیفی ممکن است شامل کل طول پروتئین مذکور نشود. در واقع، EST فقط اغلب منطقه ترجمه نشده (UTR)[1] از mRNA ژن را کد می‌کند، و چنین EST به عنوان EST غیرکدکننده شناخته شده است (شکل ۴-۵). با این وجود، زمانی می‌توان از این مشکلات اجتناب کرد که EST توسط سرهم‌بندی توالی، گاهی تا نقطه‌ای که تمام پروتئین را شناسایی کند، گسترش یابد.



```
a
Bovine EST:   1   VDHKRRGIALIFNHERFFWHLTLPNRPGTSADRDNLRRRFSDLGFEVKCFDDLRAEELLL 180
                  +DHKRRG+ALIFNHERFFWHLTLP R GT+ADRDNL RRFSDLGFEVKCF+DLRAEELLL
  Caspase 6: 22   MDHKRRGVALIFNHERFFWHLTLPERRGTNADRDNLTRRFSDLGFEVKCFNDLRAEELLL 81

Bovine EST: 181   KIHEASTASHVDADCFLCVFLSHGEGNHTYAYDAKIEIQTLTGLFKGDKCQSLVGKPKIF 360
                  KIHE ST+SH+DADCF+CVFLSHGEGNH+YAYDAKIEIQTLTGLFKGDKCQSLVGKPKIF
  Caspase 6: 82   KIHEVSTSSHIDADCFICVFLSHGEGNHVYAYDAKIEIQTLTGLFKGDKCQSLVGKPKIF 141

Bovine EST: 361   IIQACRGSQHDVPVIPLDVVDHRTDTPDVNLTQVDAASVYTLPAGADFLMCYSVA      525
                  IIQACRGSQHDVPV+PLDVVDH+TD  D N+TQVDAASVYTLPAGADFLMCYSVA
  Caspase 6: 142  IIQACRGSQHDVPVVPLDVVDHQTDKLD-NVTQVDAASVYTLPAGADFLMCYSVA      195
```

شکل ۴-۴) تفسیر توالی EST روده گاو. a) EST گاوی که حاوی ۱۷۵ اسیدآمینه است (۵۲۵ نوکلئوتید) بیش از ۸۹ درصد طول کاسپاز۶ موش را شناسایی نموده است. توالی‌های متفاوت با رنگ قرمز مشخص شده‌اند. شماره‌گذاری EST گاوی از ۱ تا ۵۲۵ به مربوط به نوکلئوتیدهاست. در مقابل، شماره‌گذاری کاسپاز۶ موش از ۲۲ تا ۱۹۵ مربوط به اسیدهای آمینه است. b) طرح شماتیکی از همردیفی توالی‌های EST گاوی با کاسپاز۶ موش است

با مقایسه مستقیم توالیEST بین موجودات مختلف، ژن‌هایا پروتئین‌های مشابه یا حتی جدید ممکن است شناسایی شوند. به طور کلی، با این حال، توصیه نمی‌شود که در سطح نوکلئوتیدی تلاش شود (به عنوان مثال، با BLASTn) زیرا شباهت کمی بین گونه‌ها با توجه

---

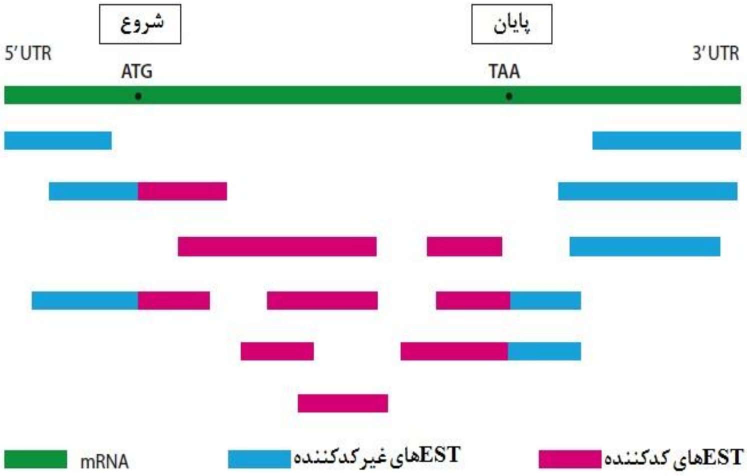
[1] Untranslated Region (UTR)

به نحوه استفاده کدونی[1]وابسته به گونه وجود دارد (فصول ۱ و ۷). با این حال، توالی‌ها به طور معمول محافظت بیشتری در سطح اسیدآمینه نشان می‌دهند. بنابراین، توالی‌ها باید بعد از ترجمه توالی نوکلئوتید با تمام شش قالب خوانش، مقایسه شوند. برای این منظور،tBLASTx، که به طور خودکار هم ترجمه و هم مقایسه پایگاه داده را انجام می‌دهد، انتخاب خوبی است (فصل ۳). با این حال، گاهی لازم است که پایگاه داده‌های بزرگ نیز مورد بررسی قرار گیرند. یک نمونه جالب از مقایسه در مقیاس بزرگ، ارزیابی توالیEST از کرم‌های انگلی مختلف است. در دانشگاه واشنگتن در سنت لوئیس،یک پروژه تعیین توالی نماتد انگل در حال انجام است که در آن بیش از ۳۰۰۰۰۰ EST از کرمک‌های انگلی مختلف در حال توالی‌یابی است. با مقایسه این مجموعه داده‌ها، امکان پیدا کردن ژن‌هایی که در همه نماتدها وجود دارند، مهیا می‌گردد. چنین رویکردی مورد استفاده قرار گرفته است تا روابط تکاملی را در راسته نماتدها مشخص کنیم (بلاکستر، ۱۹۹۸).

با استفاده از داده‌هایEST، اعضای جدید خانواده پروتئین نیز می‌توانند شناسایی شوند. این روش برای شناسایی پروتئین کینازهای جدید در مجموعه داده‌های EST نماتد در شکل ۴-۶ نشان داده شده است. برای شروع، توالی پپتید یک پروتئین کیناز شناخته شده (به عنوان مثال از موش) با پایگاه داده EST (ماننددbEST) مقایسه می‌شود. اگر توالی EST نماتد با همسانی بالا به کینازموش پیدا شود، احتمال دارد که EST یک پروتئین کیناز را کد کند. برای تعیین اینکه آیا پروتئین کیناز شناخته شده جدید است، EST باید با یک پایگاه داده پروتئینی یا نوکلئوتیدی غیرتکراری مقایسه شود. اگر هیچ توالی یکسانی شناسایی نشده باشد، در واقعیک عضو جدید از خانواده پروتئین کیناز یافت شده است.

---

[1] Codon Usage



شکل ۴-۵) توالی‌های EST از قطعات کدکننده و غیرکدکننده mRNA مشتق شده است

٤-٥ کشف انواع پیرایش

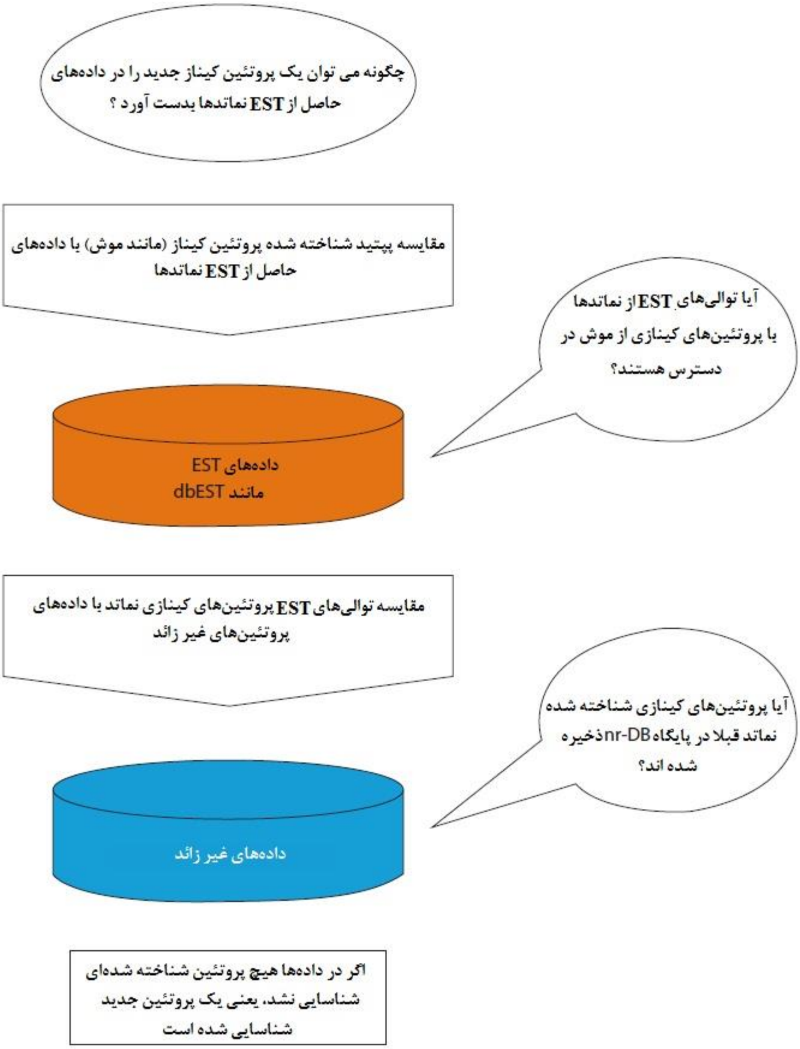
ESTها علاوه بر کمک به شناسایی ژن‌های جدید، همچنینمی‌توانند انواع جایگزین پیرایش[1] ژن را شناسایی کنند. انواع مختلفی از پیرایش جایگزینی[2]می‌توانند بعد از رونویسی ژن و در طول پردازش نسخه‌برداری اولیه RNA بوجود آیند. در طول پیرایش[3]، اینترون‌های غیرکدکننده از رونویسی اولیه حذف می‌شوند و اگزون‌های باقیمانده با هم ترکیب می‌شوند تا mRNA بالغ را تشکیل دهند (فصل ۱). در طی پیرایش جایگزینی، یک اگزون، جایگزین دیگریمی‌شود، در نتیجه یکmRNA جدید ایجاد می‌شود. به این ترتیب،mRNAهای مختلف، که پروتئین‌های مختلف را کد می‌کنند، می‌توانند از رونوشت RNA اصلی اولیه ایجاد شوند(شکل۴-۷). بنابراین، پیرایش جایگزینی استراتژی کارآمدی برای تولید پروتئین‌های مختلف از یک ژن است. اعتقاد بر این است که فرم‌های جایگزین برای یک سوم تا دو سوم همه ژن‌های انسانی وجود دارد (یئو و همکاران، ۲۰۰۴). به عنوان مثال، دو رونوشت mRNA برای گیرنده Fc شناخته شده است که در ایمونولوژی اهمیت دارند. طی پیرایش جایگزینی، دامین

---

[1] Splice
[2] Alternative Splicing
[3] Splicing

سیتوپلاسمی گیرنده برای فرم دوم تبادلمی شود. از آنجا که دامین‌های سیتوپلاسمی تکی برای انتقال پیام، حیاتی هستند، پیرایش جایگزینی دامین‌هایی را با اعمال سلولی بسیار متفاوت ایجاد می‌کنند.

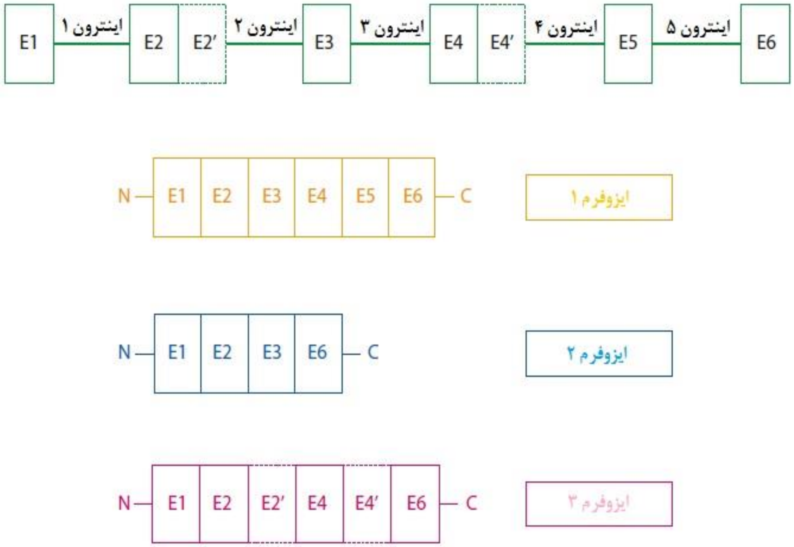
ESTهای مشتق شده از mRNAهای کاملاً پردازش شده، می‌تواند نکات ارزشمندی برای شناسایی انواع پیرایش ناشناخته ارائه دهند. EST با پایگاه داده‌های نوکلئوتیدی که حاوی اطلاعات برای رونوشت‌های mRNA است (به عنوان مثال GenBank) یا با پایگاه داده‌های پروتئینی (به عنوان مثال UniProt) مقایسه می‌شوند. در مواردی که توالی‌های یکسان دیگر در چند منطقه متفاوت هستند، مثلاً با حذف یا اضافه، می‌تواند شواهدی برای انواع پیرایش متناوب باشد. از طریق چنین مقایسه‌های EST با توالی‌های شناخته شده در پایگاه داده‌های عمومی، تعداد زیادی از ژن‌های متناوب تقسیم شده در حال حاضر کشف شده است. در دانشگاه کالیفرنیا در لس‌آنجلس، دو پایگاه داده به نام ASAP و ASAP2 از پروژه تفسیر پیرایش جایگزینی ایجاد شده است که در آن ژن‌های متناوب که توسط توالی‌های EST شناسایی شده‌اند، ذخیره می‌شوند. همچنین، بسیاری از برنامه‌های پیش‌بینی ژن مانند GrailEXP از توالی‌هایEST استفاده می‌کنند تا ژن‌ها را از ژنوم‌های توالی‌یابی شده به طور صحیح پیش‌بینی کنند و اطلاعات مربوط به سایت‌های پیرایش [grailexp] را استخراج کنند.



شکل ۴-۶) استراتژی شناسایی اعضای جدید خانواده‌های پروتئینی

٤-٦ علل ژنتیکی برای تفاوت‌های فردی

یک ویژگی ژنوم یوکاریوتی، وجود جهش یا تغییرات ژنتیکی است. این تغییرات مسئول تفاوت‌های فردی در یک جمعیت هستند. شایع‌ترین تغییرات، چندشکلی تک نوکلئوتیدی (SNP)[1]ناشی از تبادل تک نوکلئوتید است. چندشکلی‌های دیگر عبارتند از: حذف و اضافه‌های کوتاه (چندشکلی حذف و اضافه[2]) و تغییرات به دلیل توالی‌های تکراری (تکرارهای کوتاه تکراری[3]).



شکل ۴-۷) پیرایش جایگزینی. تولید چندین نسخه mRNA از یک ژن با تلفیق چندین اگزون. حرف (E) همان پیرایش جایگزینی است.

یک کنسرسیوم از نهادهای تجاری و غیرتجاری تقریباً ۱/۸ میلیون SNP را در ژنوم انسان شناسایی کرده است (توریسون و استین، ۲۰۰۳). بسیاری از اینSNPها خارج از ژن‌ها قرار دارند و بنابراین عملکرد سلولی را تغییر نمی‌دهند. با این حال، سایر SNPها در داخل ژن‌ها قرار دارند و مسئول ظهور فنوتیپ‌ها هستند. مثال‌هایی از انواع فنوتیپ مانند رنگ چشم و مو، و

---

[1] Single-Nucleotide Polymorphisms (SNPs)

[2] Deletion Insertion Polymorphisms

[3] Short Tandem Repeats

شرایط بیماری هستند. SNPهای مهم عملکردی، با مقایسه ظاهر یک فنوتیپ با فراوانی یک SNP خاص کشف می‌شوند. اگر یک همبستگی پیدا شود احتمال دارد که این SNP مسئول فنوتیپ باشد. از آنجا که افراد به طور تصادفی برای این تجزیه و تحلیل همبستگی انتخاب شده است، این استراتژی ساده‌تر و سریع‌تر از تحلیل نسل کلاسیک است، زیرا در تحلیل کلاسیک ظهور فنوتیپ‌ها باید در یک خانواده در بیش از چند نسل کشف شود.

یک نمونه از بیماری مبتنی بر SNP، فنیل‌کتونوریا[1] است که در آن تجزیه فنیل‌آلانین مختل می‌شود. جهش‌های نقطه‌ای در آنزیم فنیل‌آلانین هیدروکسیلاز منجر به غیرفعال‌سازی آنزیم می‌شود. بسیاری از SNPهای مختلف در آنزیم فنیل آلانین هیدروکسیلاز کشف شده‌اند و در پایگاه داده Phylnalanine Hydroxylase Locus Knowledge Base[pahdb] جمع‌آوری شده‌اند. به علت فعالیت آنزیمی از دست رفته، فنیلآلانین در مغز نوزادان و کودکان تجمع یافته و در نهایت به نقص مغزی منجر می‌شود. بنابراین نوزادان در بسیاری از کشورها برای سطح بالای فنیل‌آلانین مورد بررسی قرار می‌گیرند. علائم بیماری توسط یک رژیم غذایی ضعیف فنیل‌آلانین قابل پیشگیری هستند و به افراد متولد شده امکان زندگی عادی را می‌دهند.

چندشکلی ژنتیکی همچنین می‌تواند یک مزیت باشد. یک مثال از این مورد، حساسیت افتراقی افراد به عفونت توسط ویروس نقص‌ایمنی بدن انسان ۱ (HIV-1) است. علاوه بر پروتئین سطح CD4، این ویروس به گیرنده‌های دیگری مانند گیرنده chemokine CCR5 برای ورود به سلول نیاز دارد. یک جهش از این گیرنده با حذف ۳۲ نوکلئوتید در سال ۱۹۹۶ کشف شد. این جهش منجر به تغییر در چارچوب خوانش و سپس تغییر در ترجمه پروتئین غیرفعال شده است که دیگر در سطح سلول وجود ندارد. انسان‌هایی که برای این جهش هموزیگوت هستند (هر دو آلل‌ها مختل می‌شوند) بیشتر به عفونت HIV-1 مقاوم هستند. کسانی که برای جهش هتروزیگوت (یک آلل عملکردی) هستند، بعداً ایدز را توسعه خواهند داد و امید به زندگی بیشتری نسبت به افرادی که فاقد جهش تغییرچارچوب می‌باشند، دارند.

در جمعیت قفقازی ایالات متحده آمریکا، این چندشکلی در فراوانی ۱ درصد هموزیگوت است و ۲۰ درصد دیگر آلل هتروزیگوت دارند. متأسفانه، در بین جمعیت آفریقا و شرق آسیا، این چندشکلی به ندرت یافت می‌شود (برگر و همکاران ۱۹۹۹).

---

[1] Phenylketonuria

SNPها همچنین نشانگرهای ژنومی عالی هستند زیرا در کل ژنوم پراکنده شده و با تراکم بالا (به طور متوسط هر ۳۰۰ تا ۵۰۰ نوکلئوتید در ژنوم انسان) یافت می‌شوند. علاوه بر این، SNPها دارای فراوانی جهش کم بین نسل‌ها هستند و با روش‌های عملکرد بالا، قابل تشخیص هستند. بنابراین، SNPها اجازه تولید نقشه‌های ژنتیکی دقیق با وضوح بالا را می‌دهند. این وضوح، کشف ژن‌های بیماری را آسان می‌کند، به خصوص اگر ژن‌های متعددی مسئول ظهور بیماری‌های پیچیده مانند سرطان یا دیابت باشند. روش‌هایی برای تشخیص SNP یا بررسی ژنوتیپ وجود دارند. اگر نوکلئوتیدهای غیریکسان موجود باشد، بررسی ژنوتیپ با ریزآرایه[1] بر اساس اصلی است که دمای غیرطبیعی شدن رشته‌های DNA ترکیبی، کاهش می‌یابد. مزیت این روش با عملکرد بالا، این است که آن را برای تجزیه و تحلیل همزمان و موازی بسیاری از توالی‌هاممکن می‌سازد. تکنیک‌های دیگر، برای شناسایی SNPها بر اساس واکنش‌های آنزیمی‌است که اختصاصیت بسیار بالایی را برای سوبسترا نشان می‌دهند و بنابراین دقیق‌تر از روش‌های مبتنی بر هیبریداسیون هستند. یک تکنیک ژنتیکی مبتنی بر آنزیم، معمولاً پایروسیکوئنسینگ[2] است. قطعات کوتاه DNA، در زمان واقعی بدون نیاز به گام‌های زمان‌بر تصفیه ژل، توالی‌یابی می‌شوند. مزیت این روش این است که کل همسایگی SNP توالی‌یابی می‌شود و به عنوان یک کنترل داخلی برای واکنش توالی‌یابی انتخاب می‌شود. یک تکنولوژی مبتنی بر آنزیم جایگزین، طویل شدن پرایمر با تک-باز است که نتایج کمی دقیق را با هزینه میانگین فراهم می‌کند. توالی‌های الیگونوکلئوتید کوتاه منحصراً در کنار SNP هیبرید می‌شوند. این الیگونوکلئوتیدها سپس به عنوان پرایمر برای پلیمرازها استفاده می‌شوند که نوکلئوتید نشاندار شده را در موقعیت SNP قرار می‌دهند. سپس نوکلئوتید وارد شده با استفاده از روش‌های رنگی یا طیف‌سنجی جر م[3]، ردیابی می‌شود. علاوه بر این، SNPها همچنین می‌توانند به صورت in silico تعیین شوند، یعنی، با مقایسه گرافیکی توالی EST از افراد مختلف. تبادل نوکلئوتیدی با استفاده از چنین همردیفی چندگانه، بسیار آسان است. با این حال، هنگام توصیف SNPهای جدید با استفاده از ESTها، توصیه می‌شود نهایت احتیاط صورت پذیرد زیرا این تفسیر می‌تواند حاوی اشتباهات توالی‌یابی به صورت SNPها باشد.

---

[1] Microarray
[2] Pyrosequencing
[3] Mass Spectroscopy

پایگاه داده dbSNP مخزن سایت NCBI جهت بررسی چندشکلی است. هر ورودی شامل جزئیات تنوع ژنتیکی، نوکلئوتید مجاور و فراوانی چندشکلی است. همچنین شامل اطلاعاتی مربوط به روش تجربی و شرایط مورد استفاده برای شناساییSNP می‌باشد. dbSNP حاوی حدود ۷۸۰ میلیون چندشکلی از ۵۳ موجود زنده مختلف است که ۵۴۵ میلیون از آن انسانی هستند (سپتامبر ۲۰۱۶). علاوه بر این، یک مجموعه از SNPهای انسان را می‌توان در پایگاه GWAS مرکزی پیدا کرد، که قبلاً به عنوان پایگاه داده تنوع ژنوم انسانی شناخته شده است. این ورودی‌های SNP تحت بررسی کیفی بیشتر قرار گرفته‌اند و به طور کامل تفسیر شده‌اند.

۱-٦-٤ فارماکوژنتیک

فارماکوژنتیک (یا فارماکوژنومیک)، تغییرات ژنتیکی است که در اثر واکنش متفاوت بیماران در مواجهه با داروها رخ می‌دهد. یک مطالعه در ایالات متحده آمریکا در سال ۱۹۹۴ نشان داد که ۲/۲ میلیون بیمار از عوارض جانبی جدی دارو رنج می‌برند و بیش از ۱۰۰ هزار بیمارجان خود را از دست دادند. بنابراین، احتمال بیشتر مرگ از عوارض جانبی دارو نسبت به عفونت‌های ویروسی وجود دارد. بر این اساس، توانایی پیش‌بینی اینکه چگونه یک بیمار ممکن است قبل از شروع درمان به دارو واکنش نشان دهد، پیشرفت قابل توجهی خواهد بود.

اینکه چگونه یک بیمار به داروها پاسخ می‌دهد یک فرآیند پیچیده است که شامل پروتئین‌های مختلفی از جمله گیرنده‌ها و آنزیم‌هایی است که به ترتیب به داروها متصل شده و آن را متابولیزه می‌کنند. تنوع ژنتیکی در چنین پروتئین‌هایی می‌تواند باعث کاهش اتصال یا عدم وجود متابولیسم دارو شود. اهمیت چندشکلی، در پروتئین‌های خانواده سیتوکروم P450 بیشتر مشهود است. به عنوان مثال، آنزیم CYP2D6 مسئول متابولیسم ۲۰ تا ۲۵ درصد از تمام داروهای تجویزی است. جهش در ژن CYP2D6 می‌تواند بر میزان متابولیسم داروها تأثیر بگذارد. بسته به نوع جهش، می‌توان واکنش بیماران را با متابولیسم دارو بصورت زیاد، گسترده، متوسط و کم تشخیص داد. بدیهی است، چندشکلی ژنتیکی به طور واضحی واکنش فردی بیمار به داروها را تحت تاثیر قرار می‌دهد. از آنجایی که SNPها اغلب از شایع‌ترین تغییرات ژنتیکی هستند، جستجوی SNPها که متابولیسم یا اثر دارو را تحت تاثیر قرار می‌دهند، اهمیت زیادی برای فارماکوژنتیک دارد.

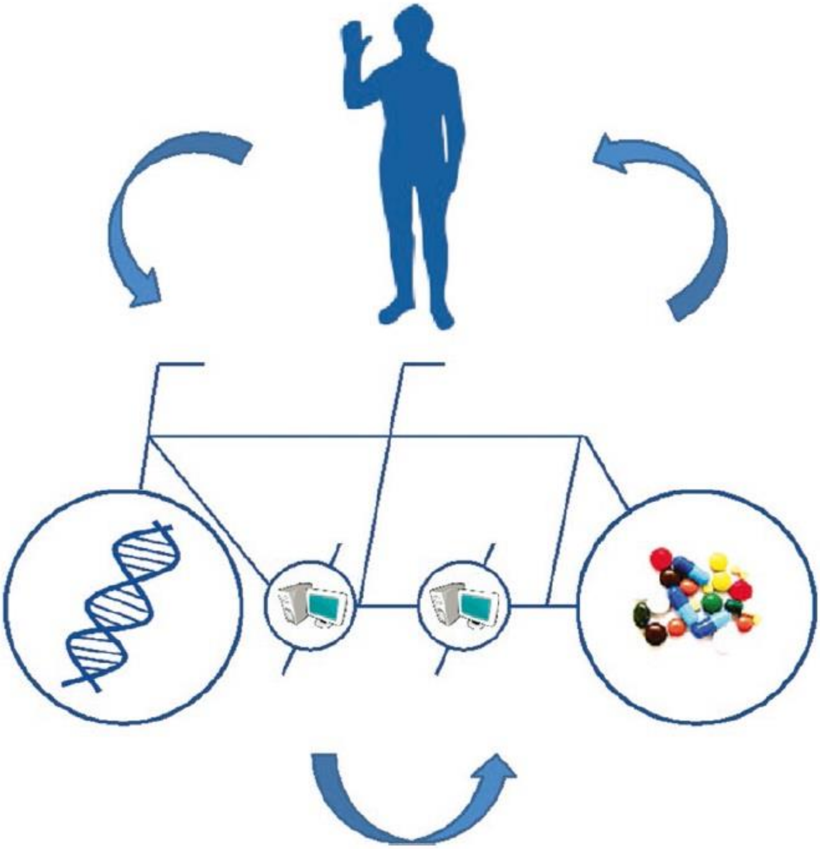
همانطور که گفته شد هدف اصلی فارماکوژنتیک‌ها، پیش‌بینی اثرات جانبی ناخواسته دارو پیش از درمان است. یک پیش نیاز مهم این است که توسعه تست‌های تشخیصی برای وابستگی ژنتیکی بیماران و نحوه واکنش آن‌ها به یک دارو خاص درک شود. در چنین آزمایش‌های تشخیصی، ژنوتیپ هر بیمار تثبیت می‌شود، یعنی، آیا پروتئین‌های مربوطه مانند آنزیم‌های متابولیزه‌کننده دارو، چندشکلی‌های متمایز را نشان می‌دهند. بیماران پس از تست‌های تشخیصی می‌توانند به گروه‌های مربوطه دسته‌بندی شوند و یک درمان مناسب بر اساس ژنوتیپ آن‌ها انتخاب شود (شکل ۴-۸ و ۴-۹). این عمل به عنوان روش داروی طبقه‌بندی شده اشاره می‌شود، زیرا درمان بهینه و متناسب با هر بیمار متعلق به گروه دارای واکنش خاص مشخص می‌گردد. یک مثال که قبلاً در بسیاری از کشورها تمرین شده است، شیمی‌درمانی بیماران مبتلا به سرطان خون حاد لمفاتیک (ALL)[1] است. مرکاپتوپورین و تیوگوانین[2] اغلب به عنوان داروهایی استفاده می‌شوند که به DNA سلول‌های در حال تکثیر (بویژه سلول‌های سرطانی) متصل می‌شوند و منجر به مرگ آن‌ها می‌شود. یک آنزیم مسئول متابولیسم این ترکیبات، تیوپورین-S- متیل ترانسفراز[3] است. مطالعات بالینی نشان داده است که چندشکلی ژنتیکی به شدت بر فعالیت تیوپورین-S- متیل ترانسفراز تأثیرمی‌گذارد و بنابراین سمیت و کارآیی مرکاپتوپورین و تیوگوانین را تأیید می‌کند. بیماران دارای کمبود در تیوپورین-S- متیل ترانسفراز این داروها را در سلول‌های خونی در غلظت‌های بالا جمع می‌کنند و در نهایت باعث مرگ می‌شوند. در مقابل، در بیماران مبتلا به فعالیت تیوپورین-S - متیل ترانسفراز بالا، مرکاپتوپورین و تیوگوانین باید در دوزهای بالاتر استفاده شود. بنابراین، هر بیمار برای چندشکلی در ژن کدکننده تیوپورین-S - متیل ترانسفراز مورد بررسی قرار گرفته و موثرترین دوز قبل از درمان با مرکاپتوپورین و تیوگوانین مشخص می‌شود.

---

[1] Acute Lymphatic Leukemia (ALL)

[2] Mercaptopurine and Thioguanine

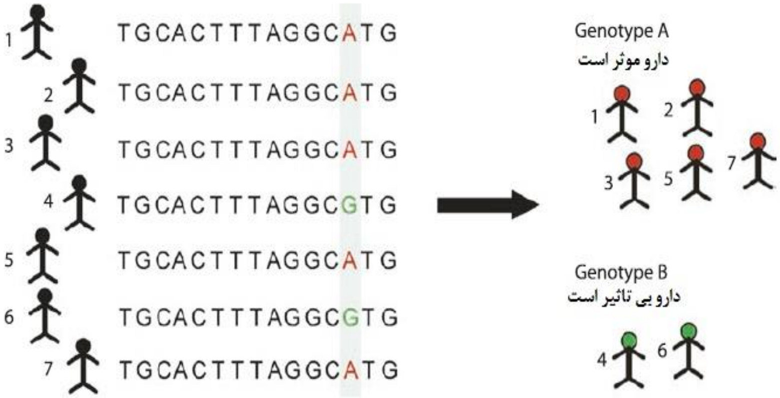
[3] Thiopurine-S-Methyltransferase



شکل ۴-۸) فارماکوژنتیکس: تشخیص و درمان درکنار یکدیگر قرار می‌گیرند. استعداد ژنتیکی بیمار اثر یک دارو را تحت تاثیر قرار می‌دهد. تجزیه و تحلیل استعداد ژنتیکی بیمار می‌تواند به انتخاب یک داروی مناسب کمک کند.

علاوه بر بیماران در کلینیک، در تحقیقات فارماکولوژی نیز از فارماکوژنتیک استفاده شده است. قبل از تأیید برای استفاده در بیماران، هر نامزد دارویی جدید باید در مطالعات بالینی گسترده با استفاده از معیارهای ایمنی و کارآمد مورد آزمایش قرار گیرند. فارماکوژنتیک امکان جدا کردن بیمارانرا در واکنش به نوع درمان نشان می‌دهد یا شخصی که ممکن است عوارض جانبی ناخواسته را قبل از شروع هر مطالعه تجربه کند، تعیین می‌کند. این فرآیند، شانس رسیدن دارو به بازار را برای بیمارانی که از مصرف داروهای فاقد عوارض جانبی خطرناک سود

می‌برد، افزایش می‌دهد. لیستی از داروهایی که منع مصرف دارند بجز برای بیمارانی که مورد آزمایش قرار گرفته‌اند، در وب سایت Verband Forschender Pharmaunternehmen یافت می‌شود. علاوه بر این، فارماکوژنتیک سبب توسعه داروهای جدید برای گروه‌های مختلف بیمارانی می‌شود که به درمان‌های موجود پاسخ نداده‌اند و می‌تواند به طبقه‌بندی درمان کمک می‌کند. بیماران با متابولیسم بسیار سریع که یک دارو را می‌توانند متابولیزه کنند، یک داروی جایگزین یا دوز بالاتری از همان دارو را دریافت می‌کنند. بر این اساس، بیماران با متابولیسم آهسته، که در آن به سطح پلاسمای خطرناک دارو می‌رسد، می‌توانند با استفاده از یک داروی جایگزین یا دوز پایین‌تری از همان دارو درمان شوند. چنین ناسازگاری‌های دارویی می‌تواند در مقایسه با پیش‌دارو، به عنوان مثال داروهایی که به وسیله متابولیزه شدن (مانند: تاموکسیفن[1]) یا داروهایی که پیش‌دارو نیستند (داروهای ضدافسردگی مانند: میرتازاپین[2])، به داروهای فعال تبدیل می‌شوند. بعضی از شرکت‌ها مانند Humatrix AG، برای ارزیابی متناسب بودن داروها برای بیماران در آزمایش‌های تشخیصی تخصص یافته‌اند.



شکل ۴-۹) ژنوتایپینگ بیماران با استفاده از شناسایی SNPها

همچنین باید در نظر داشت که واکنش‌های فردی به داروها تنها می‌تواند تا حدی با تغییرات ژنتیکی و عوامل دیگر،که بر اثر و ایمنی دارو اثر می‌گذارند، توضیح داده شوند(اورت، ۲۰۱۶). این موضوع شامل عوامل متعددی مانند: سن بیمار، وضعیت تغذیه، مصرف الکل،

---

[1] Tamoxifen
[2] Mirtazapine

وضعیت میکروبیوم بیمار، اینکه آیا شرایط همسو با سایر بیماری‌ها یا همسو با سایر داروها مصرف می‌شوند یا خیر، می‌باشد. علاوه بر این، وجود تنوع ژنتیکی نیز می‌تواند، منجر به تنوع متابولیکی شود.

اگر ما می‌خواهیم میزان موفقیت داروهای فردی را افزایش دهیم، بنابراین ما باید نه تنها موقعیت ژنتیکی بیمار، بلکه پروفایل متابولیسمی فرد را نیز در نظر داشته باشیم. پروفایل متابولیک[۱] با استفاده از روش‌های فیزیکوشیمیایی در حال حاضر حدود ۱۰ سال است که ادامه دارد. در سال‌های اخیر، اصطلاحات متابونومیکس[۲] و متابولومیکس[۳] معرفی شده است. با این حال، مطالعات اولیه نشان داد که پروفایل‌های متابولیک دو بیمار پس از دارو با همان دارو به طور چشمگیری وابسته به پروفایل متابولیک بیماران قبل از دارو است. میکروبیوم روده بیمار بر متابولیسم دارو اثر می‌گذارد. این دانش منجربه یک رشته جدید، به نام فارماکومتابونومیکس[۴] شد. این نظریه پیش‌بینی اثر دارو براساس مشخصات متابولیک بیمار را پیش از اعمال دارو با استفاده از مدل‌های ریاضی پیش‌بینی می‌کند. استفاده از فارماکوژنومیکس و فارماکومتابونومیکس باید به کیفیت بالاتری از داروهای فردی منجر شود (اورت، ۲۰۱۶).

۲-٦-٤ پزشکی شخصی و نشانگرهای زیستی

تنظیم درمان بیمار در معرض ابتلا به بیماری ژنتیکی بیمار و مشخصات متابولیسم فرد اغلب به عنوان پزشکی شخصی[5] شناخته می‌شود. این اصطلاح را می‌توان اغلب در ادبیات علمی یافت که حدود سال ۲۰۰۰ شروع شد و پس از آن زمان اهمیتش را به دست آورد، اگرچه تعریف واضحی از این واژه مطرح نشد، که هنوز جای تفسیر دارد. شلیدن و همکاران، ۲۰۱۳، با مقایسه ۶۵۳ نشریه علمی با استفاده از رویکردهای واژگانی تعارف، درک مشترکی را به دست آوردند. بر اساس تحقیقات خودشان، پزشکی شخصی، تلاش می‌کند طبقه‌بندی را بهینه کند، یعنی ارزیابی ریسک فعلی و درمان براساس دانش اطلاعات بیولوژیک و دانش نشانگرهای زیستی در سطح مسیرهای متابولیک مولکولی، ژنتیک، پروتئومیکس و متابولومیکس. در واقع،

---

[1] Metabolic Profiling
[2] Metabonomics
[3] Metabolomics
[4] Pharmacometabonomics
[5] Personalized Medicine

این تعریف به نوعی دست و پا گیر است. در نهایت، این بدان معنی است که خصوصیات بیولوژیکی فردی برای درمان هر بیمار در نظر گرفته می‌شود. به این ترتیب، توجه ویژه به نشانگرهای زیستی[۱] صورت می‌گیرد، که در میان سایر روش‌ها بر اساس ژنتیکی تعیین می‌شود.

مارکرهای زیستی به سادگی پارامترهایی هستند که می‌توانند برای تشخیص، پیش‌آگهی و درمان استفاده شوند. نمونه‌های معمول شناخته شده، پارامترهای خون نگاره (مربوط به خون)[۲] هستند که توسط پزشکان برای تشخیص و نظارت بر موفقیت درمان و یا تغییر درمان استفاده می‌شود. با توجه به پیچیدگی‌های بسیار زیادی از بیماری‌ها مانند سرطان، چنین نشانگرهای زیستی، دیگر کافی نیست. امروزه نشانگرهای زیستی مورد نیاز است که منجر به یک تصویر دقیق‌تر می‌شود. در اینجا ما ارتباطی با فارماکوژنتیک پیدا می‌کنیم، جایی که دانش استعداد ژنتیکی بیمار، مانند یک چندشکلی در تیوپورین-S- متیل ترانسفراز، به عنوان یک نشانگر زیستی جهت بهینه‌سازی درمان استفاده می‌شود. در روش نسبتاً مشابه، تحقیقات نشانگرهای زیستی به دنبال مناطقی در DNA ژنومی، mRNA یا پروتئین‌هاست که با بیماری مرتبط هستند. هنگامی که چنین نشانگرهای زیستی تثبیت می‌شوند، می‌توانند برای تشخیص، پیش‌آگهی و درمان استفاده می‌شوند. دانش ژنوم بیمار برای نشانگرهای زیستی ژنتیکی، بسیار مهم است. تا همین اواخر، توالی‌یابی ژنوم‌های فردی غیرممکن بوده است.

فقط با ظهور توالی‌یابی نسل بعدی (NGS)[۳]استفاده از توالی ژنوم‌ها یک روش تشخیص مناسب می‌باشد، زیرا که می‌توانست در فقط طی چند روز تنها با پرداخت هزینه چند هزار دلاری به جای ۱۰ سال و در یک هزینه تقریبی ۳ میلیارد دلاری توالی‌یابی کند.

چگونه یک مارکر زیستی را شناسایی می‌کنیم؟ یک روش، مطالعه مرتبط با کل ژنوم (GWAS) است. هدف از GWAS شناسایی آلل‌هایی است که با بیماری‌های مشخصی ارتباط دارند، یعنی آلل‌هایی که در صورت وجود برخی بیماری‌ها را پدید می‌آورند و در صورت عدم وجود، این بیماری یافت نمی‌شوند. اگر چنین همبستگی پیدا شود، در ابتدا تنها یک ارتباط برقرار می‌شود. مطالعات بیوشیمیایی و بیولوژی مولکولی برای تعیین اینکه آیا این یک علت ایجادکننده است یا خیر، انجام می‌شود. برای یک GWAS، دو گروه مطالعه تشکیل می‌شوند، یکی با افرادی که ویژگی خاصی دارند، مثلاً یک بیماری، و دیگری از افرادی است که این

---

[1] Biomarkers

[2] Hemograms

[3] Next-Generation Sequencing (NGS)

ویژگی را ندارند. سپس نمونه‌های DNA هر دو گروه برای تغییرات ژنتیکی مورد تجزیه و تحلیل قرار می‌گیرند. یا کل ژنوم یا فقط مناطق خاص نشانگر، یعنی SNPهای تعریف شده، تجزیه و تحلیلمی‌شوند. پیشرفت تکنیکی اخیر در توالی‌یابی DNA (بخش ۴-۶-۳) و هزینه‌های کاهش‌یافته، در حال حاضر باعث توالی‌یابی بیشتر ژنوم بیمار می‌شود. از طرفی می‌توان از GWASها برای اهداف تشخیصی استفاده نمود، به عنوان مثال، در فارماکوژنتیک (بخش ۴-۶-۱)، همچنین از این روش برای اهداف پیش‌بینی شده، مانند: جستجو برای ارتباطات ویژگی آلل در ژنوم بیمار، حتی اگر بیماری هنوز ظاهر نشده است، نیز بهره جست. با وجود شناسایی چنین ارتباطی، به این معنا نیست که بیماری باید سرانجام ظاهر شود؛ فقط به این معنی است که احتمال خاصی وجود دارد که ظهور خواهد کرد. به عنوان مثال، هموکروماتوسیس[1] را در نظر بگیرید که به جهش هموزیگوت در ژن HFE متصل است. احتمال این که بیماری در واقع ظاهر شود تنها ۳۰-۵۰ درصد است، یعنی از ۱۰۰ بیمار که نشان‌دهنده جهش در ژن HFE است، فقط ۳۰-۵۰ بیمار علائم بالینی بیماری را نشان می‌دهند. به این ترتیب روشن می‌شود که تعداد روزافزون ژنوم شناخته شده انسان نه تنها برای بیماران و جامعه سودمند است، بلکه سؤالات اجتماعی و اخلاقی را نیز مطرح می‌کند. مثلاً چگونه بیمار می‌تواند با دانستن یک عامل خطر مقابله کند، یا شرکت‌های بیمه با این اطلاعات چه باید بکنند؟ این کتاب درسی برای پاسخ به چنین سؤالی مناسب نیست. با این حال، مهم است که به خاطر داشته باشید که علیرغم همه‌ی رضایت در مورد امکانات فنی، گفتگوی عمومی ضروری است.

۳-٦-٤ توالی‌یابی نسل بعدی (NGS)

همانطور که قبلاً ذکر شد، NGS باعث توالی‌یابی سریع کل ژنوم می‌شود. علاوه بر این، همچنین ممکن است RNA را توالی‌یابی کنیم (RNA-Seq، بخش ۴-۲-۲)، انواع پیرایش و مکان‌های قابل پیرایش را شناسایی کنید و مقدار mRNA را دقیقاً تعیین کنید. همچنین به افراد اجازه می‌دهد تنوع میکروبی را در انسان یا محیط مطالعه کنند. NGS، به همین دلیل، به یک ابزار بسیار مهم در تحقیقات روزمره تبدیل شده است. روش‌های متعددی در دسترس هستند که از اصل مشابهی پیروی می‌کنند. در مرحله اول یک کتابخانه DNA ایجاد می‌شود.

---

[1] Hemochromatosis

این کتابخانه شامل قطعات کوتاه DNA است که طول آن‌ها با طول DNA کوتاه از توالی شناخته شده در هر دو انتهای'۵ و '۳ طویل شده‌اند (تکثیر شده). این آداپتورها[۱] در مرحله بعدی برای ترمیم قطعات DNA به محیط واکنش جامد و تکثیر آن‌ها استفاده می‌شود. در این مرحله، از چندین روش استفاده می‌شود که در نهایت به صورت خوشه‌ای از قطعات DNA مشابه می‌باشد. سپس توالی دقیق در خوشه‌های منفرد تشکیل می‌شود. آخرین مرحله، ارائه‌ی داده‌ها است و تمام روش‌ها داده‌ها را به شکل تراشه‌های DNA ارائه می‌دهند. تفاوت اصلی در سیستم‌های مختلف در جزئیات فنی توالی‌یابی قرار دارد. در اصل، چهار نوع سیستم توالی‌یابی را می‌توان تعریف کرد:

- پایروسیکوئنسینگ: در طی واکنش توالی، یک پیروفسفات آزاد می‌شود، که از طریق توالی از واکنش‌های شیمیایی منجر به انتشار نور می‌شود. این نور توسط یک دوربین شناسایی می‌شود. سپس بازها اضافه می‌شوند و دوربین تشخیص می‌دهد که آیا نور منتشر می‌شود یا خیر. قبل از اضافه کردن باز بعدی، مرحله شستشو انجام می‌شود.
- توالی‌یابی با سنتز[۲]: این روش شامل استفاده از نوکلئوتیدهایی است که به یک ترمیناتور[۳] و یک رنگ فلورسنت متصل می‌شوند. بعد از افزودن نوکلئوتید، رنگ فلورسنت منتشر شده و طول موج فلورسنت منتشر شده، ثبت می‌شود. در نهایت، ترمیناتور حذف می‌شود و نوکلئوتید بعدی اضافه می‌شود.
- •توالی‌یابی با اتصال[۴] : به جای یک DNAپلیمراز، از ۱۶ پروب مختلف الیگونوکلئوتیدی استفاده می‌شود. هر یک از پروب‌های نوکلئوتیدی یکی از چهار رنگ مختلف فلورسنت را در انتهای '۵ خود حمل می‌کند. هر اکتامر شامل دو باز اختصاصی و شش باز عمومی است. برای توالی‌یابی، یک پرایمر خاص به آداپتور توالی DNA متصل شده و پروب اتصالی با استفاده از DNAلیگاز متصل می‌شود. پس از یک مرحله شستشو، سیگنال فلورسنت ثبت می‌شود و سه باز آخر DNA و رنگ‌های فلورسنت حذف می‌شوند. اینکار هفت بار انجام می‌شود و به دنبال آن یک گام

---

[1] Adapters

[2] Sequencing by Synthesis

[3] Terminator

[4] Sequencing by Ligation

غیرطبیعی شدن رخ می‌دهد. این فرآیند با استفاده از آغازگر شروع می‌شود که با یک نوکلئوتید جایگزین می‌شود. پنج آغازگر مختلف در مجموع استفاده می‌شود.

- توالی‌یابی یون نیمه‌هادی[1]: این روش مشابه پایروسیکوئنسینگ است. به جای ثبت رهاسازی یک پیروفسفات، رهاسازی پروتون ردیابی می‌شود. خوشه‌های DNA به یک تراشه نیمه‌هادی متصل هستند که قادر به اندازه‌گیری pH اطراف آن هستند. هنگامی‌که یک نوکلئوتید اضافه می‌شود، یک پروتون آزاد می‌شود که منجر به تغییر مقدار pH می‌گردد و می‌تواند توسط تراشه نیمه‌هادی شناسایی شود.

هریک از روش‌ها دارای مزایا و معایبی می‌باشند که شامل، طول خواندن مختلف، هزینه‌های واکنش‌گر، نرخ خطا، زمان اخذ و پوشش است. پوشش به معنای تعداد خواندن در یک سرهم‌بندی توالی است که برای تولید یک دنباله مرجع لازم است. برای یک ژنوم کامل، حداقل پوشش، ۳۰ است. امروزه تنها دو روش می‌توانند به این سطح پوشش برای ژنوم انسان دست یابند: توالی‌یابی با سنتز و توالی‌یابی با اتصال. روش پایروسیکوئنسینگ برای ژنوم باکتریایی و برای یوکاریوتی‌های ساده، مانند *Arabidopsis thaliana* مناسب است.

مقدار داده‌های تولید شده توسط روش توالی‌یابی NGS چالش بزرگی است. حتی فایل فشرده شده FASTQ (یک فرمت فایل تخصصی است که در هر توالی شامل شناسه توالی، خود توالی در فایل FASTA، و دو خط اضافی، یکی برای نظرات و یکی برای نمرات کیفیت) به راحتی به اندازه ۲۰۰ گیگابایت برای یک ژنوم انسان به سطح پوشش ۶۰ می‌رسد. بنابراین پروژه با ۱۰-۲۰ ژنوم از حدود ۴ ترابایت فضای دیسک استفاده می‌کند. بنابراین، نه تنها ذخیره‌سازی جزئی نیست، بلکه انتقال این مقدار داده‌ها بین گروه‌های تحقیقاتی مختلف یک چالش است. یک راه حل مناسب استفاده از سیستم ابر[2] است. مؤسسه ملی بهداشت (NIH) دارای دو سیستم ابر است که به نام Biowolf و Helix نام‌گذاری شده است. در اروپا، EMBL بر روی یک سیستم ابر به نام لکه ابر مارپیچ[3] کار می‌کند.

یکی دیگر از چالش‌ها، همردیفی دوباره یا نقشه‌برداری خوانش کوتاه نسبت به ژنوم مرجع است. به دلیل طول کوتاه خوانش، آن‌ها با چند موقعیت ژنوم مرجع متناسب می‌شوند. علاوه بر این، ژنوم مرجع بزرگ است و بنابراین پیدا کردن موقعیت صحیح را مشکل می‌سازد. به دلیل

---

[1] Ion Semiconductor Sequencing

[2] Cloud System

[3] Helix NebulaCloud

خطاهای توالی‌یابی و ماهیت SNPها، تغییرات خاصی از فرآیند نقشه‌برداری ضروری است. بعداً می‌توان خطاها را از واریانت‌های واقعی تشخیص داد. الگوریتم‌های متعددی برای نقشه‌برداری خواندن در دسترس هستند، مانند BWA، Bowtie، SNP-o-matic، NextGenMap و BLAT، همچنین یک لیست کامل از انواع الگوریتم‌ها را می‌توان در وب سایت HTS-Mapper پیدا نمود. خروجی بسیاری از برنامه‌های mapper در قالب SAM/BAM است، جایی که فرمت BAM یک نسخه باینری (دوتایی) فشرده شده از فرمت SAM قابل خواندن برای انسان را فراهم می‌آورد. فایل‌های BAM را می‌توان نمایه‌سازی نمود، تا اجازه دسترسی سریع به هر منطقه از یک توالی را فراهم کند. ابزارهای ویژه، مانند ابزار SAM، برای تجزیه و تحلیل، اصلاح و تجسم توالی‌ها امکان‌پذیر است.

هنگامی‌که نقشه‌برداری کامل می‌شود، اطلاعات ژنوم را می‌توان تجزیه و تحلیل کرد، مانند: انواع نوکلئوتید تکی مثل SNPها که شناسایی می‌شوند. همچنین برای این مرحله، چندین ابزار در دسترس هستند، مثلاً MAQ، SAMtools، VariationHunter و destruct. یک مرور کلی از ابزارها و روش‌ها را می‌توان در پایگاه Wikibook توالی‌یابی نسل بعدی (NGS) پیدا کرد، که به طور مداوم به روز می‌شود.

۴-۶-۴ پروتئوژنومیکس

با ظهور NGS، به سرعت مشخص شد که تعدادی از انواع پیرایش‌ها و چندشکلی‌های نوکلئوتیدی باید منجر به تعداد زیادی تغییرات در پروتئوم شود که در پایگاه داده استاندارد ذخیره شوند. هدف پروتئوژنومیکس[1] بررسی ارتباط واقعی بین ژنوم و پروتئوم است. برای این منظور، پروتئین‌ها با تولید یک اثر انگشت پروتئینی با استفاده از طیف‌سنجی جرمی (MS) توالی‌یابی می‌شوند. اگر یک اثر انگشت حاصل از آزمایش مشابه، قابل رویت (نظری) باشد، پروتئین‌ها یکسان هستند و توالی پروتئین ناشناخته مشخص می‌شود. پایگاه داده‌های اثر انگشت نظری پروتئین بر اساس داده‌های NGS که ژنوم را به پروتئوم متصل می‌کنند، ساخته شده است. این روش، در سال ۲۰۰۴ ساخته شده است. در دهه ۱۹۹۰ و دهه ۲۰۰۰ پروتئومیکس شاتگان استفاده می‌شد، که در آن داده‌های MS برای جستجو پایگاه‌های پروتکل استفاده شده است. در سال ۲۰۰۴، جاف و همکاران (جاف و همکاران، ۲۰۰۴) از ترجمه شش

---

[1] Proteogenomics

قالب از ژنوم مایکوپلاسما به عنوان پایگاه داده پروتئین استفاده کردند و واژه پروتوئژنومیکس را تعریف کردند. این مفهوم به سرعت برای موجودات پیچیده‌تر مورد استفاده قرار گرفته است و در حال حاضر در ترکیب با NGS اهمیت حیاتی برای شناسایی و مطالعه انواع پروتئین‌های انسانی در تحقیقات بیولوژیکی و پزشکی دارد (شینکمن و همکاران، ۲۰۱۶).

انواع مختلفی از داده‌های نوکلئوتیدی برایبکار بردن این روش مناسب می‌باشند. داده‌های اولیه EST به سه یا شش قالب خوانش ترجمه می‌شوند. اگر از داده‌های ژنومی استفاده شود، آن‌ها به شش قالب خوانده می‌شوند. همچنین، داده‌های RNA-Seq نیز مورد استفاده قرار می‌گیرند، علاوه بر این داده‌های توالی‌یابی ریبوزومی، که مولکول‌های mRNA به ریبوزوم متصل می‌شوند، نیز استفاده می‌شوند. پایگاه‌های داده اختصاصی که بر ویژگی‌های خاص متمرکز شده‌اند، مانند: انواع پیرایش یا SNPها نیز مورد استفاده قرار می‌گیرند (شینکمن و همکاران، ۲۰۱۶، نسویزسکی، ۲۰۱۴).

اگرچه تعدادی از انواع پروتئین‌ها تاکنون کشف شده‌اند، ولی هر دو روش براساس قطعات خاص هستند، به عنوان مثال، در مورد NGS، قطعات DNA یا RNA، و در مورد پروتئوژنومیکس، پروتئین‌های هضم‌شده توسط آنزیم. بنابراین توالی کامل و بدون نقص نمی‌تواند با اطمینان ۱۰۰ درصد نشان داده شود. بنابراین، ممکن است که انواع بیشتری کشف نشده باقی بماند. با این حال، هر دو روش در طول زمان بهبود می‌یابند، و باعث کشف توالی‌های دست نخورده می‌شود.

سایت‌های مفید

cap. http://doua.prabi.fr/software/cap3
dbest. https://www.ncbi.nlm.nih.gov/dbEST/
dbgss. https://www.ncbi.nlm.nih.gov/dbGSS/
dbsnp. https://www.ncbi.nlm.nih.gov/SNP/
dbsts. https://www.ncbi.nlm.nih.gov/dbSTS/
ebi-gwas. http://www.ebi.ac.uk/gwas/
grailexp. http://compbio.ornl.gov/grailexp/
gwas. http://www.gwascentral.org/
helix-nebula. http://www.helix-nebula.eu/usecases/embl-use-case
homologene. http://www.ncbi.nlm.nih.gov/homologene/
hts-mapper. http://www.ebi.ac.uk/~nf/hts_mappers/
humatrix. https://www.humatrix.de/
image. http://imageconsortium.org/
nematode. http://www.nematode.net/

ngs-knowledge-base. https://goo.gl/HIaY1W
ngs-movie. https://www.youtube.com/watch?v=jFCD8Q6qSTM
nhgri-gwas. https://www.genome.gov/20019523/
nih-biowolf. https://hpc.nih.gov/
pahdb. http://www.pahdb.mcgill.ca/
phrap. http://www.phrap.org/
pyrosequencing. http://www.pyrosequencing.com/
sam-tools. https://en.wikipedia.org/wiki/SAMtools
stackpack. http://genoma.unsam.edu.ar/stackpack.old/index.html
stratipharm. http://www.stratipharm.de/
unigene. http://www.ncbi.nlm.nih.gov/UniGene/
vfa-personalisiert. http://www.vfa.de/personalisiert/
wikibookhttps://en.wikibooks.org/wiki/Next_Generation_Sequencing_%28NGS%29

منابع

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656.
2. Berger EA, Murphy PM, Farber JM (1999) Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol 17:657–700.
3. Blaxter M (1998) Caenorhabditis elegans is a nematode. Science 282:2041–2046.
4. Everett JR (2016) From metabonomics to pharmacometabonomics: the role of metabolic profiling in personalized medicine. Front Pharmacol 7:297 und darin enthaltene Referenzen.
5. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866–5878.
6. Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59–77.
7. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
8. Nesvizhskii AI (2014) Proteogenomics: concepts, applications, and computational strategies. Nat Methods 11:1114–1125.

9. Schleidgen S, Klingler C, Betram T, Rogowski WH, Marckman G (2013) What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics 14:55.
10. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM (2016) Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize Human Proteomic variation. Annu Rev Anal Chem (Palo Alto, Calif) 9:521–545.
11. Thorisson GA, Stein LD (2003) The SNP Consortium website: past, present, future. Nucleic Acids Res 31:124–127.
12. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.
13. Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. Genome Biol 5(10):R74.

فصل پنجم

ساختارهای پروتئینی و طراحی دارو مبتنی بر ساختار

۵-۱ ساختار پروتئین

پروتئین‌ها ماکرومولکول‌هایی هستند که زیرواحدهای مونومری آن‌ها از ۲۰ نوع اسیدآمینه طبیعی تشکیل شده‌اند. اسیدهای‌آمینه از طریق پیوند پپتیدی (تولید شده بعد از آزاد شدن آب) برای تشکیل یک پلی‌پپتید (فصل ۱) به هم متصل می‌شوند. پلی‌پپتیدها می‌توانند از سه تا چند صد اسیدآمینه تشکیل شوند. توالی اسیدآمینه یک پروتئین مشخص، همچنین که به عنوان ساختار اولیه شناخته می‌شود، از نظر ژنتیکی تعیین می‌شود. در هنگام ترجمه براساس اطلاعاتی که در mRNA کدگذاری شده است، ثابت می‌شود.

خواص یک زنجیره پلی‌پپتیدی گسترده متناظر با یک بخش مقطعی از آن دسته از اسیدهای‌آمینه مربوطه است، یعنی، عملکرد پروتئین را نمی‌توان تنها از ساختار اولیه تعیین کرد. این خواص به ترتیب مکانی اسیدآمینه‌ها نسبت به یکدیگر بستگی دارد. زنجیرهای پلی‌پپتیدی به صورت خود به خودی به ساختارهای ثانویه و سپس به ساختارهای سه‌بعدی (3D) می‌پیچند. ساختار ثانویه شامل دو ویژگی اصلی ساختاری، مارپیچα[۱] و صفحه β[۲] است. این عناصر ساختاری از طریق عناصر غیرتکراری به نام حلقه‌ها[۳]، به هم متصل می‌شوند که از گردش‌های نامنظم عناصر ساختار ثانویه سوم تشکیل شده است. این حالت ترکیبی از موقعیت زنجیره‌های جانبی اسیدآمینه و ستون پپتید از ساختار ثانویه است که ساختار سوم را تشکیل می‌دهد. اگر یک پروتئین از چندین زیرواحد تشکیل شده باشد، ارتباط این زیر واحدها با تشکیل پروتئین عملکردی، ساختار چهارم نامیده می‌شود.

عملکرد پروتئین به وسیله ساختار 3D آن وساطت می‌شود، که اگر شناخته شود، باعث استنباط عملکرد می‌شود. پیش‌بینی قابل اعتماد از ساختار سوم پروتئین، فقط بر اساس ساختار اولیه، حداقل در آینده نزدیک هنوز امکان‌پذیر نیست. همچنین تعیین تجربی ساختار هنوز دشوار است و تعداد ساختارهای شناخته شده پروتئین نسبتاً کوچک است. بنابراین پیش‌بینی عملکرد پروتئین به ساختار سوم یا چهارم محدود است. با این حال، پروتئین‌ها انواع مختلفی از ویژگی‌های ساختاری و توپولوژیکی را نشان می‌دهند که می‌تواند برای پیش‌بینی خواص و عملکرد آن‌ها استفاده شود. بسیاری از این ویژگی‌ها می‌توانند از ساختار اولیه توسط روش‌های

---

[1] α-Helix
[2] β-Sheet
[3] Loops

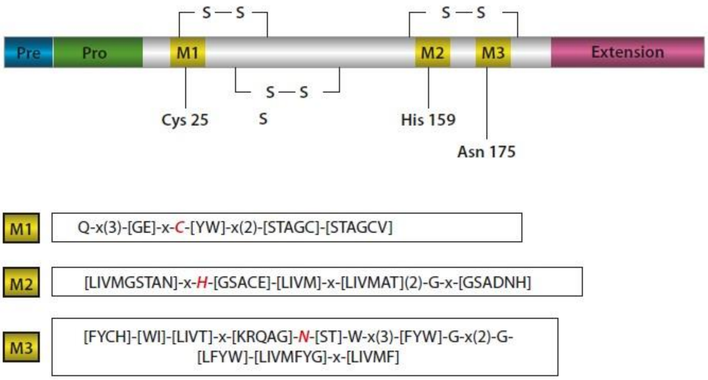
محاسباتی استنباط یا پیش‌بینی شوند. برخی از این خواص و پیش‌بینی‌های در موارد زیر بحث می‌شوند.

۵-۲ سیگنال پپتیدها

برای بسیاری از پروتئین‌ها، محل سنتز پروتئین همان محل عملکرد پروتئین نمی‌باشد. این امر در مورد پروتئین‌های بین غشایی، پروتئین درون شبکه آندوپلاسمی و پروتئین‌هایی که ترشح شده یا وارد لیزوزوم می‌شوند، اعمال می‌شود. قبل از فعال شدن، این پروتئین‌ها ابتدا باید به محل عمل منتقل شوند و این امر با یک پپتید تشخیص سیگنال برای سیستم حمل و نقل سلولی تسهیل می‌شود. سیگنال تشخیص یک توالی رهبر پایانه N (سیگنال پپتید[1]) است که تقریباً حدود ۱۵-۳۰ اسیدآمینه در پایانه‌ی N پروتئین بالغ قرار دارد (شکل ۵-۱). با توجه به فرضیه سیگنال گونتر بلوبل و دیوید ساباتینی[2] (بلوبل و ساباتینی، ۱۹۷۱)، سیگنال پپتید توسط یک سیگنال شناسایی می‌شود، و زنجیره پلی‌پپتید نوظهور را از طریق غشای شبکه آندوپلاسمی هدایت می‌کند. به محض اینکه سیگنال پپتید غشا را گذراند، به صورت اختصاصی از پلی‌پپتید نوپا با آنزیم سیگنال پپتیداز جدا می‌شود. پروتئین‌هایی که همراه سیگنال پپتیدها هستند را پری‌پروتئین‌ها[3] می‌نامند همچنین پروتئین‌هایی که حاوی پروپپتیدها[4] هستند نیز پری‌پروپروتئین[5] نامیده می‌شوند. بر خلاف سیگنال پپتید، پروپپتیدها به روش پروتئولیتیک حذف می‌شوند تا منجر به فعال‌سازی پروتئین شوند (شکل ۵-۱).

---

[1] Signal Peptide
[2] Signal Hypothesis of Günter Blobel and David Sabatini
[3] Preproteins
[4] Propeptides
[5] Preproproteins



شکل ۵-۱) مثال فوق تصویر شماتیک از پری‌پروپروتئین می‌باشد که با سیستئین پروتئازهای خانواده پاپائین نشان داده شده است. اسیدآمینه‌های کاتالیتیک تراید (Cys25, His159, Asp175) هر کدام در داخل رشته‌های دنباله‌ای مشخص از پروتئازهای سیتئین قرار دارند (M1–M3). فقط عملکرد تعداد کمی از پروتئازهای سیستئینکه داراییک افزونه پایانهC هستند که هنوز مشخص نیست

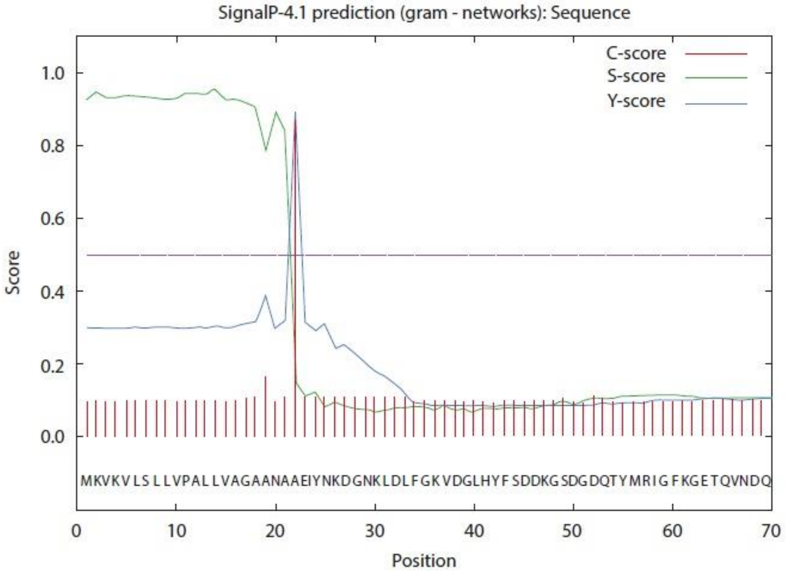
وجود یک سیگنال پپتید،یک نشانه مهم برای محل عمل پروتئین‌ها است. این دانش به نوبه خود می‌تواند به بهبود عملکرد کمک کند و بنابراین در تعیین اینکه آیا پروتئینیک مولکول هدف مناسب است کمک می‌کند. به همین دلیل، روش پیش‌بینی حضور سیگنال پپتید در ساختار اولیه توسعه یافته است. یک مثال، برنامه SignalP از مرکز تحلیل توالی زیستی (CBS)[1]در دانشگاه فنی دانمارک است (پترسون و همکاران، ۲۰۱۱). تشخیص سیگنال پپتید توسط ذرات تشخیص سیگنال به دلیل توالی اسیدآمینه محافظت شده نیست بلکه به خواص فیزیکی و شیمیایی بستگی دارد. سیگنال پپتید معمولاً شامل سه قسمت است. منطقه اول (منطقه n) شامل ۱ تا ۵ اسیدآمینه با بار مثبت است، منطقه دوم (منطقه h) از ۵ تا ۱۵ اسیدآمینه آب‌گریز ساخته شده است و منطقه سوم (منطقه c) ۳ تا ۷ عدد اسیدهای‌آمینه قطبی اما عمدتاً بدون بار است. بنابراین روش همردیفی کلاسیک توالی برای پیش‌بینی سیگنال پپتیدها مناسب نیست. برنامه SignalP در نسخه چهارم فعلی خود، بر اساس استفاده از شبکه‌های عصبی[2] است. با استفاده از روش‌های یادگیری ماشین، ویژگی‌های مجموعه داده‌های آموزشی با توالی‌های شناخته شده آموخته می‌شود و می‌تواند برای پیش‌بینی داده‌های

---

[1] Center for Biological Sequence Analysis (CBS)
[2] Neural Networks

ناشناخته مورد استفاده قرار گیرد. بنابراین شبکه‌های عصبی آموزش داده می‌توانند خواص اسیدهای آمینه را در توالی‌های ناشناخته قضاوت کنند، در نتیجه باعث تشخیص توالی سیگنال می‌شود. SignalP از دو شبکه عصبی مختلف استفاده می‌کند، زیرا سیگنال پپتید و مارپیچ‌های بین‌غشایی (۵-۵-۳) به سختی از هم قابل تشخیص هستند. بنابراین یک شبکه عصبی با توالی‌های سیگنال پپتید و دیگر شبکه عصبی با توالی‌های مارپیچ‌های بین‌غشایی آموزش داده می‌شوند. با استفاده از این روش، میزان مثبت کاذب سیگنال پپتیدهای پیش‌بینی شده می‌تواند به حداقل برسد.

قبل از اینکه تجزیه و تحلیل آغاز شود، مهم است که گروه موجود درست را انتخاب کنید زیرا باکتری‌های گرم منفی، باکتری‌های گرم مثبت و یایوکاریوت‌ها هستند. شکل ۵-۲ خروجی گرافیکی برنامه SignalP برای پروتئینC غشاء خارجی (پیش‌ساز) از *Salmonella typhimurium*، (OMPC-SALTY P0A263) را نشان می‌دهد. نمره C برای نمره محل تجزیه ارائه شده است که در شناخت محل شکاف بین سیگنال پپتید و توالی پروتئین آموزش داده می‌شود و محل شکافت SPase را پیش‌بینی می‌کند.



شکل ۵-۲) شکل خروجی داده‌های پایگاه SignalP در CBS

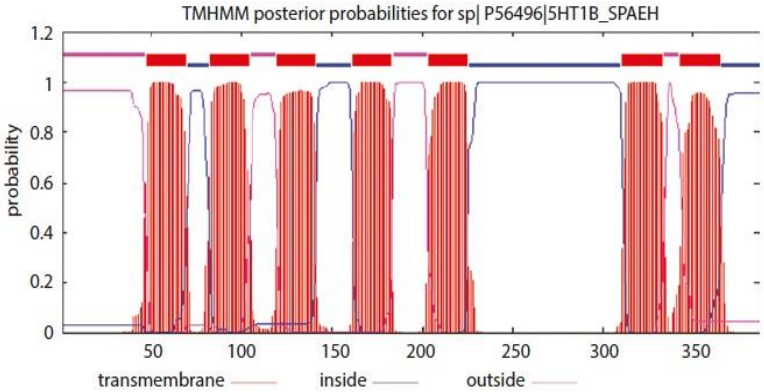
حرف I. بیانگر حداکثر نمره C در موقعیت اسیدآمینه اول پروتئین بالغ را نشان می‌دهد، که یک موقعیت پشت محل برش است. نمره S، نمره سیگنال پپتید است، برای تمایز سیگنال پپتیدها و سایر توالی‌ها آموزش داده می‌شود و اگر اسیدآمینه مربوطه بخشی از سیگنال پپتید باشد، مقدار بالایی دارد. بنابراین، اسیدآمینه پروتئین بالغ امتیاز S کمی دارد. نمره Y (نمره جایگاه برش ترکیب شده) یک میانگین هندسی از مقادیر مطلق نمره C و گرادیان نمره S است و نشان می‌دهد که در کدام محل نمره C بالا است و نمره S دارای نقطه انحنا است. تجزیه و تحلیل این سه نمره نشان می‌دهد که محل تجزیه احتمالی بین اسیدهای‌آمینه ۲۱ و ۲۲ است. علاوه بر این، دو شاخص دیگر نیز محاسبه می‌شوند. میانگین S، که بیانگر میانگین مقادیر S تمام اسیدهای‌آمینه سیگنال پپتید است. در نتیجه، اگر یک سیگنال پپتید وجود داشته باشد، این مقدار باید بالا باشد. نمره D، میانگین محاسبه مقدار میانگین S و حداکثر مقدار نمره‌ی Y است. اگر یک سیگنال پپتید پیش‌بینی شده باشد، این مقدار نیز بالا خواهد بود.

۵-۳ پروتئین‌های بین غشایی

غشاهای بیولوژیک حاوی پروتئین‌های داخلی هستند که دارای عملکردهای مختلفی در سلول هستند مانند عملکرد گیرنده‌های سطح سلولی. ادغام به لیپید غشا دو لایه توسط برهمکنش‌های هیدروفوبی (آبگریز) بین پروتئین و ساختارهای زنجیره غیرقطبی لیپیدها انجام می‌شود. گروه‌های قطبی از لیپیدها، پیوند هیدروژنی و پیوندهای یونی را با پروتئین می‌سازند. بنابراین پروتئین‌های غشایی داخلی همیشه مولکول‌های آمفی‌فیلی (دوگانه)[۱] هستند که دارای مناطق هیدروفیلی (آبدوست)[۲] و لیپوفیلی (چربی‌دوست)[۳] هستند. این پروتئین‌ها به صورت نامتقارن در غشا قرار دارند، یعنی، برخی از پروتئین‌های غشایی تنها در یک طرف غشا قرار می‌گیرند، در حالی که سایر پروتئین‌ها به طور کامل به غشا نفوذ می‌کنند و در دو طرف خارج سلولی و داخل سلولی قرار می‌گیرند. دومی، پروتئین‌های بین‌غشایی[۴] نامیده می‌شوند. دامین‌های بین‌غشایی آب‌گریز معمولاً توسط مارپیچ‌های آلفا تشکیل می‌شوند.

---

[1] Amphiphilic
[2] Hydrophilic
[3] Lipophilic
[4] Transmembrane Proteins



شکل ۵-۲) شکل خروجی داده‌های پایگاه TMHMMدرCBS

پیش‌بینی پروتئین‌های بین‌غشایی برای طبقه‌بندی و تعریف عملکرد، همانطور که قبلاً برای سیگنال پپتید شرح داده شده است، اهمیت زیادی دارد. برنامه TMHMM سرور پایگاهCBS در دانمارک، می‌تواند دامین‌های بین‌غشایی را پیش‌بینی کند. TMHMM مبتنی بر یک مدل مارکوف مخفی (HMM)[۱]است که برای تشخیص مارپیچ‌های بین‌غشایی آب‌گریز آموزش داده شده است. همچنین این برنامه جهت‌گیری دامین‌های منفرد در غشاء (درون سلولی یا خارج سلولی) و طبع آن، کل پروتئین را پیش‌بینی می‌کند.

خروجی گرافیکی چنین پیش‌بینی با TMHMM درشکل۵-۳ برای دامین‌های بین‌غشایی گیرنده جفت پروتئینG (GPCR)[۲] گیرنده ۵-هیدروکسی تریپتامین- 1B[۳] رت مول Spalax leucodon ehrenbergi (5H1B-SPAEH)[۴] نشان داده شده است. این نوع GPCRها از پروتئین غشایی داخلی با هفت دامین بین‌غشایی تشکیل شده‌اند. در نمودار، احتمال مارپیچ بین‌غشایی و محل قرارگیری داخل سلولی یا خارج سلولی آن در طول توالی اسیدآمینه قرار گرفته است. علاوه بر این، در قسمت بالای شکل، شماتیکی از توپولوژی پروتئین نیز ارائه شده است. نمایش گرافیکی احتمالات، همچنین باعث شناختن مارپیچ‌های بین‌غشایی نسبتاً اندک نیز می‌شوند.

---

[1] Hidden Markov Model (HMM)

[2] G Protein-Coupled Receptor (GPCR)

[3] 5-hydroxytryptamine-1B Receptor

[4] Mole Rat *Spalax leucodon ehrenbergi* (5H1B-SPAEH)

۵-٤ تحلیل ساختارهای پروتئینی

همانطور که قبلاً ذکر شد، در حال حاضر امکان پیش‌بینی ساختار سه‌بعدی پروتئین از توالی اسیدآمینه نمی‌باشد و در آینده نیز امکان‌پذیر نخواهد بود. بنابراین برای تعیین ساختارهای پروتئینی باید روش‌های تجربی مورد استفاده قرار گیرند. دو روش اصلی عبارتند از: کریستالوگرافی اشعه ایکس[1] و طیف سنجی رزونانس مغناطیسی هسته‌ای با کیفیت وضوح بالا (NMR)[2]. روش سوم، استفاده از میکروسکوپ الکترونی برای پروتئین‌های بزرگ است. در مجموع، علیرغم پیشرفت فنی بسیار زیاد، این روش‌ها هنوز هم بسیار وقت‌گیر و پرهزینه هستند و تفکیک وضوح موفق ساختار کریستالی برای هر پروتئین تضمین نمی‌شود.

۵-٤-۱ مدل سازی پروتئین

یک روش مفید و سریع برای پیش‌بینی ساختار، مدل‌سازی همولوژی پروتئین‌ها بر اساس همولوژی توالی است. رویکرد مبتنی بر این واقعیت است که پروتئین‌های مربوطه در یک خانواده پروتئینی که دارای درجه بالایی از تشابه توالی اسیدآمینه هستند نیز پیچش مشابه پروتئینی دارند (مانند پروتئازهای سیستئین خانواده کاتپسین[3]) (همچنین به شکل ۵-۵ و ۵-۶ مراجعه شود). پروتئین‌هایی که ساختار سه‌بعدی آن‌ها قبلاً شناخته شده، به عنوان پروتئین‌های مرجع یا الگوهستند. ابتدا توالی اسیدآمینه پروتئینی که باید مدل‌سازی شود با توالی پروتئین(های) مرجع با استفاده از همردیفی دوگانه یا چندگانه (در مورد چندین پروتئین مرجع) مقایسه می‌شوند. برای توالی‌هایی با همسانی بیش از ۷۰ درصد، ساختار مدل شده را می‌توان بسیار دقیق پیش‌بینی کرد. با این حال، برای دنباله‌ای با همسانی کمتر از ۳۰ درصد، مشکلات مدل‌سازی بوجود می‌آید. همسانی‌های توالی از مناطق محافظت‌شده ساختاری (SCR)[4] اغلب بالاتر از حلقه‌های کمتر محافظت‌شده است و هر دو بر درجه همسانی توالی کامل تأثیر می‌گذارند. جالب توجه است که نواحی با حفاظت کمتر اغلب در سطح پروتئین

---

[1] X-ray Crystallography
[2] High-resolution Nuclear Magnetic Resonance(NMR) Spectroscopy
[3] Cysteine Proteases of the Cathepsin Family
[4] Structurally Conserved Regions (SCRs)

یافت می‌شوند و اثر نسبتاً کوچکی بر روی SCRها دارند، که در داخل پروتئین یافت می‌شود و در اکثر مراکز فعال قرار دارند.

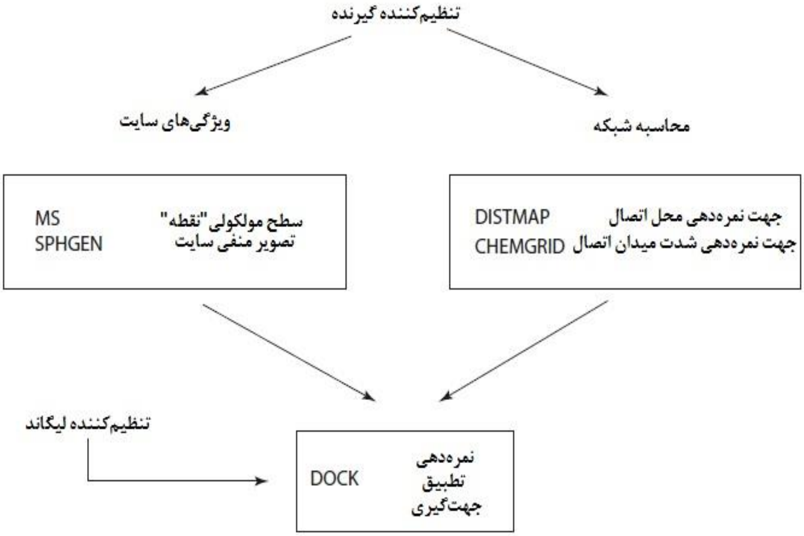
برای شناسایی SCRها در پروتئین‌های مرجع، همردیفی ساختاری توالی‌های اسیدآمینه بر اساس ساختار ثانویه انجام می‌شود. توالی که باید مدل‌سازی شود، بر روی الگوهای جهت‌گیری شده آرایش یافته، و سپس مختصات فضایی SCRها به توالی مدل منتقل می‌شود. مختصات حلقه‌ها معمولاً از مناطق مشابه ساختارهای پروتئینی دیگر گرفته می‌شود. جهت‌گیری فضایی زنجیره‌های جانبی اسیدهای‌آمینه فردی در SCRها در قالب‌ها نگهداری می‌شوند. برای تمام زنجیره‌های جانبی محافظت نشده، موقعیت احتمالی آماری به دست می‌آید. روند مدل‌سازی همولوژی هم با محاسباتی انجام می‌شود که منجر به کمینه‌سازی مدل و بررسی ارتباط ساختاری مدل پروتئینی حاصل می‌شود. از سرور SWISSMODEL مؤسسه بیوانفورماتیک در لوزان سوئیس برای محاسبه خودکار مدل‌های همولوژی استفاده می‌شود (بیاسینی و همکاران، ۲۰۱۴). در مورد پروتئین‌هایی با شباهت توالی بالا، مدل محاسبه شده اغلب با کیفیت بالا است.

۲-۴-۵ تعیین ساختارهای پروتئین با روش‌هایی با عملکرد بالا

تعداد ساختارهای پروتئینی تعیین شده تجربی که در تنها بایگانی جهانی ساختارهای ماکرومولکول‌های زیستی، بانک داده‌های پروتئینی (PDB)[1] ذخیره می‌شود، در دهه‌های اخیر به شدت افزایش یافته است (وستبروک و همکاران، ۲۰۰۳). در سال ۱۹۷۲، PDB تنها یک ساختار داشت، در سال ۱۹۹۲ تعداد آن تقریباً ۱۰۰۰ بود و تا آوریل ۲۰۰۳ مقدار اطلاعات آن به ۲۰۶۲۲ افزایش یافت. در نوامبر ۲۰۱۶، PDB شامل ۱۲۴۰۲۹ ساختار بوده است.

---

[1] Protein Data Bank (PDB)



شکل ۵-۴) طرح شماتیک از نحوه عملکرد برنامه DOCK

این افزایش قابل توجه در اطلاعات به طور عمده به فرایند تکنولوژی شامل خودکار شدن و روش‌های با راندمان بالا برای حل ساختار مربوط می‌شود. ابتکار ساختار پروتئین[1]، یکی از مهمترین عوامل این پیشرفت بود. این ابتکار یک کنسرسیوم علمی‌بین‌المللی از ابتکارهای ملی مختلف از ژاپن، آمریکای شمالی و اروپا بود. هدف چیزی نبود جز اینکه از نظر ساختاری، همه پروتئین‌های کد شده در ژنوم مهم‌ترین موجودات (آرکی باکتری، یوباکتری‌ها، و یوکاریوت‌ها) حل شوند.

برای حل ساختارها، تجزیه و تحلیل ساختاری اشعه ایکس و طیف سنجی NMR در فرمت با عملکرد بالا استفاده می‌شود. برای کاهش تعداد ساختارهای پروتئینی که به صورت آزمایشی حل می‌شوند، تنها نمایندگان خانواده‌های مختلف پروتئینی مورد بررسی قرار گرفتند. ایده اصلی این بود که پروتئین‌ها را می‌توان به خانواده‌های پروتئینی تقسیم کرد و این تشابه توالی معمولاً به تشابه ساختاری منجر می‌شود. نتیجه‌گیری این است که تعدادی از پیچش‌های مختلف ساختاری پروتئین‌های موجود در طبیعت باید محدود باشند. یک برآورد این است که بین ۱۰۰۰۰ تا ۳۰۰۰۰ خانواده پروتئین وجود دارد، و این حاوی تقریباً ۱۰۰۰-۵۰۰۰

---

[1] Protein Structure Initiative

پیچش‌های مختلف پروتئین است. در این بین، حدود ۱۴۰۰ پیچش در حال حاضر شناخته شده است. با این حال باید در نظر داشت که ساختارهای پروتئینی مشابهی به صورت اجتناب‌ناپذیر، عملکردهای مشابهی ندارند و ساختارهای مختلف پروتئین نیز ممکن است عملکرد مشابهی داشته باشند. به عنوان مثال، پروتئازهای سیستئین به سه گروه ساختاری متفاوت بر اساس الگوهای پیچش پروتئین تقسیم می‌شوند: پروتئازهای شبه پاپاین[1]، پروتئازهای ویروس پیکورنا[2] و کاسپازها[3].

برای رسیدن به اهداف بلندپروازانه ابتکار ساختاری پروتئین، استراتژی به شرح زیر بود:

۱- تمام توالی‌های شناخته شده پروتئین به خانواده‌های پروتئین با استفاده از روش‌های بیوانفورماتیک گروه‌بندی شدند.

۲- نمایندگان هر خانواده پروتئین با مقدار کافی توسط روش‌های بیولوژیکی مولکولی تولید شدند.

۳- ساختارهای پروتئینی این نمونه‌ها با استفاده از کریستالوگرافی پروتئین و طیف سنجی NMR آزمایش شدند.

۴- تمام ساختارهای دیگر پروتئینی خانواده‌های پروتئینی مرتبط، با مدل‌سازی همولوژی تولید شدند.

با استفاده از این روش، مقدار زیادی از الگوهای پیچش جدید پروتئینی شناسایی شدند و در نتیجه، نقش مهمی‌در شفاف‌سازی عملکرد همه پروتئوم‌های شناخته شده ایفا کرد. در عین حال، مزایای تحقیقات دارویی مدرن مورد بررسی قرار گرفت، از آنجا که بسیاری از ساختارهای پروتئینی بدون تفسیر عملکرد حل شده بودند. با این حال، نتایج در آینده ارزش فراوانی خواهد داشت. بنابراین ابتکار کنونی، که کنسرسیوم ژنومیکس ساختاری نامیده می‌شود، بیشتر بر راه حل ساختار پروتئین‌های مرتبط با بیماری متمرکز است. باید این ساختارها را برای طراحی منطقی دارو مبتنی بر ساختار استفاده کرد و به صورت معنی‌داری از پیشرفت داروها حمایت کرد (بورلی و بونانو، ۲۰۰۲).

---

[1] Papainlike Proteases

[2] Picorna Virus Proteases

[3] Caspases

۵-۵ طراحی مبتنی بر ساختار دارو

از توالی‌یابی ژنوم‌های کامل و تولید اطلاعات بیولوژیکی مربوط، یک رویکرد مدرن برای تحقیقات فارماکولوژیک ایجاد شده است. برای شروع توسعه یک داروی جدید، ابتدا باید هدف دارو (در عمل، یک پروتئین) که نقش مهمی‌در بیماری ایفا می‌کند، شناسایی شود (فصل ۷). بعد از اینکه عملکرد هدف به صورت تجربی تایید شد (اعتبار سنجی هدف دارو)، مواد شیمیایی با مولکول کوچک شناسایی می‌شوند که بر عملکرد پروتئین به گونه‌ای اثر می‌گذارد که بیماری را تسکین دهند یا درمان کنند. ممانعت اختصاصی یک آنزیم توسط مهارکننده‌های شیمیایی می‌تواند یک نمونه باشد.

فناوری‌های همپوشانی رویکردهای کامپیوتری مانند بیوانفورماتیک، کموانفورماتیک و طراحی مولکولی، به اجزای ضروری تلاش‌های کشف داروی جدید تبدیل شده است. این استراتژی‌ها برای شناسایی و اعتبار سنجی اهداف دارویی و همچنین برای غربال‌گری و طراحی مولکول‌های کوچک، ضروری است. همچنین ساختار سه‌بعدی هدف دارو از اهمیت ویژه‌ای برخوردار است که برای طراحی منطقی داروی بر اساس ساختار است. برای مثال، غربال‌گری مجازی، که برهمکنش اهداف پروتئین را با ماهیت شیمیایی در کتابخانه‌های بزرگ آزمایش می‌کند، رویکردی تثبیت شده است که در بیشتر استراتژی‌های کشف شده گنجانده شده است. بر خلاف آزمایش‌های انجام شده در آزمایشگاه، غربال‌گری مجازی اتوماتیک می‌شود، که در یک کامپیوتر انجام می‌شود و بسیاری از مواد شیمیایی را می‌توان برای طیف فعالیت آن‌ها، به صورت نسبتاً سریع آزمایش کرد. مهمترین رویکردها، بررسی غربال‌گری مبتنی بر فارماکوفور[۱] (وولبر و همکاران، ۲۰۰۸) و داکینگ[۲] آن‌ها (کیتچن و همکاران، ۲۰۰۴) است.

واژه داکینگ، تفسیر تصویری مدرن از مفهوم قفل و کلیدی است که در سال ۱۸۹۴ توسط امیل فیشر (فیشر ۱۸۹۴) ارائه شده است. اختصاصیت کمپلکس گیرنده-لیگاند با مکمل بودن هندسی و فیزیکوشیمیایی هر دو به وجود می‌آید. تناسب القایی، نوع دیگری از این فرضیه است، جایی که هندسه محل اتصال، با اتصال لیگاند سازگار می‌شود. بهترین برنامه‌های شناخته شده مورد استفاده، DOCK است که توسط ایروین کونتز در دانشگاه کالیفرنیا، سان فرانسیسکو (اوینگ و کونتز ۱۹۹۶)[dock]، GOLD از مرکز داده کریستالوگرافی کمبریج

---

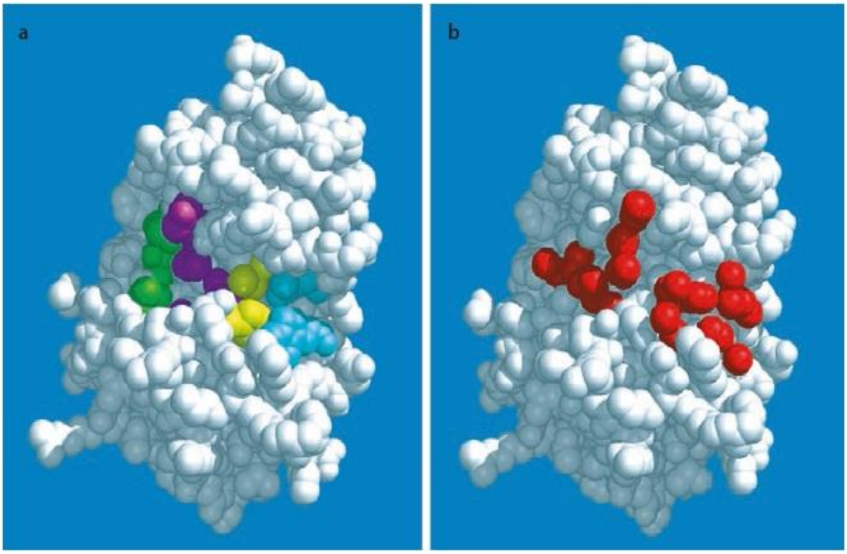
[1] Pharmacophore

[2] Docking

(جونز و همکاران، ۱۹۹۷)، FlexX از BioSolveIT GmbH در Sankt Augustin (راری و همکاران، ۱۹۹۶) و Autodock که در مؤسسه تحقیقاتی اسکریپس(موریس و همکاران ۲۰۰۹) توسعه یافت.

۵-۵-۱ مثال داکینگ با استفاده از DOCK

با DOCK، می‌توان تمام جهت‌های ممکن یک لیگاند و گیرنده آن را تولید کرد. به عنوان مثال، ساختار پروتئینی یک آنزیم با یک مرکز فعال تعریف شده به طور مشخص می‌تواند یک گیرنده معمولی را بسازد. ساختار لیگاند از پایگاه داده مولکول‌های شیمیایی مانند Available Chemicals Directory شروع می‌شود. در مثال نشان داده شده، پروتئاز سیستئین شبه-کتپسین L[1] از لارو مرحله سوم کرم فلیاریایی *Brugia pahangi* به عنوان گیرنده عمل می‌کند. این آنزیم برای پوست‌اندازی و توسعه این انگل مهم است. ساختار پروتئین توسط مدل‌سازی همولوگ انجام شد.



شکل ۵-۵) شکل شماتیک کروی از پروتئاز سیستئین شبه-کتپسین

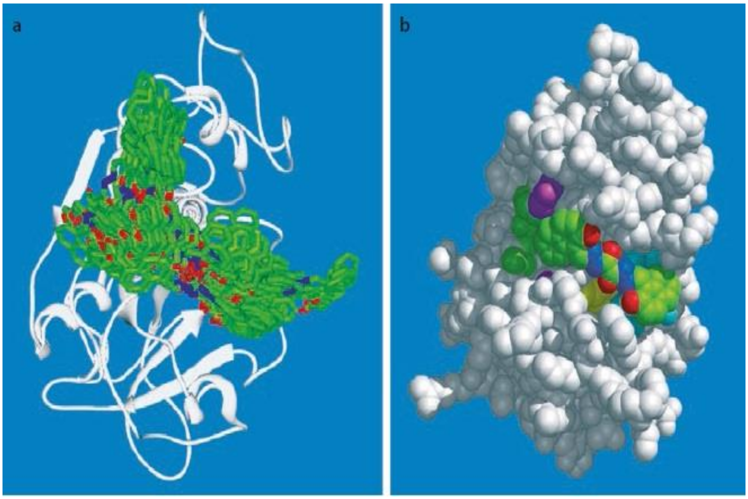
---

[1] Cathepsin L-like Cysteine Protease

۱- اولین قدم، مشخص کردن مرکز فعال (شکل ۵-۴) است. برای انجام این کار، سطح مولکولی مرکز فعال ابتدا (زیربرنامه MS) تولید می‌شود و به یک تصویر منفی (زیربرنامه SPHGEN) تبدیل می‌شود. حوزه‌های همپوشانی پس از آن در مرکز فعال قرار می‌گیرند (شکل ۵-۵). در نهایت مرکز کره‌ها توسط اتم‌های لیگاند جایگزین خواهد شد.

۲- در مرحله دوم، محاسبه پارامترهای فیزیکی، شیمیایی و توپولوژیکی در هر نقطه‌ی گره یک شبکه فضایی انجام می‌شود (محاسبات شبکه) تا نمره محاسبه شود که می‌تواند یک نمره تماس بر اساس تناسب لیگاند یا نمره میدان نیرو باشد.

۳- پس از این محاسبات، اتصال واقعی می‌تواند انجام شود. این حالت را می‌توان در دو حالت، حالت DOCK ساده یا حالت DOCK جستجو، انجام داد. در حالت ساده، DOCK تمام جهت‌گیری‌های ممکن یک لیگاند تک را در مرکز فعال تولید می‌کند (شکل ۵-۶). در حالت جستجو، پایگاه‌داده‌های بزرگ مولکول‌های شیمیایی جستجو می‌شود. برای انجام این کار، بهترین جهت‌گیری هر لیگاند ابتدا تولید می‌شود و سپس به عنوان یک امتیاز نسبی در مقایسه با سایر لیگاندها ذخیره می‌شود. اتصالات با نمرات بالاترین رتبه برای اندازه، تناسب و تعامل با مرکز فعال مورد بررسی قرار می‌گیرد. سپس بهترین ترکیبات می‌توانند به صورت آزمایشی در سنجش‌های مناسب آزمون شوند.



شکل ۵-۶) شکل شماتیک کروی از پروتئاز سیستئین شبه-کتپسین

برای مثال سیستئین پروتئاز *Brugia pahangi* (لکایله و همکاران، ۲۰۰۲) یک پایگاه داده شیمیایی از مهارکنندگان سیستئین پروتئاز شناخته شده با DOCK مورد بررسی قرار گرفت و ترکیبات هیدرازید شناخته شده برای مهار سیستئین پروتئازها یا نگل‌های *Trypanosoma cruzi*، *Trypanosoma brucei*، *Leishmania major* و *Plasmodium falciparum* نمرات بسیار بالایی داشتند. سپس اتصال هیدرازیدها شناسایی شده، در حالت ساده DOCK برای شناسایی مهارکننده‌های امیدوارکننده بیشتر مورد بررسی قرار گرفت (شکل ۵-۶).

آزمایش‌های بعدی با بهترین مهارکننده‌های پیش‌بینی شده، مانع توسعه لارو مرحله سوم عفونی به لارو مرحله چهارم شدند (سلزر، ۲۰۰۳).

۵-۵-۲ داکینگ با استفاده از GOLD

GOLD یکی دیگر از نرم‌افزارهای گسترده داکینگ است. کنفورماسیون لیگاند در ناحیه پیوند با استفاده از یک الگوریتم ژنتیکی محاسبه می‌شوند که بر تکامل ژنتیک طبیعی (شکل ۵-۷) استوار است. کنفورماسیون سه‌بعدی مولکول از طریق زاویه‌های پیچشی آن، که به عنوان یک بیت‌وکتور (کروموزوم) ذخیره می‌شوند، توصیف می‌شود. مشابه طبیعت، این نشان دهنده ژنوتیپ کنفورماسیون لیگاند است. فرایندهای تکاملی با جهش بیت‌های تکی یا مبادله بخش‌های تکی bitvector دو کروموزوم مختلف شبیه‌سازی می‌شوند. بنابراین، کنفورماسیون سه‌بعدی بر اساس تغییر تصادفی کروموزوم‌آن تغییرمی‌کند. بنابراین، کنفورماسیون سه‌بعدی (فنوتیپ) بر اساس کروموزوم ایجاد خواهد شد و در محل اتصال قرار می‌گیرد. هر موقعیت اتصال با استفاده از تناسب آن بر اساس یک عمل نمره‌دهی به دست می‌آید. فقط موقعیت‌هایی که برهمکنش‌های مطلوب برای پروتئین تناسب بالایی نشان می‌دهد، برای دور بعدی الگوریتم ژنتیکی انتخاب می‌شود. این مراحل تا زمانی که نمره پایدار به دست می‌آید تکرار می‌شود.

برهمکنش پروتئین-پروتئین[1] بین تیوردوکسین ردوکتاز (TrxR) و سوبسترای تیوردوکسین (Trx) از *Mycobacterium tuberculosis* یک هدف جدید برای مبارزه با سل است. نرم‌افزار داکینگ GOLD برای شناسایی اولین مهارکننده‌های این برهمکنش پروتئین–پروتئین با موفقیت مورد استفاده قرار گرفت (کخ و همکاران، ۲۰۱۳). برهمکنش‌های پروتئین–

---

[1] Protein–Protein Interaction

پروتئین به طور کلی یک چالش خاص است، زیرا بیشتر آن‌ها بر اساس سطح پروتئین‌ها بدون هیچ پاکت پیوند عمقی درگیر هستند. یک تجزیه و تحلیل دقیق از ساختار اشعه ایکس موجود، یک نقطه جالب از هدف را بیان کرد. یک زنجیر جانبی آرجین بیرون کره تیورودوکسین مانند یک گروه لنگر به نظر می‌رسد (شکل ۵-۸، قسمت بالا)، همراه با یک شکاف هیدروفوبیک و یک اتصال هیدروژن اضافی در نزدیکی آن. GOLD به طور موفقیت‌آمیزی برای غنی‌سازی مولکول‌ها از یک کتابخانه مجازی از ۶/۵ میلیون ترکیب به کار می‌رود که احتمالاً هر دو برهمکنش اتصالات هیدروژنی توصیف شده را نشان می‌دهد (شکل ۵-۸، بخش پایین). ۱۷۰ تا از بهترین رتبه‌بندی در یک آزمایش بیوشیمیایی مورد ارزیابی قرار گرفتند که ۱۸ مولکول ممانعت نشان دادند. به طور کلی این نشان دهنده میزان ضربه ۱۰/۵ درصد است. با وجودی که می‌توان انتظار میزان ضربه بیشتری را داشت که با استفاده از رویکردهای داکینگ در نگاه اول به نظر می‌رسد، این یک نتیجه قابل توجه در مقایسه با جایگزین ممکن است. فقط ساختار اشعه ایکس در ابتدای مطالعه شناخته شد. بنابراین، تمام ۶/۵ میلیون ترکیب مورد آزمایش قرار گرفتند تا اولین مهارکننده‌ها را پیدا کنند. تلاش‌های تجربی در مقایسه با ۱۷۰ ترکیب آزمایش شده غیرمتناسب بود.

۵-۵-۳ مدل‌سازی فارماکوفور و تحقیقات

دانستن ساختار سه‌بعدی یک گیرنده برای داکینگ در غربال مجازی ضروری است. در غیاب یک ساختار سه‌بعدی یا مدل همولوژی، نمی‌توان از روی نمایش مجازی استفاده کرد. تا زمانی که برخی از لیگاندهای گیرنده شناخته شده‌اند، غربال‌گری مجازی براساس مدل فارماکوفور می‌تواند انجام شود. یک مدل فارماکوفور یک مفهوم انتزاعی است که پتانسیل برهمکنش یک مولکول با پروتئین هدفش را در نظر می‌گیرد و آرایش فضایی خواص لیگاند که مسئول اتصال هستند را توصیف می‌کند (وولبر و همکاران، ۲۰۰۸). برای ساختن یک مدل فارماکوفور، می‌توان از چندین مهارکننده شناخته شده یا لیگاند فعال پروتئین بر اساس خواص فارماکوفور استفاده کرد. ویژگی‌های احتمالی فارماکوفور عبارتند از گیرنده‌ها و دهندگان هیدروژنی، سیستم‌های هیدروفوبی و آروماتیکی و بارها (شکل ۵-۹). آرایش فضایی ویژگی‌های فردی برای بازیابی مدل فارماکوفور یا فرضیه فارماکوفور مورد تجزیه و تحلیل قرار می‌گیرد. با تحلیل

کتابخانه‌های مجازی بر اساس این مدل فارماکوفور، مولکول‌هایی می‌توانند شناسایی شوند که الگوی برهمکنش فضایی مشابه را با فعالیت مشابه بالقوه نشان می‌دهد.

تمام کنفورماسیون‌های ممکن سه‌بعدی از کتابخانه مجازی باید برای فرایند غربال‌گری نهایی ایجاد شود، زیرا آرایش فضایی ویژگی‌های فارماکوفور در مدل فارماکوفور تناسب یافته و در مقایسه با آن قرار گرفته است. همپوشانی با نمره توصیف می‌شود و مولکول‌های با توافق بالا می‌توانند به عنوان لیگاندهای بالقوه برای ارزیابی تجربی مورد استفاده قرار گیرند. زمان محاسبه شده کاهش یافته در مقایسه با داکینگ، سود عظیمی‌از جستجوهای فارماکوفور است. به همین دلیل آن‌ها اغلب برای کاهش اندازه کتابخانه‌های مجازی و فیلتر کردن برای داکینگ بعدی استفاده می‌شوند. نرم‌افزار برای مدل‌سازی فارماکوفورها و جستجوهای فارماکوفور برای مثال شامل موارد زیر است: مدل‌سازی فارماکوفور MOE، Phase (دیکسون و همکاران ۲۰۰۶) و Ligandscoutمی‌باشد (وولبر و همکاران، ۲۰۰۷).

هنگامی که ساختار سه‌بعدی گیرنده شناخته شده است، مدلسازی فارماکوفور مبتنی بر گیرنده نیز می‌تواند مورد استفاده قرار گیرد. علاوه بر این، ساختارهای پیچیده پروتئین-لیگاند می‌توانند برای ایجادیک مدل فارماکوفور که شامل اطلاعاتی درباره ساختار پروتئین و لیگاند شناخته شده باشد، استفاده شوند (وولبر و همکاران، ۲۰۰۷).

٤-٥-٥ موفقیت طراحی منطقی داروی مبتنی بر ساختار

سؤالیکه اغلب پرسیده می‌شود این است که آیا چنین روش‌های مجازی در واقع منجر به دارو می‌شود. پاسخ به وضوح بله است. مثال‌های بیشتری وجود دارد که در اینجا می‌توان ذکر کرد که فن‌آوری‌های مجازی به طور قابل توجهی به توسعه داروها کمک کرده‌اند. با این حال باید توجه داشت که توسعه یک داروی یک فرآیند سختگیرانه است که شامل مراحل مختلفی می‌شود. طراحی منطقی دارو تنها یک گام اول در راه طولانی به یک داروی قابل فروش است.

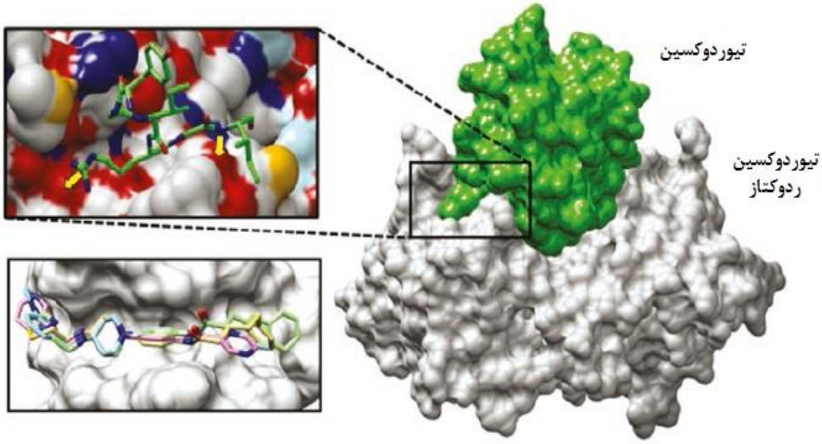
شکل ۵-۷) الگوریتم ژنتیک استفاده شده توسط نرم افزار GOLD

دورزولامید[۱] (نام تجاری Trusopt توسط مِرک از سال ۱۹۹۵) که برای درمان گلوکوم[۲] استفاده شده استیک مهارکننده انهیدراز کربنیک[۳] است که به عنوان اولین دارو از برنامه‌ای که شامل طراحی منطقی مبتنی بر ساختار است، آغاز شد. مثال دوم، کاپتوپریل[۴]، داروئی است که فشار خون را کاهش می‌دهد،که ساختار سرب آن بر اساس یک ماده طبیعی است که آنزیم تبدیل‌کننده آنژیوتانسین(ACE)[۵]را مهار می‌کند. انالاپریل[۶]، یکی دیگر از مهار کننده‌های مؤثر ACE، توسعه بیشتر کاپتوپریل است. نمونه‌های بیشتر بازدارنده‌های پروتئازی HIV، ساکیناویر و ریتوناویر (Norvir)[۷] به ترتیب از شرکت روش و ابو ت[۸] هستند؛ مهارکننده تیروزین کیناز Gleevec[۹] از Novartis، که در بیماران دارای سرطان خون استفاده می‌شود؛ و

---

[1] Dorzolamid
[2] Glaucoma
[3] Carbonic Anhydrase Inhibitor
[4] Captopril
[5] Angiotensin-Converting Enzyme (ACE)
[6] Enalapril
[7] Saquinavir and Ritonavir (Norvir)
[8] Roche and Abbott
[9] Tyrosine Kinase Inhibitor Gleevec

مهارکننده‌های نورامینیداز تامیفلو[1]، از Roche و رلنزا، از GlaxoSmithKline، که هرگز بدون طراحی داروهای منطقی توسعه نمی‌یافتند (کلب، ۲۰۱۳).



شکل ۵-۸) تریوردوکسین ردوکتاز (خاکستری) و بستر آن تیرودوکسین (سبز). برش بالا نشان دهنده نقطه هدف برای غربالگری مجازی مبتنی بر docking است(فلش های زرد: تعامل هیدروژنی). برش پایینی نشان می دهد که اتصال دهنده ها از چهار مهار کننده یک داربست مشترک را به اشتراک می گذارند

نمونه‌هایی وجود دارد که برنامه DOCK با موفقیت مورد استفاده قرار گرفته است. مطالعاتی با پروتئازهای سیستئینی به طور خاص مؤثر بوده‌اند. با استفاده از مدل‌های DOCK و همولوژی پروتئازهای سیستئین *Leishmania major*، مولکول‌های کوچکی شناسایی شدند که این آنزیم‌های مورد هدف دارو را متوقف می‌کنند و مانع از نمو پروماستیگوت و آماستیگوت لیشمانیا[2] در کشت سلولی بدون آسیب به سلول‌های میزبانمی‌شوند. در یک مدل موش از لیشمانیوز، پیشرفت عفونت به طور قابل ملاحظه‌ای به تأخیر افتاد (سلزر و همکاران، ۱۹۹۷، ۱۹۹۹). نتایج مشابهی در مدل‌های حیوانی برای پروتئازهای سیستئین *Plasmodium falciparum* (شنای و همکاران ۲۰۰۲)، *Trypanosoma cruzi* (انگل و همکاران ۱۹۹۸) و *Schistosoma mansoni* (عبداله و همکاران ۲۰۰۷) به دست آمد. برای *Trypanosoma*

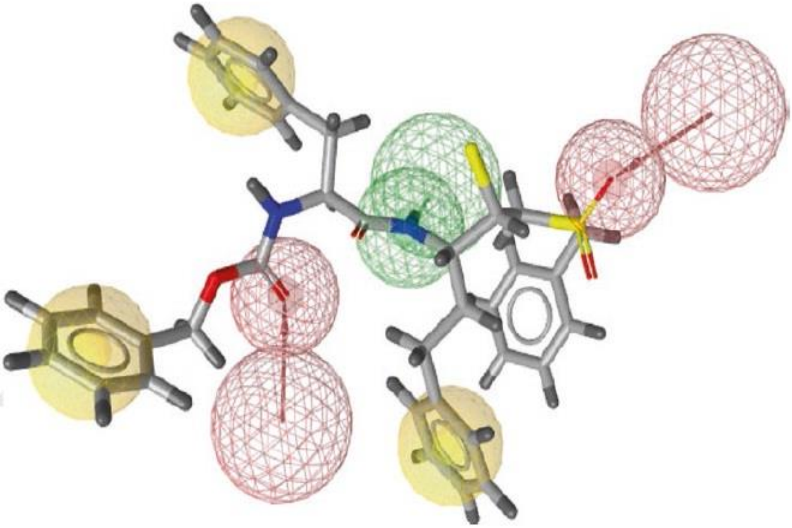
---

[1] Neuraminidase Inhibitors

[2] Promastigote and Amastigote Leishmania

cruzi، موفقیت مهارکننده‌های سیستئین پروتئاز، مرحله‌ای را برای آزمایش‌های بالینی در برابر بیماری چاگاس[1] تعیین کرده است (بار و همکاران، ۲۰۰۵).

طراحی منطقی نیز برای توسعه مهارکننده‌های پروتئازوم انگل‌ها استفاده شد. پروتئازوم یک کمپلکس چندجزئی پروتئاز است که برای مثال، فرآیندهای مهم چرخه سلولی را تنظیم می‌کند. تجزیه و تحلیل دقیق اختصاصیت سوبسترا و ساختار پروتئین منجر به مهارکننده‌های پروتئازوم انتخابی*Plasmodium falciparum* شد (لی و همکاران، ۲۰۱۶). این مهارکننده‌ها قادر به مهار رشد انگل in vivo بدون تأثیر بر سلول‌های میزبان می‌باشند. یکی دیگر از مهارکننده‌های پروتئازوم، قادر به مهار پروتئازوم کینتوپلاستیده‌ا[2] است. تمام انگل‌ها در مطالعات in vivo با استفاده ازیک مدل موش کشته شدند (خره و همکاران، ۲۰۱۶).



شکل ۵-۹) نماینده یک مدل فارماکوفور. ویژگی‌های فارماکوفور بصورت حوزه‌های رنگی نشان داده شده است. خواص آروماتیک (زرد)، گیرنده هیدروژن پذیرنده (قرمز) و اهداکننده (سبز) نشان داده شده است.

---

[1] Chagas Disease

[2] Proteasomeof Kinetoplastids

سایت‌های مفید

chimera. https://www.cgl.ucsf.edu/chimera/
dock. http://dock.compbio.ucsf.edu/
expasy. http://www.expasy.org
flexx. https://www.biosolveit.de/FlexX/
gold. https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/
ligandscout. http://www.inteligand.com/ligandscout/
moe. https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm
pdb. http://www.rcsb.org/
phase. https://www.schrodinger.com/phase
signalp. http://www.cbs.dtu.dk/services/SignalP/
spdbv. http://www.expasy.org/spdbv/
swiss-model. https://swissmodel.expasy.org/
swiss-prot. http://www.expasy.org/sprot/
tmhmm. http://www.cbs.dtu.dk/services/TMHMM/
uniprot. http://www.uniprot.org

منابع

1. Abdulla MH, Lim KC, Sajid M, McKerrow JH, Caffrey CR (2007) Schistosomiasis mansoni: novel chemotherapyusing a cysteine protease inhibitor. PLoS Med 4:e14.
2. Barr SC, Warner KL, Kornreic BG, Piscitelli J, Wolfe A, Benet L, McKerrow JH (2005) A cysteine protease inhibitor protects dogs from cardiac damage during infection by Trypanosoma cruzi. Antimicrob Agents Chemother 49:5160–5161.
3. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(W1):W252–W258.
4. Biobel G, Sabatini DD (1971) In: Manson LA (ed) Biomembranes. Plenum, New York, pp 193–195.
5. Burley SK, Bonanno J (2002) Structuring the universe of proteins. Ann Rev Genomics Hum Genet 3:243–262.
6. Dixon SL, Smondyrev AM, Rao SN (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem Biol Drug Des 67:370–372.
7. Engel JC, Doyle PS, Hsieh I, McKerrow JH (1998) Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection. J Exp Med 188:725–734.

8. Ewing TJA, Kuntz ID (1996) Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem 18:1175–1189.
9. Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges 27:3189–3232. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748.
10. Khare S, Nagle AS, Biggart A, Lai YH, Liang F, Davis LC, Barnes SW, Mathison CJ, Myburgh E, Gao MY, Gillespie JR, Liu X, Tan JL, Stinson M, Rivera IC, Ballard J, Yeh V, Groessl T, Federe G, Koh HX, Venable JD, Bursulaya B, Shapiro M, Mishra PK, Spraggon G, Brock A, Mottram JC, Buckner FS, Rao SP, Wen BG, Walker JR, Tuntland T, Molteni V, Glynne RJ, Supek F (2016) Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness. Nature 537(7619):229–233.
11. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935–949.
12. Klebe G (2013) Drug design. Springer, Heidelberg.
13. Koch O, Jäger T, Heller K, Khandavalli PC, Pretzel J, Becker K, Flohé L, Selzer PM (2013) Identification of M. tuberculosis thioredoxin reductase inhibitors based on high-throughput docking using constraints. J Med Chem 56(12):4849–4859.
14. Lecaille F, Kaleta J, Brömme D (2002) Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design. Chem Rev 102:4459–4488.
15. Li H, O’Donoghue AJ, van der Linden WA, Xie SC, Yoo E, Foe IT, Tilley L, Craik CS, da Fonseca PC, Bogyo M (2016) Structure- and function-based design of Plasmodium-selective proteasome inhibitors. Nature 530(7589):233–236.
16. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J Comput Chem 16: 2785–2791.
17. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786.
18. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489.
19. Selzer PM (2003) Structure-Based-Rational-Drug-Design: Neue Wege der modernen Wirkstoffentwicklung. In: Lucius R, Hiepe T, Gottstein B (eds) Grundzüge der allgemeinen Parasitologie. Parey, Berlin.

20. Selzer PM, Chen X, Chan VJ, Cheng M et al (1997) Leishmania major: molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors. Exp Parasitol 87:212–221.
21. Selzer PM, Pingel S, Hsieh I, Ugele B et al (1999) Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. Proc Natl Acad Sci U S A 96:11015–11022.
22. Shenai BR, Semenov AV, Rosenthal PJ (2002) Stage-specific antimalarial activity of cysteine protease inhibitors. Biol Chem 383:843–847.
23. Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The protein data bank and structural genomics. Nucleic Acids Res 31:489–491.
24. Wolber G, Dornhofer AA, Langer T (2007) Efficient overlay of small organic molecules using 3D pharmacophores. J Comput Aided Mol Des 20(12):773–788.
25. Wolber G, Seidel T, Bendix F, Langer T (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 13(1–2):23–29.

فصل ششم

تجزیه و تحلیل عملکردی ژنوم

۱-٦ شناسایی عملکرد سلولی محصولات ژن

نخستین توالی‌یابی ژنوم انسان در سال ۲۰۰۱ توسط پروژه ژنوم انسان[۱] منتشر گردید. در آن زمان تعداد ژن‌های انسان بین ۳۰۰۰۰ تا ۳۵۰۰۰ ژن تخمین زده شده بودند. امروزه مشخص شده است که ژنوم انسانی از نظر فیلوژنیک بسیار جوان بوده و تفاوت بسیار زیادی بین تعداد ژن‌ها و اندازه ژنوم نمایان می‌باشد (Ezkurdia *et al.* 2014). آن محتوی ۱۹۰۰۰تا ۲۰۰۰۰ ژن با اندازه ۳/۳ گیگاباز می‌باشد (مراجعه شود به فصول ۴ و ۷). هر سلول انسان به استثنای اسپرم و تخمک یک سری کاملی از ژن دارند. بدیهی است با این حال یک سلول خونی در مورفولوژی و فیزیولوژی خود از سلول کبدی متفاوت می‌باشد. بنابراین چگونه این تفاوت‌ها می‌تواند توضیح داده شود اگر مواد ژنتیکی همه این سلول‌ها یکسان می‌باشد؟ پاسخ ساده است. هر ژن در هر سلول رونویسی و بیان نمی‌شود. بدین معنی که تنها آن پروتئین‌هایی که مورد نیاز هستند در سلول در یک زمان معین در طول عمر سلول وجود دارند. بنابراین پروتئوم[۲] یک سلول یا بافت وابسته به نوع سلول و وضعیت سلول است.

در اصل، نظم پایه ژن (ژنوتیپ) برای تغییر بیان ژن و در نتیجه تغییرات فنوتیپ میبایستی از طریق جهش تغییر داده شود. اما در دهه‌های اخیر نشانداده شده است که عوامل محیطی می‌توانند با کمک تغییر بدون سازگاری توالی نوکلئوتیدی ژن‌ها، بیان ژن روی فنوتیپ تأثیر گذار باشند که این سازگاری بیان ژن اپی‌ژنتیک[۳] نام دارد و یک نقش اساسی در فعال‌سازی و غیرفعال‌سازی ژن‌ها ایفا می‌کند. ویژگی این تغییر اپی‌ژنتیک متأثر از عوامل محیطی مانند تغذیه و تنش می‌باشد. DNAی هسته‌ای به شکل آزاد نبوده بلکه به صورت کروماتین[۴] نمایان بوده که به عنوان واحدهای ساختمانی کروموزوم معرفی می‌شود (Allis and Jenuwein 2016). واحد اصلی تکراری کروماتین نوکلئوزوم[۵] است جایی که اطراف DNA بوسیله هشت پروتئین هیستونی[۶] پیچیده شده است. بر اساس فشردگی بسته‌بندی و تجمع انفرادی نوکلئوزوم، یک ژن قادر است فعال یا غیرفعال شود. در حالت فعال یوکروماتین[۷] و در

---

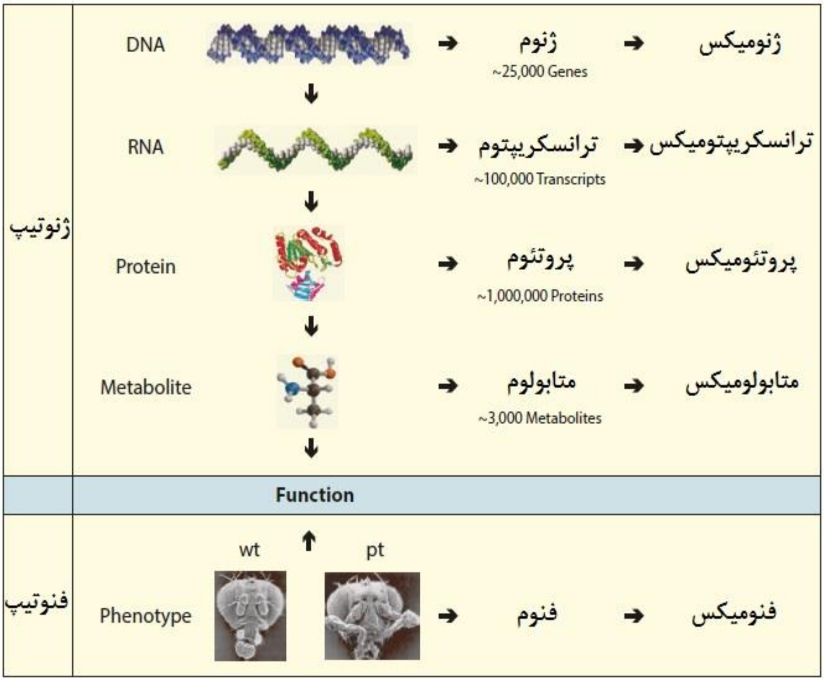
[1] Human Genome Project
[2] Proteome
[3] Epigenetic
[4] Chromatin
[5] Nucleosome
[6] Histone Proteins
[7] Euchromatin

حالت غیرفعال هتروکروماتین[1] نامیده می‌شود. این وضعیت فعال‌سازی می‌تواند متأثر از تغییر زنجیره‌های تک هیستونی باشد. به عنوان مثال استیلاسیون[2] لیزین زنجیره‌های جانبی اجازه تعامل و ارتباط با پروتئین‌های حاوی برومودامین[3] را می‌دهد. اتصال این پروتئین‌ها اجازه دسترسی نوکلئوزوم را داده و فعالیت رونویسی را افزایش می‌دهد. درحال حاضر، طیف گسترده‌ای از تغییرات و ترکیبات ممکن شناخته شده است، بطوری‌که از آن به عنوان کد هیستون[4] نام برده می‌شود.

همچنین فقط درک ژنوم و ژن‌ها کافی نیست تا توضیح دهد یک ژن چگونه است یا یک سلول یا موجود چگونه کار می‌کند. برای درک سیتم پیچیده بیولوژیکی بایستی تنظیم و بیان ژن‌ها، میزان کمی متابولیت‌ها و اثرات نقص ژنی بر روی فنوتیپ موجودات مطالعه شود. علاوه بر شناخت ژن‌ها بایستی عملکرد محصولات ژن هم شناخته شود. مطالعه این پیچیدگی اغلب سیستم زیستی[5] نامیده می‌شود که برای درک کامل موجودات زنده و پویایی موجودات در کل سیستم بیولوژیکی تلاش می‌کند. هدف به دست آوردن تصویر تلفیقی از همه فرایندهای منظم در همه سطوح از ژنوم تا پروتئوم و متابولوم[6] و از رفتار پروتئین تکی (منفرد) تا اندامک و بیومکانیک موجود کامل می‌باشد. روش‌های نوین برای تجزیه و تحلیل عملکردی ژنوم (ژنومیکس عملکردی)[7] ترانسکریتومیکس[8]، پروتئومیکس[9] و متابولومیکس[10] عنوان شده است(شکل ۱-۶).این‌ها معمولا روش‌های پرهزینه و پرمتقاضی در بحث مدیریت و تجزیه و تحلیل داده‌ها می‌باشند. این روش‌ها با تجزیه و تحلیل فنوتیپ‌های موجودات مدل و سلول‌ها در شرایط درون شیشه (این‌ویترو)[11]، همچنین در یک فرمت با توان بالا می‌باشد. در مطالعه فنوم، همه فنوتیپ‌ها را توصیف نموده و و تجزی و تحلیل آن فنومیکس[12] نامیده می‌شود.

---

[1] Heterochromatin
[2] Acetylation
[3] Bromodomain Containing Proteins
[4] Histone Code
[5] Biological System
[6] Metabolome
[7] Functional Genomics
[8] Transcriptomics
[9] Proteomics
[10] Metabolomics
[11] in vitro
[12] Phenomics



شکل ۱-۶) همبستگی میان ژنوتیپ و فنوتیپ. از ژنوم تا ترانسکریپتوم، پروتئوم و متابولوم تا فنوم. مثال در بخش ژنوتیپ از گونه انسان خردمند (*Homo sapiens*) و مثال بخش فنوتیپ از مگس سرکه (*Drosophila melanogaster*) استفاده شده است

٦-١-١ ترانسکریپتومیکس

متاسفانه نقش بسیاری از پروتئین‌ها بر اساس توالی نوکلئوتیدها به تنهایی ناشناخته است. هرچند اطلاعات مربوط به تنظیم و بیان ژن می‌تواند بینشی به عملکرد ژن در سلول‌ها، بافت‌ها و موجودات ارائه دهد. برای مثال وقتی یک ژن مثلا در سلول‌های عضلانی بیان می‌شود استنباط می‌شود که محصول ژن احتمالا برای فیزیولوژی این نوع سلول مهم است. بسیاری از تکنیک‌ها مانند نورترن بلات[1] برای تجزیه و تحلیل تنظیم و بیان ژن وجود دارند، که یک روشی برای تعیین بهره‌وری هیبریداسیون mRNAها در ژل آگارز یا واکنش زنجیره‌ای پلیمراز

---

[1] Northern Blot

معکوس (RT-PCR)[۱]، یک تکنیک برای تکثیر توالی نوکلئوتیدی اختصاصی مشتق شده از mRNAها است. این روش‌ها هر چند اجازه تجزی و تحلیل همزمان فقط تعداد کمی ژن را می‌دهد و برای تجزی و تحلیل مقدار زیادی داده مناسب نیستند. بنابراین ضروری است فرایندها با راندمان بالا که تجزیه و تحلیل با زمان مناسب برای تمام ترانسکریپتوم‌ها را فراهم می‌کند توسعه دهیم.

۱-۱-۱-٦ ریزآرایه DNA

یک مثال از روش‌های با کارایی بالا ریزآرایه DNA[۲] است که برای تعیین بیان ژن سلولی مناسب است زیرا قادر است پروفایلی از هر سلول بر اساس بیان ژن بوجود آورد. این روش همچنین به عنوان پروفایل بیان ژن معرفی شده است. ماده محافظ جامد ریزآرایه می‌تواند شامل یک صفحه شیشه‌ای با چندین هزار لکه[۳] اسیدنوکلئیک در کنار هم بر روی آن باشد (شکل ۲-۶). علاوه بر آن مواد دیگری مانند غشاء نایلونی[۴]نیز می‌تواند به عنوان محافظ استفاده شود. هر لکه DNA شامل نسخه‌های بسیاری از یک DNA منحصر به فرد تک‌رشته‌ای[۵] است که به یک ژن اختصاصی مربوط می‌شود (Holloway *et al*. 2002).

بسیاری از تکنیک‌ها برای تولید ریزآرایه DNA استفاده شده است.در اصل یک تمایزی میان آرایه الیگونوکلئوتیدی[۶] و ریزآرایه cDNA وجود دارد. در آرایه الیگونوکلئوتیدی توالی کوتاه با طول۲۰تا ۵۰ نوکلئوتید بطور مستقیم روی سطح مواد محافظ سنتز می‌شوند (شکل ۶-۲).این فرآیند شامل فتولیتوگرافی[۷] است که برای محصولات نیمه‌رسانا توسعه یافته است و هم اکنون نیز در صنعت کامپیوتر استفاده می‌گردد. اسلاید شیشه‌ای بوسیله اتصال دهنده‌هایی[۸]پوشیده شده است تا امکان تشکیل پیوندکووالانسی با نوکلئوتیدها فراهم شود. اتصال‌دهنده‌ها با یک گروه نشاندار[۹]محافظت می‌شوند تا از اتصال غیراختصاصی نوکلئوتیدها

---

[1] Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)
[2] DNA microarray
[3] Spot
[4] Nylon Membranes
[5] Unique single-stranded DNA
[6] Oligonucleotide arrays
[7] Photolithography
[8] Linkers
[9] Photolabile

جلوگیری کنند. با بکارگیری انتخابی گروه ضدنشاندار، گروه نشاندار حذف می‌شود و در نتیجه بطور اختصاصی مجموعه آرایه‌های فعال انتخاب می‌شود. سپس سطح آرایه با یک محلول نوکلئوتیدی که حاوی فقط یک نوکلئوتید اختصاصی مانند dATP است آغشته می‌گردد. در موقعیت‌هایی که توسط گروه ضدنشاندار، فعال‌سازی انجام می‌شود، نوکلئوتید می‌تواند به طور مستقیم به اتصال‌دهنده مواد محافظ متصل شود. نوکلئوتیدها خود را نیز در انتهای ′۵ با یک محافظ نشاندار محافظت می‌کنند و این دوباره باید فعال‌سازی گردد قبل از آن به دنبا آن واکنشی رخ دهد. بنابراین با بکارگیری تکرارهای چندتایی ضدنشاندار می‌توان یک آرایه الیگونوکلئوتید تولید کرد. این تکنیک قادر به تولید بسته فشرده‌ای از ریزآرایه‌ها با بیش از ۲۵۰۰۰۰ لکه الیگونوکلئوتید در هر سانتی‌متر مربع می‌باشد. در سال ۱۹۴۴، آفیمتریکس[1] نخستین شرکت بود که چیپ‌های DNA تجاری (آفیمتریکس) موجود را معرفی کرد.

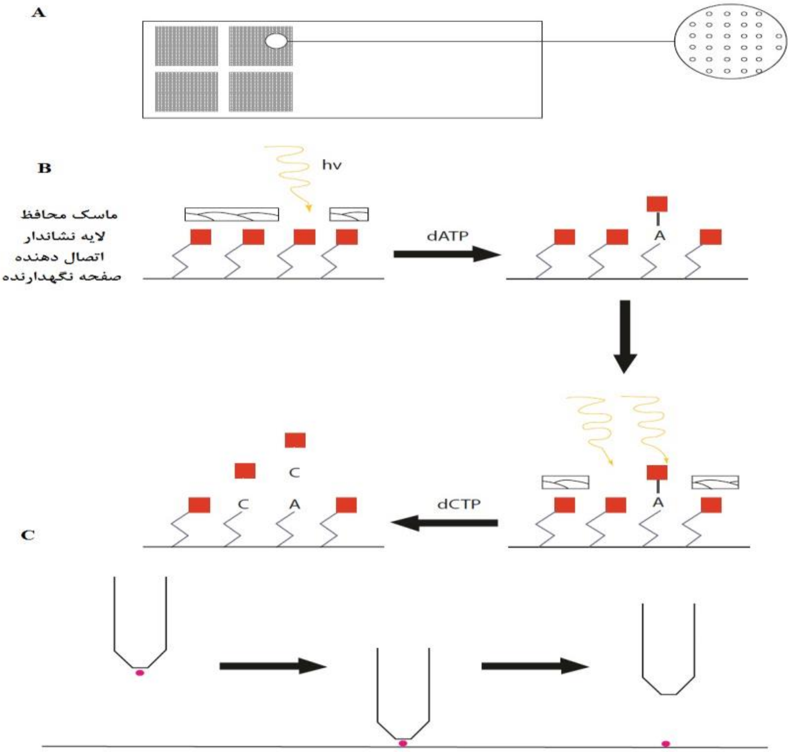
در مقابل، برای آرایه‌های cDNA، کاوشگرهای cDNA طولانی‌تر بر روی صفحه محافظ قرار می‌گیرند (شکل ۲-۶). در این روش ابتدا cDNA با طولی حدود چند صد نوکلئوتید به وسیله PCR در آزمایشگاه تولید می‌شود. پس از آن در مقیاس کوچک بکار گرفته می‌شوند همانگونه که لکه‌های DNA با استفاده از روبات بر روی صفحه محافظ آرایه قرار گرفته و بوسیله اشعه فرابنفش تثبیت می‌شوند. برخی از تولیدکنندگان از روبات‌های ردیاب با روش‌های مختلف استفاده می‌کنند. یک روش ریز لکه[2] است، به طوریکه محصولات PCR بطور مستقیم با روش مویینگی بر روی صفحه محافظ آرایه قرار می‌گیرد. روش دیگر ریز اسپره[3]کردن است، به این ترتیب که محلول cDNA اسپری می‌شود. شبیه یک چاپگر جوهرافشان است اما بدون نازل لمسی محافظ آرایه. با تراکمی بیش از ۲۵۰۰ لکه DNA در هر سانتی‌مترمربع می‌توان آرایه‌های cDNA نیز حاصل گردد.

تکنولوژی آرایه cDNA در بسیاری از آزمایشگاه‌های تحقیقاتی به دلیل مقرون به صرفه بودن محبوب واقع شده است. همچنین یک انعطاف‌پذیری در انتخاب نوع مواد (موجود زنده، بافت‌ها یا سلول‌ها) نیز وجود دارد. نوع دیگری از ریزآرایه‌ها یک آرایه الیگونوکلئوتیدی است که توسط تراکم شدید از نقاط با کیفیت بالا مشخص می‌شود. به دلیل تراکم بالا، برای هر ژن، چندین الیگونوکلئوتید می‌توانند بر روی آرایه قرار گیرند و اجازه کنترل نتایج و افزایش دقت

---

[1] Affymetrix
[2] Microspotting
[3] Microspraying

این آرایه‌ها را می‌دهد. از معایب این فناوری این می‌باشد که آرایه‌ها در خانه تولید نمی‌شوند و می‌بایستی از کارخانه خریداری شوند که اغلب بطور قابل ملاحظه‌ای گران هستند. علاوه بر این بستگی دارد که آرایه‌هایی که توسط شرکت سازنده ارائه می‌شوند انحصاری نشده باشند.



شکل ۲-۶) ریزآرایه DNA. A) ریزآرایه DNA که شامل چندین هزار لکه نوکلئیک‌اسید و آرایه‌هایی با تراکم بالا است. B) تصویربرداری از تولید یک آرایه الیگونوکلئوتیدی با استفاده از فوتولیتوگرافی. C) محلول‌های cDNA نقاطی هستند در تولید آرایه بر روی صفحات ریزآرایه به کمک روبات‌ها

- اجرای یک آزمایش پروفایل بیانی با cDNA

در بسیاری از مطالعات پروفیل بیان، الگوی بیان ژن در دو نوع جمعیت سلولی مثلا سلول‌های سالم (سلول‌های نوع A) با سلول‌های توموری (سلول‌های نوع B) مقایسه می‌کند

(شکل ۳-۶). مرحله نخست جداسازی RNAکل[1] از هر دو جمعیت سلول است. mRNA به کمک آنزیم ترانسکریپتاز معکوس رونویسی می‌شود و همزمان بوسیله افزودن نوکلئوتیدهایی که با رنگ‌های متفاوت فلورسنت متصل شده است، نشان‌گذاری می‌شوند. معمولا cDNA کنترل (سلول‌های سالم) با رنگ سیستئین۳ و cDNA نمونه (سلول‌های توموری) با رنگ سیستئین۵ نشان‌گذاری می‌شوند. سیستئین۳ و سیستئین۵ به ترتیب نور را در طیف سبز و قرمز نشر می‌دهند. این روش به عنوان روش نشان‌گذاری مستقیم[2] معرفی می‌شود. در مقابل، روش‌های نشان‌گذاری غیرمستقیم تنها از مقدار کمی از مواد اولیه که در دسترس هست استفاده می‌کنند. در این مورد، نوکلئوتیدهای تغییریافته در طی سنتز cDNA با رنگ‌های خاصی با میل ترکیبی زیاد، ترکیب و استفاده می‌شوند.

مخازن cDNA نشاندار، مخلوط و واسرشت[3] شده و رشته‌های cDNA تک‌رشته سپس با ریزآرایه DNA انکوبه می‌شوند. DNA در درون مخازن به مولکول‌های تک‌رشته‌ای DNA مکمل، که آرایه را تشکیل می‌دهند، هیبرید می‌شود. فعال‌سازی لیزری ریزآرایه در فرکانس رنگی بوسیله اسکن کمی مقدار نور ساطع شده، میزان اتصال DNA را اندازه‌گیری می‌کند. در نتیجه دو تصویر یکی به رنگ سبز و یکی در محدوده طول موج قرمز حاصل می‌شود. اگر هر دو باهم قرار بگیرند، یک تصویر ادغام شده با نقاط رنگی ایجاد می‌شود (شکل ۳-۶).

اگر ژن‌ها به صورت متفاوتی بیان شده باشند، به عنوان مثال، در یک جمعیت سلولی، میزان بیشتری از یک mRNA اختصاصی وجود داشته باشد، لکه‌ها قرمز یا سبز ظاهر می‌شوند. اگر cDNA نشاندار شده با سیستئین ۵ متصل باشد، لکه‌های قرمز ظاهر می‌شود به عنوان مثال افزایش بیان این ژن‌ها در سلول‌های سرطانی با کنترل مقایسه شده است. بر عکس، اگر ژن‌ها در سلول‌های سرطانی نسبت به سلول‌های کنترل کمتر بیان شوند، لکه‌ها فلورسنت سبز خواهند داشت. اگر cDNA فلورسنت قرمز و سبز به میزان مساوی با DNA نشاندار شده هیبرید شوند، لکه‌ها زرد رنگ به نظر خواهند رسید. این بدین معنی است که ژن‌های مربوطه در سلول‌های سرطانی و کنترل (شاهد) در میزان یکسانی بیان شده‌اند. لکه‌هایی که با هیچ کدام از cDNAها مکمل نیستند در مخزن، با رنگ مشکی به چشم می‌خورند. بنابراین واضح است که بیان نسبی ژن بین دو نمونه به یک میزان می‌باشد. تعیین مقدار مطلق برای آرایه‌های

---

[1] total RNA

[2] Direct Labeling

[3] Denatured

cDNA امکان‌پذیر نمی‌باشد. اما آرایه‌های الیگونوکلئوتیدی متفاوت عمل می‌کنند تا اجازه کمی‌سازی (اندازه‌گیری) مطلق را بدهند.

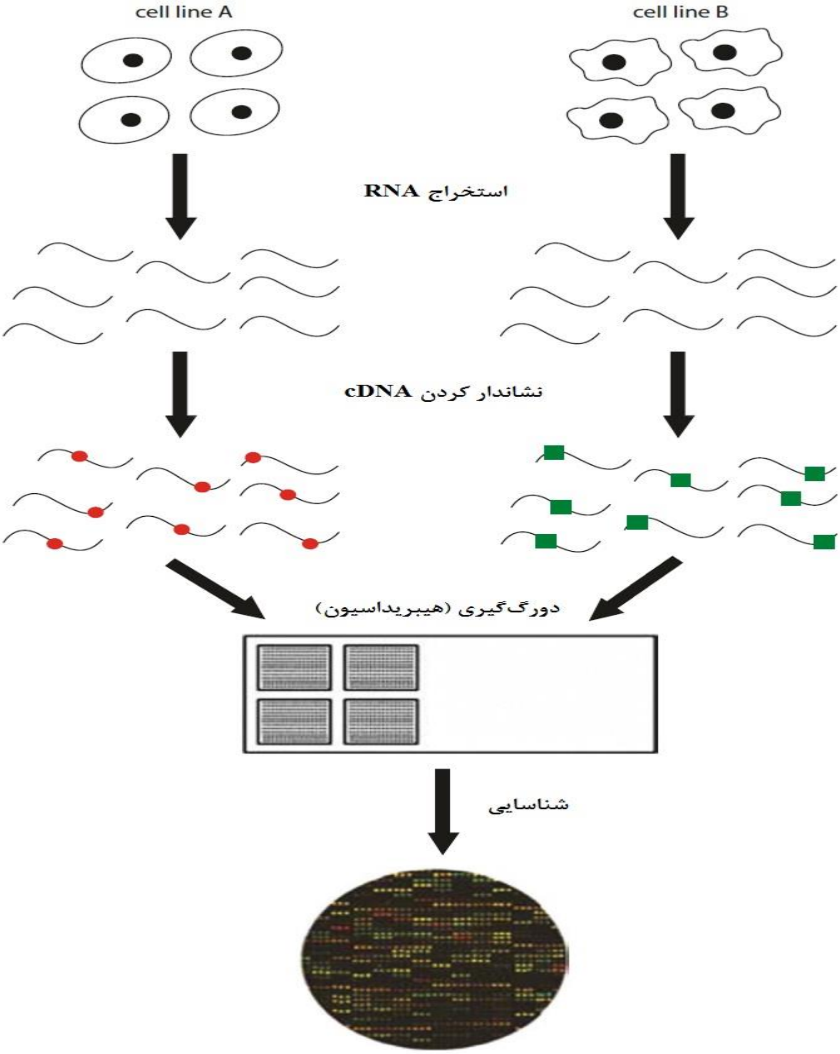
- تفسیر آزمایش پروفایل بیان ژن

اگرچه ایده‌های مربوط به ریزآرایه ساده است، ولی کاربرد و تجزیه و تحلیل نتایج آن‌ها پیچیده می‌باشد. این امر از منابع مختلف خطا ناشی می‌شود که شامل خطاهای آماری، که براساس نوسانات تصادفی است و نمی‌توانند تحت تأثیر قرار گیرند و خطاهای سیستماتیک، که منجر به اندازه‌گیری انحرافات می‌شود. چنین اشتباهات سیستماتیک می‌تواند ناشی از تنظیمات نادرست ابزار و یا تغییر شرایط محیطی (مانند نوسانات دمایی یا رطوبتی) در طول کار باشد.

خطاها را می‌توان با طراحی آزمایشات تجربی مناسب به حداقل رساند. همچنین خطاهای آماری را نیز با تکرار آزمایشات می‌توان کمتر نمود. نمونه‌ها باید هر روز بصورت تازه آماده شوند تا اطمینان حاصل شود که آزمایشات به صورت مستقل انجام می‌شود. خطاهای سیستماتیک را می‌توان با استفاده از طراحی یک آزمایش تجربی و آزمایشات کنترل (شاهد) پیشرفته، به حداقل رساند. یک نمونه از آزمایشات کنترل، روش تعویض رنگ[1] است. که در آن cDNAها با یک رنگ متفاوت از رنگی که در آزمایشات اصلی استفاده شده است، نشان‌گذاری می‌شود. به طورخاص، اگر در آزمایش اصلی، cDNA سلول‌های سرطان و سلول‌های کنترل (شاهد) به ترتیب با سیستئین ۵ و سیستئین ۳ نشان‌گذاری شود، در آزمایش کنترل به روش تعویض رنگ، cDNA سلول‌های سرطانی با سیستئین ۳ و سلول‌های کنترل با سیستئین ۵ نشان‌گذاری می‌شوند. به دلیل اینکه آماده‌سازی cDNA برای هر دو نوع آزمایش کنترل تعویض رنگ و آزمایش اصلی، یکسان است و فقط برچسب‌ها متفاوت‌اند، نتایج مشابهی می‌بایست به دست آید. با استفاده از آزمایشات کنترل تعویض رنگ می‌تواند خطایی که در طول نشان‌گذاری نمونه‌ها رخ داده است مورد بررسی قرار گیرد، که اگر چنین باشد میزان خطا نیز در تجزیه و تحلیل نتایج مورد بررسی قرار خواهد گرفت (Churchill 2002).

---

[1] Dye Swapping



شکل ۳-۶) مقایسه میزان بیان ژن در دو لاین سلولی متفاوت به عنوان بخشی از آزمایش پروفایل بیانی با استفاده از ریزآرایه cDNA

تفسیر داده‌ها با تجزیه و تحلیل شکل‌هایی که بوسیله اسکنر ریزآرایه ایجاد شده شروع می‌شود. شدت هر لکه می‌بایستی برای تبدیل به مقدار عددی (کمی) اندازه‌گیری شود. این مرحله پیچیده و دشوار است. بیش از هزار لکه باید بطور یکنواخت شناسایی شود. برای انجام این کار پارامترهای لکه‌ها و شدت فلورسنت در دو کانال بایستی اندازه‌گیری شود و هر دو با پس‌زمینه مقایسه شود. لکه‌های غیرمعمول که دارای شکل‌های نامنظم و یا حاوی توده‌های قرمز و سبز رنگ هستند را برای تجزیه و تحلیل بهتر باید نادیده گرفت. همه این فرآیندها معمولاً توسط نرم‌افزار اسکنر ریزپردازنده انجام می‌شود. با توجه به تعداد زیاد تولیدکنندگان ریزآرایه و فرآورده‌ها و پروتکل‌های مختلف و تنظیم پیچیده تجربی (تعداد زیاد مراحل انفرادی) تعجب‌آور نیست که اطلاعات ریزآرایه حاوی خطاهای سیستماتیک باشند. مثال‌های بسیاری از توزیع ناهموار هیبریداسیون در یک آرایه وجو دارد که منجر به رنگ‌آمیزی ناهمگن در برخی مناطق آرایه می‌شود یا از رنگ‌ها با نیمه‌عمر متفاوت استفاده شود که می‌تواند حین اندازه‌گیری شدت لکه، منجر به اشتباهاتی شود. برای جبران چنین خطاهای سیستماتیکی، مقادیر پروفایل بیان باید نرمال شود. نرمال‌سازی بر اساس این فرضیه که اکثر ژن‌ها در نمونه‌ها به صورت متفاوتی بیان نمی‌شوند صورت می‌گیرد. نرمال‌سازی نه تنها نتایج را تعدیل می‌کند بلکه همچنین مقایسه آزمایشات را در روزها و آزمایشگاه‌های مختلف را تضمین می‌کند. چندین الگوریتم برای نرمال‌سازی وجود دارد و آن‌ها معایب و مزایایی متفاوتی دارند. انتخاب نوع الگوریتم به تجربه و ترجیح محقق بستگی دارد (Quackenbush 2001).

برای مدت‌ها موضوع بحث این بوده است که آیا پلت‌فرم‌های تولیدکنندگان که از همه قابل قیاس هستند، متفاوت می‌باشند. با وجود این همه نگرانی، چندین محقق نشان دادند که با راه‌اندازی آزمایشگاهی مناسب، ممکن است بررسی مقایسات امکان‌پذیر باشد. با این حال، استفاده از پروتکل‌های استاندارد و کنترل‌های کافی ضروری است (Ji and Davis 2006). برای نظارت بر کنترل کیفیت این محصولات، دو کنسرسیوم صنعت ریزآرایه و آژانس‌های ایالات متحده آمریکا به همراه اعضای گروه پژوهشی دانشگاه، تشکیل شده است. پروژه کنترل کیفیت ریزآرایه (maqc)[۱] کنترل‌های استاندارد را ایجاد می‌کند که هدف آن تسهیل مقایسه آزمایش‌های ریزآرایه می‌باشد. همچنین کنسرسیوم کنترل RNA خارجی (ercc)[۲] اهداف مشابهی را نیز دنبال می‌کند. ERCC کنترل‌های RNA خارجی را توسعه می‌دهد و RNA

---

[1] The MicroArray Quality Control Project (maqc)

[2] The External RNA Controls Consortium (ercc)

جدا شده آزمایشگاهی قبل از سنتز cDNAاضافه می‌شود. به این ترتیب، میزان نتایج یک آزمایش ریزآرایه که با معیارهایی تعریف شده است می‌تواند تأیید شود.

گام بعدی در تجزیه و تحلیل داده‌ها، شناسایی ژن‌هایی است که به طورقابل توجهی بین دو نمونه، متفاوت بیان می‌شوند. برای سادگی در ریزآرایه اولیه، فرض بر آن است که تمام آن ژن‌ها که در نمونه‌ها متفاوت بیان می‌شوند حداقل دوبار بیان متفاوتی داشته باشند. امروزه روش‌های آماری پیچیده‌تری برای شناسایی آن ژن‌هایی که تفاوت معنی‌داری در سطوح بیان دارند استفاده می‌شود. فواید استفاده از این روش‌ها، تفاوت‌اندک معنی‌داری در شناسایی ژن‌ها، در سطوح مختلف بیان دارند. بعد از این تجزیه و تحلیل آماری تعداد نهایی ژن‌هایی که به طور متفاوت بیان می‌شوند، به دست می‌آید. مهمتر از همه این‌ها، نتایج می‌بایستی بوسیله روش‌های مستقل مانند تجزیه و تحلیل نوترن بلات تایید شوند (2002 Slonim).

تعیین بیان متفاوت ژن‌های انفرادی تنها جنبه جالب ریزآرایه نیست بلکه شناسایی الگوهای پروفایل بیان ژن نیز از مزایای دیگر آن بشمار می‌رود که جالب توجه است.ایده این است که ژن‌هایی که متعلق به یک مسیر(Pathway) هستند و یا به یک سری محرک‌های زیست‌محیطی واکنش نشان می‌دهند نیز بررسی شده و درنتیجه یک پروفایل بیانی مشابهی نمایش می‌دهند. با استفاده از تجزیه و تحلیل خوشه‌ای، ژن‌هایی با پروفایل‌های بیانی مشابه می‌توانند به گروه‌ها یا خوشه‌ها ترکیب شوند (شکل ۴-۶). چنین تحلیلی را برای ۱۶۴ ژن باکتریایی که به ۱۳ خوشه تقسیم می‌شوند نشان می‌دهد. تجزیه و تحلیل‌های خوشه‌ای بینشی ارزشمند در بررسی عملکرد پروتئین‌ها فراهم می‌آورد. اگر ژن‌هایی که محصولات‌شان نقش و عملکردی مشخص نداشته باشند به ژن‌ها با خوشه‌های که اخیراً شناخته شده‌اند، گروه‌بندی می‌شوند. سپس می‌توان یک عملکرد مشابه یا یک مسیر بیولوژیک مشترک را برای ژن ناشناخته، نشان داد. همچنین پروتئین‌های ناشناخته نیز می‌توانند به طورخاص، مورد بررسی قرار گیرند. هر آزمایش پروفایل بیان، مقدار زیادی داده را تولید می‌کند. یک آزمایش می‌تواند شامل ده‌ها ریزآرایه، که به نوبه خود شامل چندین هزار لکه است باشد. بنابراین، نتایج چند صدهزار و یا حتی میلیون‌ها اندازه‌گیری با استفاده از پایگاه داده‌های خاص که در آن داده‌ها می‌توانند در هر زمانی ذخیره و بازیابی شوند، باید مدیریت و تجزیه و تحلیل شود. به عنوان مثال پایگاه‌های داده بیان ژن در مرکز ملی اطلاعات بیوتکنولوژی (NCBI) (geo)[1] و بیان

---

[1] National Center for Biotechnology Information (NCBI) [geo]

آرایه در موسسه بیوانفورماتیک اروپا (EBI) (arrayexpress)[1] می‌باشند. علاوه بر این می‌توان داده‌های خام را بدون پردازش، تحت پروتکل‌ها و شرایط آزمایشی مناسب اجرا نمود. این داده‌ها باید حداقل اطلاعات مربوط به یک آزمایش ریزآرایه [miame] که در آن حداقل الزامات، برای تفسیر صریح و باز تولید قابل اعتماد آزمایش‌های ریزآرایه تعریف شده‌اند را برآورده نماید (Brazma *et al*. 2001).

به طورخلاصه، انجام آزمایش‌های ریزآرایه، اجزا بیوانفورماتیکی پیچیده را برای آزمایش‌کننده فراهم می‌آورد. خوشبختانه، انواع مختلفی از نرم‌افزارها وجود دارند که تجزیه و تحلیل داده‌ها را ساده می‌کنند. یک برنامه شناخته شده تجاری برای تجزیه و تحلیل داده‌های ریزآرایه، برنامه همردیف‌سازی GeneSpring GX می‌باشد. بسته‌های نرم‌افزاری که اغلب استفاده می‌شوند و در آکادمی محیط‌زیست توسعه یافته‌اند شامل: هدایت‌گر زیستی (bioconductor)، دنباله TM4 (tm4) و الگوهای ژنی(GenePattern) می‌باشند.

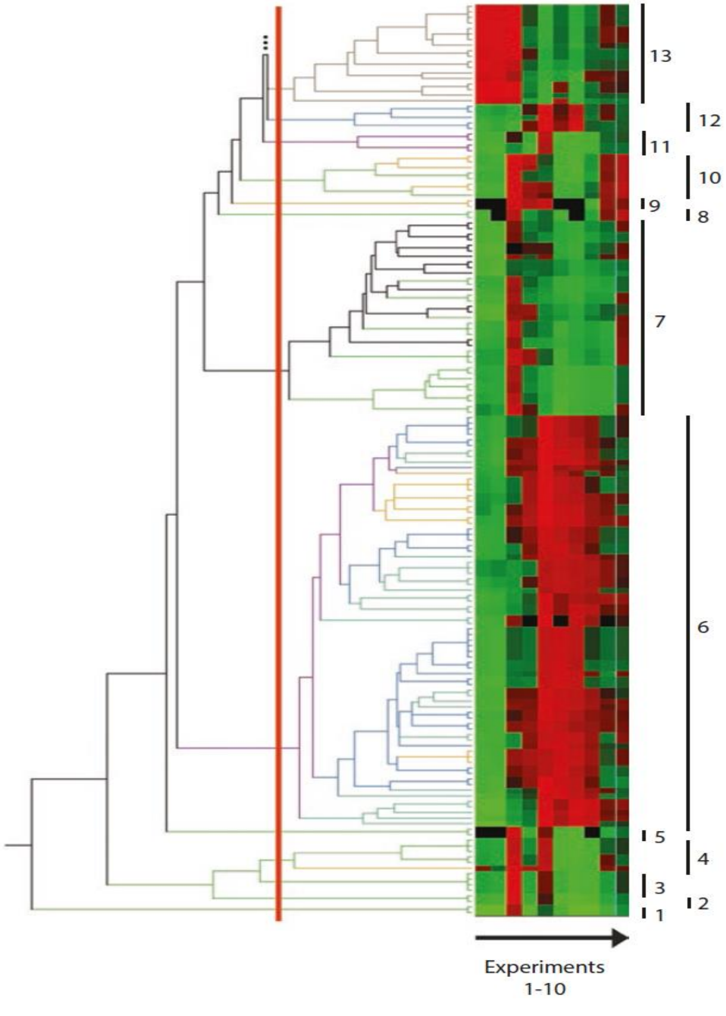
علاوه‌بر پروفایل بیان، انواع برنامه‌های دیگر برای ریزآرایه وجود دارد که اهمیت بیشتری از آن حاصل می‌شود به عنوان مثال: بخش بررسی تومور در علم پزشکی (Gershon 2005). درمان بهینه یک بیمار مبتلا به سرطان به شدت وابسته به یک تشخیص دقیق است که در حال حاضر اساس آن ترکیبی از داده‌های بالینی و هیستوپاتولوژیک می‌باشد. با این حال، در برخی موارد، تشخیص دقیق دشوار است زیرا تومورها اغلب خواص غیرمعمولی دارند. در چنین مواردی، ریزآرایه می‌تواند به طبقه‌بندی تومورها با توجه به پروفایل بیان ژن آن‌ها کمک کند. لوسمی حاد[2] یک مثال از بیماری سرطان است. این سرطان لکوسیت می‌تواند با استفاده از داده‌های بالینی و مورفولوژیکی برای تشخیص به دو زیر گروه لوسمی لمفوبلاستی حاد (ALL)[3] و لوسمی مایلوئید حاد (AML)[4] تقسیم شود. تمایز این انواع زیرگروه ضروری است، چراکه هرکدام با تیمارهای مختلف شیمیایی درمان می‌شوند. یک تحقیق اولیه این است که آیا با تشخیص مولکولی می‌توان نتایج قابل اعتمادی به دست آورد که مورد آزمون قرار گیرد و به مقایسه ریزآرایه DNA با روش‌های کلاسیک کمک کند (Golub *et al*. 1999). پروفایل بیان ژن بیماران با روش تشخیص مولکولی مورد تجزیه و تحلیل قرار گرفت و نتایج با روش غیر

---

[1] European Bioinformatics Institute (EBI) [arrayexpress]
[2] Acute Leukemia
[3] Acute Lymphoblastic Leukemia (ALL)
[4] Acute Myeloid Leukemia (AML)

مولکولی مقایسه شدند. نتیجه نشان داد که ابزار تشخیصی ریزآرایه قابل اعتماد است. علاوه‌براین، بیمار مبتلا به لوسمی حاد غیرطبیعی، نیز مورد بررسی قرار گرفت.

در این آزمایش ابزارتشخیصی ریزآرایه نشان داد که پروفایل بیان ژن این بیمار بطور کامل از دیگر بیماران متفاوت است. مشخصات پروفایل آن به سرطان بافت عضلانی بیش از لوسمی حاد منطبق می‌باشد. همچنین بررسی‌های سیتوژنتیک نیز با تشخیص لوسمی حاد مغایر و موافق با تومور عضلانی است. با این نتایج تشخیص نهایی و شیمی‌درمانی تغییر یافت. بنابراین طبقه‌بندی تومورها بر اساس ریزآرایه DNA با تکنیک‌های تشخیصی استاندارد پشتوانه معتبرتری را فراهم می‌کند (Golub *et al*. 1999).



شکل ۴-۶) گروه‌بندی ژن‌ها با پروفایل بیانی مشابه. بیان ۵۶۲ ژن باکتریایی که در ۱۰ آزمایش مختلف اندازه‌گیری شده است. پروفایل بیان مقایسه گردید و ژن‌هایی با الگوی بیانی مشابه در یک گروه قرار گرفتند. در این شکل ۱۳ گروه با بیش از ۱۶۴ ژن نمایش داده شده است. به عنوان مثال گروه ۱۳شامل ۱۸ ژن است که بیان بالایی در۳ آزمایش نخست (قرمز) مشاهده می‌شود اما پس از آن بیان کاهش می‌یابد (سبز). نوار قرمز آستانه انتخاب شده برای تعریف یک گروه معرفی شده است.

یکی دیگر از زمینه‌های مهم برای استفاده از فناوری ریزآرایه سم‌شناسی است. تجزیه و تحلیل مواد سمی برای شناسایی عواقب مواد شیمیایی مضر در سلول‌ها طراحی شده است. به عنوان مثال، یک آنتی‌بیوتیک بالقوه جدید ممکن است نه تنها باکتری عفونی را از بین ببرد، بلکه به سلول‌های ارگان‌های بدن بیمار نیز آسیب برساند. بنابراین پتانسیل هر آنتی‌بیوتیک جدید برای خواص سمی خود با سموم موجود مقایسه می‌گردد. این مقایسه شامل پروفایل بیان ژن در ریزآرایه DNA می‌باشد.اگر پروفایل بیان ژن بین ترکیبات جدید و سم مشخص با هم، همپوشانی داشت سپس ماده جدید به عنوان ماده سمی بالقوه گروه‌بندی می‌شود. ویژگی‌های سمی با کمک ریزآرایه DNA به عنوان گروه ژنومی سمی[۱]، شناخته شده است.

۲-۱-۱-۶تجزیه و تحلیل سریالی بیان ژن

همانند فناوری ریزآرایه DNA، روش تجزیه و تحلیل سریالی بیان ژن (SAGE)[۲]، یک فناوری با توان بالا برای اندازه‌گیری بیان ژن است. روش SAGE مقایسه بیان ژن در سلول‌ها یا بافت‌های مختلف و در نتیجه شناخت ژن‌هایی با بیان متفاوت را تسهیل می‌کند. در روش SAGE به جداسازی RNA کل از سلول‌ها و بافت‌ها و تبدیل mRNA به cDNA با کمک آنزیم ترانسکریپتاز معکوس با منشأ ویروسی، نیاز دارد. با این حال، cDNAکلون نمی‌شود، بلکه به جای آن با برخی آنزیم‌های محدودکننده‌ای که DNA را در سایت‌های خاص برش می‌دهند، تیمار می‌شود. این امر منجر به تولید قطعات کوتاه از هر مخزن cDNA انفرادی با طول بین ۱۰ تا ۱۱ نوکلئوتید یعنی تگ (Tag) می‌شود. برچسب‌ها (تگ) به مولکول‌های ترتیبی بلند متصل می‌شوند و پس از آن درون پلاسمید برای توالی‌یابی کلون می‌گردند. در یک آزمایش SAGE، مقدار فراوانی یک تگ در نمونه به عنوان میزان اندازه‌گیری بیان mRNA استفاده می‌شود. به عنوان مثال، اگر ژن برچسب‌دار(تگ)، ۵ بار در یک نمونه از سلول‌های سالم، اما ۲۰ بار در یک نمونه از سلول‌های سرطانی پیدا شود، فرض بر این است که این ژن تقریباً چهار برابر در سلول‌های سرطانی افزایش بیان داشته است. نتایج SAGE را می‌توان در پایگاه داده بیان ژن (geo)[۳]ذخیره نمود. در این پایگاه، اطلاعات مربوط به هر تگ، مانند: توالی DNA، میزان

---

[1] Toxicogenomics

[2] Serial Analysis of Gene Expression (SAGE)

[3] Gene Expression Omnibus (geo)

فراوانی در بافت‌ها و یا سلول‌ها، و میزان رونویسی اختصاصی از آن تگ را می‌توان مشاهده نمود.

مزیت بزرگ SAGE نسبت به ریزآرایه این است که تمام رونوشت‌های mRNA سلول، می‌تواند مورد تجزیه و تحلیل قرار گیرد که رونوشت‌های ناشناخته را هم شامل می‌شود. در مورد ریزآرایه DNA، تنها رونوشت‌های mRNA با لکه‌های cDNA موجود در ریزآرایه، مورد تجزیه و تحلیل قرار می‌گیرد. از دیگر مزایای SAGE، قابلیت تکثیر آن‌ها بین آزمایشات است.یکی از معایب SAGE زمان زیاد مورد نیاز برای انجام آزمایشات با کارایی بالا می‌باشد. در مقابل، ریزآرایه DNA انعطاف‌پذیری بالاتری نشان می‌دهد و استفاده از آن‌ها برای تجزیه و تحلیل ژن‌های کل ژنوم در یک آزمایش مجاز می‌باشند. SuperSAGE به عنوان یک روش تکمیلی است برای جبران خطاهایی که در آن آنزیم‌های محدودکننده تولید تگ‌های بزرگتری می‌نماید، استفاده می‌شود (Matsumura *et al.* 2006). امروزه میلیون‌ها تگ قادرند در ترکیب با روش توالی‌یابی نسل جدید (NGS)[1] مورد تجزیه و تحلیل قرار گیرند.

۲-۱-٦ پروتئومیکس

اندازه‌گیری میزان mRNA به کمک ریزآرایه DNA یا SAGE، اطلاعات مهمی در مورد پتانسیل عملکرد سلولی محصولات ژن فراهم می‌کند. اندازه‌گیری میزان mRNA به تنهایی، به طور کامل و دقیق کافی نیست تا بتواند سیستم‌های بیولوژیکی پیچیده را توصیف کند. در نتیجه فعالیت‌های سلولی، مانند: فرایندهای بیولوژیکی توسط پروتئین‌های پروتئوم و نه به کمک ژن‌های ژنوم و یا رونوشت‌های mRNA، میانجی‌گری می‌شود. همگام با فناوری ریزآرایه DNA روش‌هایی با توان بالا برای تجزیه و تحلیل عملکرد پروتئین‌ها، مانند: پروتئومیکس، توسعه یافته است. پروتئومیکس به دو گروه تقسیم‌بندی می‌شود: پروتئومیکس کمی یا کلاسیک[2] و پروتئومیکس عملکردی[3]. پروتئومیکس کمی با شناسایی و اندازه‌گیری پروتئین‌های سلول سروکار دارد درحالی که هدف پروتئومیکس عملکردی، تعیین عملکرد پروتئین است. پروژه پروتئو مانسانی(HPP)[4] یک کنسرسیوم بین‌المللی با چند گروه تحقیقاتی

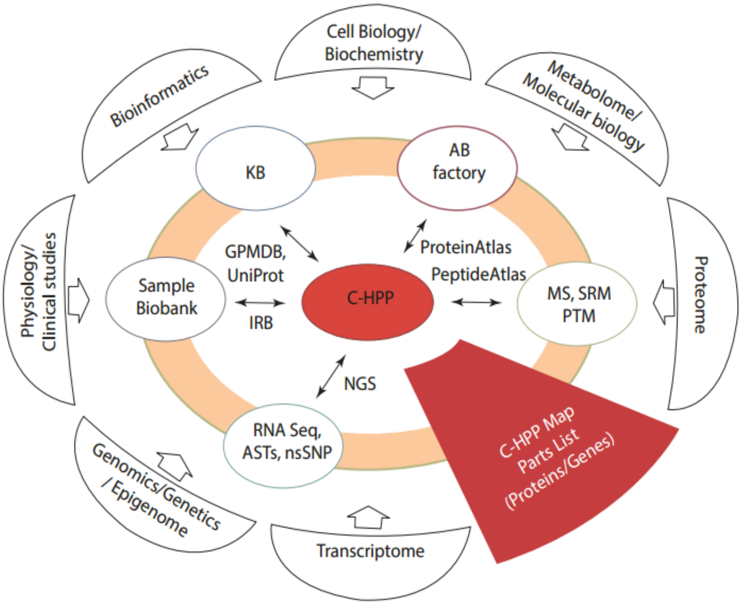
---

[1] Next-Generation Sequencing (NGS)
[2] Classical or Quantitative Proteomics
[3] Functional Proteomics
[4] Human Proteome Project (HPP)

است که قابل مقایسه با پروژه ژنوم انسان است. هدف از این پروژه، تجزیه و تحلیل سیستماتیک در پروتئوم انسان برای درک بهتر زیست‌شناسی انسان در سطح سلولی است. این امر منجربه بهبود کاربردهای دارویی می‌شود (مانند بهبود درمان و تشخیص بیماری). بخش مهمی از این پروژه با تجزیه و تحلیل پروتئوم مبتنی بر کروموزوم سرو کار دارد که عملکرد هر یک از ژن‌ها را درک و مورد تجزیه و تحلیل قرار دهد. همکاری گروه‌های تحقیقاتی مختلف در زمینه‌های ژنومیکس، ترانسکریپتومیکس، پروتئومیکس و متابولومیکس برای دستیابی به این هدف مورد نیاز است (شکل ۵-۶).

۱-۲-۱-۶ پروتئومیکس کلاسیک

پروتئومیکس کلاسیک مشابه پروفیل بیان است، به همین دلیل است که آن را نیز پروفیل پروتئین نامیدند. هر دو فناوری اجازه انگشت‌نگاری مولکولی یک سلول بر اساس ژن‌های بیان شده در سطح mRNA یا پروتئین را می‌دهد. که با مقایسه دو یا چند انگشت‌نگاری ژنی و پروتئینی با بیان متفاوت، می‌توانند شناسایی شوند. هر دو فناوری مزایا و معایبی دارند. پروفایل پروتئین در نهایت پروتئین‌هایی که نقش عملکردی در سلول دارند را تشخیص می‌دهد. همچنین، تغییرات کمی در ترکیب یک پروتئین بر اساس سنتز یا تجزیه آن قادر است اندازه‌گیری شود. سایر مزایای پروفایل پروتئین نیز قابلیت بررسی تغییرات پس از ترجمه (مانند فسفریلاسیون و گلیکوزیلاسیون) است و ترکیب پروتئین بخش‌های سلولی را تعیین می‌کند (مانند میتوکندری و هستک). یکی از معایب این است که تمام پروتئین‌ها محلول نیستند، به خصوص پروتئین‌های غشایی، و بنابراین نمی‌توان آن‌ها را تشخیص داد. محدودیت دوم، محدود بودن تشخیص پروتئین‌هایی با بیان ضعیف است که می‌توانند از دست رفته باشند. در مقابل ژنوم‌های کامل در آزمایشات ریزآرایه جدید می‌توانند مورد تجزیه و تحلیل قرار گیرند. اما هنوز این فرض وجود دارد که پروفایل بیان، که مقدار mRNA را به طور مناسبی نشان می‌دهد اغلب برای پروتئین غیرقابل توجیه است. علاوه‌براین، مقدار mRNA نمی‌تواند اطلاعات را درباره نیمه‌عمر پروتئین فراه مکند. بنابراین، در هر صورت، هر دو بیان ژن و پروفایل پروتئین باید به عنوان تکنیک‌های تکمیلی اجرا شوند.



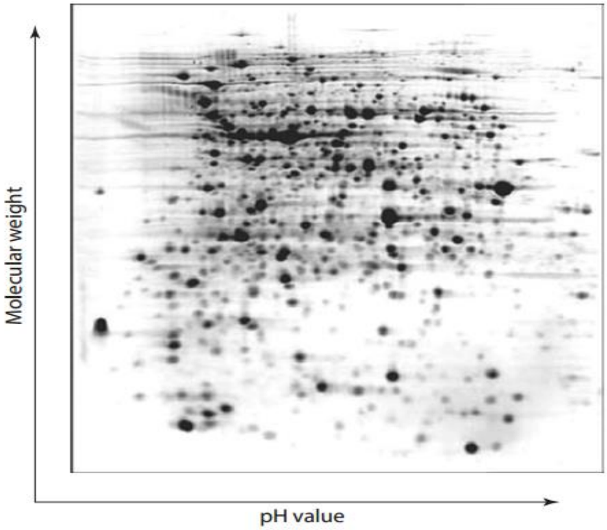
شکل ۵-۶) بخش کروموزوم مبتنی بر پروژه پروتئوم انسانی است (C-HPP)، MS اسپکتروسکوپی جرمی، AB آنتی بادی، KB بازهای شناخته شده (برگرفته از سایت http://www.c-hpp.org)

یک روش معمول برای پروفایل پروتئین شامل: الکتروفورز ژل دو بعدی با اسپکتروسکوپی جرمی است. در الکتروفورز دو بعدی پروتئین‌های سلولی ابتدا از طریق یک ماتریس (به عنوان مثال ژل پلی‌آکریل امید) بر اساس میزان حرکت خود در میدان الکتریکی جدا می‌شوند. بنابراین ممکن است بر اساس دو ویژگی ذاتی پروتئین یعنی بار و جرم جداسازی صورت پذیرد. این میزان از حرکت پروتئین به ترکیب اسیدهای‌آمینه آن بستگی دارد مثلا سیتوکروم C محتوی میزان بالایی آمینواسید بازی است و بنابراین بار مثبت در PH خنثی دارد. بار خالص پروتئین به PH محیط اطرافش بستگی دارد و PH که هم بار مثبت و هم بار منفی پروتئین آن با هم برابر هستند (به عنوان مثال بار خالص صفر) را نقطه ایزوالکتریک[1] می‌گویند.بر این اساس اگر پروتئین PHآن با PH محیط اطرافش برابر باشد در میدان الکتریکی حرکتی نخواهد کرد.از آنجایی که نقطه ایزوالکتریک از ویژگی‌های هر پروتئین است می‌توان مخلوطی از پروتئین‌ها را در شیبی از PH با استفاده از یک میدان الکتریکی از هم جدا نمود. این روش، ایزوالکتریک

---

[1] Isoelectric point (pI)

فوکوسینگ[1] نامیده می‌شود، در الکتروفورز دو بعدی به عنوان بعد نخست برای جداسازی پروتئین استفاده می‌شود. در بعد دوم پروتئین‌ها تنها بر اساس وزن مولکولی‌شان جداسازی می‌شوند. پپتیدها با وزن مولکولی پایین سریعتر از پروتئین‌های بزرگتر از منافذ ژل پلی‌آکریل امید عبور می‌کنند. در این روش بیش از ۱۰۰۰۰ پروتئین مختلف با وضوح بالا در ژل دو بعدی می‌توانند جداسازی شوند.پس از جداسازی پروتئین‌ها با استفاده از روش‌های مختلف رنگ‌آمیزی (مانند رنگ‌آمیزی با نقره یا رنگ‌های فلورسنت) می‌شوند (شکل ۶-۶). ژل‌ها سپس هضم و بار روش‌های بیوانفورماتیکی مورد تجزیه و تحلیل قرار می‌گیرند.



شکل ۶-۶) الکتروفورز ژل پلی‌آکریل امید دو بعدی (2D-PAGE): یک پروتئین در باکتری ابتدا در شیب طولانی PH (PH ۳ تا۱۰) در بعد اول و جرم مولکولی در بعد دوم جداسازی می‌شوند. سپس پروتئین‌های حل شده توسط نقره رنگ‌آمیزی می‌شود

برنامه‌هایی مانند ملانی[2] در سروراکسپسی (ExPasy) پروتئومیکس، اجازه تشخیص اتوماتیک و تعیین دقیق لکه‌های پروتئین را می‌دهند. علاوه بر این ملانی اجازه مقایسه چندین

---

[1] Isoelectric Focusing
[2] Melanie

ژل دو بعدی را نیز می‌دهد. لکه‌های پروتئینی هم مکان در ژل‌های مختلف شناسایی شده‌اند و تفاوت‌های کمی (عددی) آن‌ها بر اساس شدت نقطه، اندازه‌گیری شده‌اند. ملانی همچنین محتوی الگوریتم برای نرمال‌سازی و تجزیه و تحلیل آماری با نتایج معنی‌دار را می‌تواند داوری کند، تا پروتئین بیان شده متفاوت را شناسایی کند.

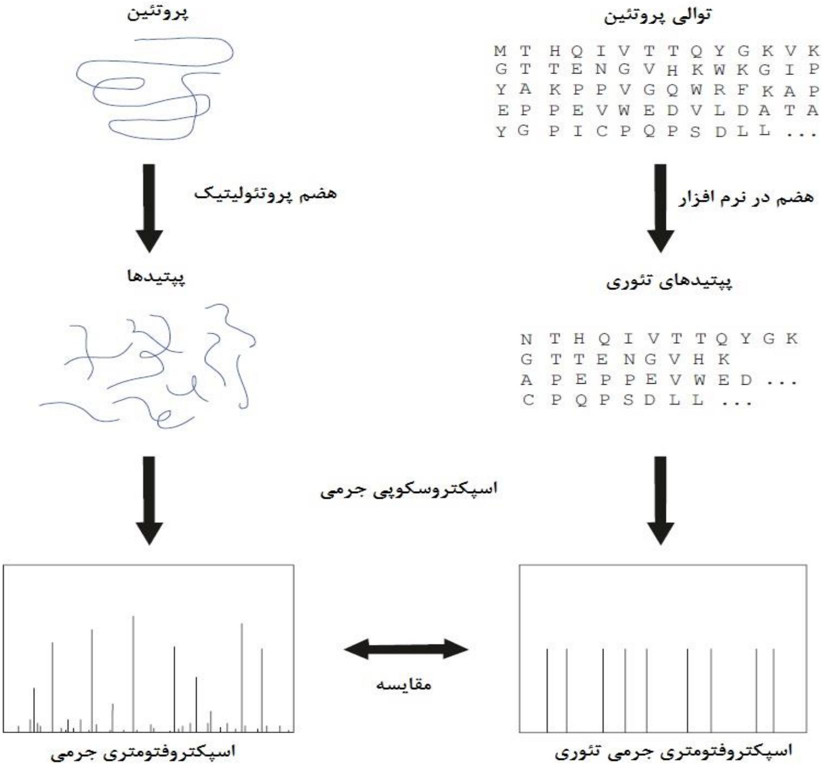
ارزیابی بیوانفورماتیک عملکرد ژل‌های دو بعدی لیستی از پروتئین‌ها بیان شده که تنها بر اساس نقاط ایزوالکتریک و توده‌های مولکولی شناخته شده‌اند. در حالی که هویت برخی از این پروتئین‌ها می‌توانند با استفاده از این اطلاعات برای اکثر پروتئین‌ها تعیین شوند، اما برای بیشتر پروتئین‌ها تعیین توالی آمینواسید ضروری است. این توالی با پایگاه داده پروتئین مقایسه می‌شود و اگر پروتئین در حال حاضر وجود داشته باشد، هویت آن می‌تواند تأیید شود. امروزه تکنیک‌های مختلفی برای تعیین توالی آمینواسید استفاده می‌شود. یک روش توالی‌یابی تجزیه ادمن[1] است که برای انجام آن مقدار زیادی پروتئین مورد نیاز است. یک روش پیشرفته آنالیز پروتئین اسپکتروسکوپی جرمی[2] بر اساس آنالیز پپتیدها از طریق جذب ماتریکس کمکی لیزر نسبت به یونیزاسیون در زمان پرواز یا اصطلاحاً (MALDI-TOF)[3] است. روش MALDI–TOF به اندازه کافی حساس و تنها به مقادیر پیکومول پروتئین نیاز دارد.لکه‌های پروتئین رنگ شده از ژل پلی‌آکریل امید خارج می‌شوند و با پروتئاز (مانند تریپسین) انکوبه می‌شوند که هر پروتئین را به الگوی پپتید اختصاصی هیدرولیز می‌نماید. پپتیدها از ژل استخراج شده و در یک MALDI–TOF مورد تجزیه و تحلیل قرار می‌گیرد. برای هر پپتید یک طیف جرم پپتیدهای اختصاصی تولید می‌شود (شکل ۷-۶) در عین حال، تمام پروتئین‌ها در یک پایگاه داده به پپتیدهای موجود در سیلیکا بر اساس تجزیه و شکست مشابه اختصاصی تریپسین هضم می‌شوند و طیف جرمی تئوری این قطعات حاصل از هضم محاسبه می‌شود. بطور آزمایشگاهی و تجربی طیف جرمی MALDI–TOF تعیین شده است و سپس با طیف جرمی تئوری مقایسه می‌شوند و نهایتاً طیف‌های جرمی یکسان انتخاب می‌شوند. از آنجا که طیف سنجی جرمی MALDI–TOF می‌تواندبرای بیش از یک پروتئین نتیجه دهد.

---

[1] Edman DegradationSequencing

[2] Mass Spectroscopy

[3] Matrix-Assisted Laser Desorption/Ionization–Time Of Flight (MALDI–TOF)



شکل ۷-۶) شناسایی پروتئین‌ها با دو منبع متفاوت از اسپکتروفتومتر جرمی در آزمایشگاه و اسپکتروفتومتر جرمی در نرم افزار

شناسایی قطعی پروتئین نیاز به طیف پپتیدهای مختلف دارد. بنابراین اگر چندین طیف جرمی که توسط MALDI تعریف شده و آن‌ها با آنچه که به لحاظ تئوری تعیین شده در توافق باشد، بنابراین پروتئینی که به لحاظ آزمایشی آنالیز شده مشابه آن پروتئینی که در پایگاه داده شناسایی شده است. تکنیک یونیزاسیون متناوب پروتئین، یونیزاسیون الکترواسپری است(ESI)[1]. ESI روشی حساس و به ویژه برای تجزیه و تحلیل ترکیبات مولکولی با وزن بالا مانند پروتئین‌ها مناسب می‌باشد. مزیت ESI بیش از MALDI است که می‌توان ESI را به

---

[1] Electrospray Ionization (ESI)

یک سیستم کروماتوگرافی مایع[1]متصل کرد. در آخر محلول‌های پروتئینی حداقل با پیچیدگی متعادل (مثلاً تعداد محدودی از پروتئین‌های مختلف) را می‌توان از هم جدا کرد و جایگزین الکتروفورز دو بعدی پر زحمت نمود. با اتصال مستقیم سیستم کروماتوگرافی مایع به اسپکتروسکوپی جرمی (LC/MS) شناسایی پروتئین تسریع پیدا می‌کند. از معایب ESI حساسیت بالای آن به آلاینده‌های قلیایی و آنچه را که تفکیک توده‌ها را دچار ابهام می‌کند. دیگر تحولات در زمینه طیف‌سنجی جرمی نیز رخداده است (Griffin *et al*. 2001).

در طیف سنجی جرمی دو طرفه (MS/MS) دو طیف‌سنج جرمی بطور پیوسته اجرا می‌شود که تا حدود زیادی حساسیت و انتخابی بودن سیست مرا بهبود می‌بخشد. به عنوان مثال نمونه‌های پروتئین با کمک ESI یونیزه می‌شوند. سپس در اولین طیف‌سنج، یون‌های جرم معلوم انتخاب و برای ایجاد قطعات بیشتر برانگیخته می‌شود و تجزیه و تحلیل جزئیات در طیف‌سنج دوم اجرا می‌شود. بنابراین جداسازی کروماتوگرافی اولیه ممکن است غیر ضروری باشد. در عمل هرچند این سیستم‌ها اغلب به عنوان بخشی از یک سیتم LC-MS/MS حتی سیستم LC-MS/MS دو بعدی جفت می‌شوند که حساسیت و انتخابی بودن آن‌ها افزایش می‌یابد.

۲-۱-۲-٦ پروتئومیکس عملکردی

هدف پروتئومیکس عملکردی، تشخیص عملکرد پروتئین‌هاست، مثلاً اثر متقابل پروتئین-پروتئین را شناسایی می‌کند. بسیاری از این فرایندهای سلولی تحت چنین اثرات متقابل کنترل و مدیریت می‌شود و شناسایی آن‌ها یک موضوع مهم برای درک عملکرد همه پروتئین‌ها می‌باشد. برای مثال باز داری آلوستریک آنزیم‌ها، تنظیم آبشارهای سیگنال انتقال[2]، مونتاژ ساختاری کمپلکس پروتئین تشکیل سیتواسکلت را می‌دهد. روش‌های متعددی اجازه تجزیه و تحلیل اثرات متقابل مانند کروماتوگرافی تمایلی[3] و سیستم هیبرید دوگانه مخمر[4] را می‌دهد. با این حال برنامه‌های کاربردی آن‌ها معمولاً با بررسی اثرات متقابل یک تعداد محدودی پروتئین تأیید می‌شود. در عین حال این روش‌ها به این حد از پیشرفت رسیده‌اند که از آن‌ها می‌توان در

---

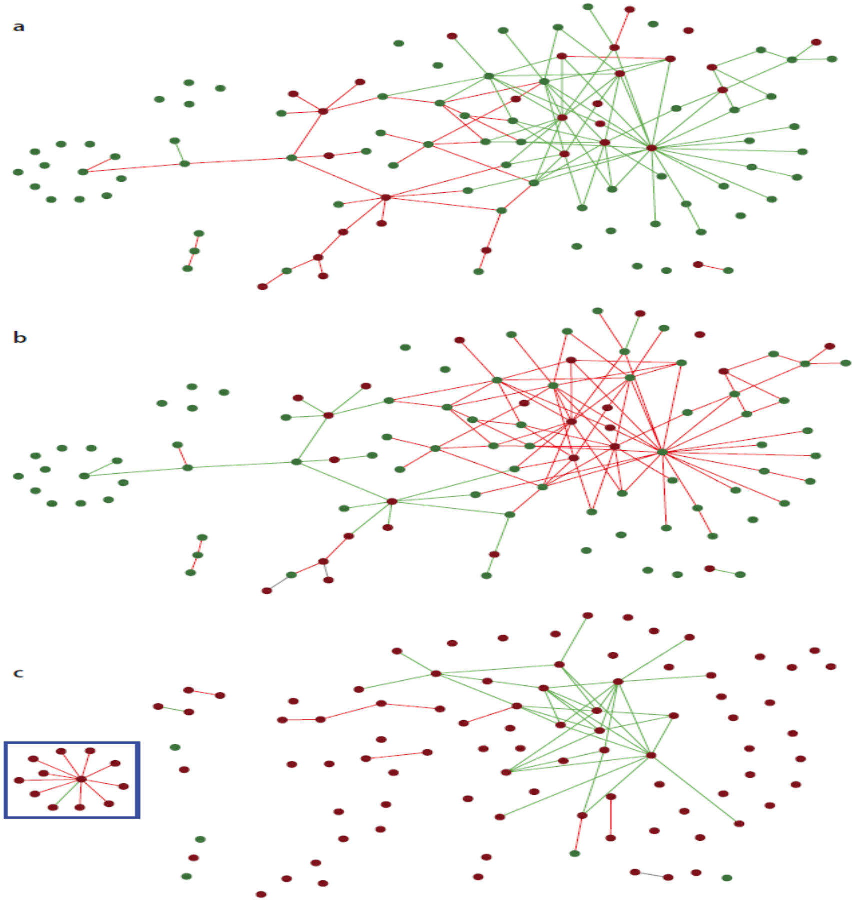
[1] LiquidChromatographic (LC)
[2] Signal Transduction Cascades
[3] Affinity Chromatography
[4] Yeast Two-Hybrid

تشریح اثرات متقابل پروتئین-پروتئین در پروتئوم‌های کامل استفاده نمود (شکل ۸-۶).در این زمینه اصطلاحات تعاملی یک موجود بکار می‌رود و این نوع از تحقیقات اینتراکتومیکس[1] نامیده می‌شود.اثرات متقابل امتزاج دو پروتئین تشخیص داده می‌شود، سیستم هیبرید دو گانه بطور معمول استفاده می‌شود (شکل ۹-۶).پروتئین X،"برای آنکه اثرات متقابل پروتئین مطلوب باشد"، به دامین متصل به DNA یک فاکتور رونویسی متصل می‌شود. سپس پروتئین X با محصولات بیانی که از ترجمه کتابخانه (پروتئین اختیاریY) cDNA حاصل شده‌اند مخلوط می‌شود که با دامین فعال‌سازی فاکتور رونویسی هم جنس (هم ماهیت) ادغام و ترکیب می‌شود. نه پروتئین X و نه پروتئین Yبه تنهایی قابلیت تشکیل یک فاکتور رونویسی کامل و عملکردی را ندارند.فقط زمانی که دو پروتئین X و Y با هم اثر متقابل دارند دو دامین کنار هم قرار می‌گیرند و یک فاکتور رونویسی عملکردی حاصل می‌شود که قادر به فعال‌سازی رونویسی ژن‌های گزارشگر می‌شود.بیان آن‌ها با کمک آزمون فعال‌سازی می‌تواند اندازه‌گیری شود و بنابراین نشان‌دهنده اثرات متقابل بین پروتئین X و Y می‌باشد. با استفاده از هیبرید دو گانه مخمر، کل پروتئوم مخمر نان (*Saccharomyces cerevi-siae*) برای اثرات متقابل پروتئین-پروتئین تجزیه و تحلیل شد و به ۴۵۴۰ اثر متقابل از ۳۲۷۸ پروتئین مختلف منجر گردید (Ito *et al.* 2001). تجزیه و تحلیل میزان زیادی از پروتئوم برای اثرات متقابل پروتئین-پروتئین مورد آزمایش قرار گرفت. تخلیص تمایلی دو طرفه (TAP)[2]یکی دیگر از فناوری‌های که برای آنالیز کمپلکس‌های چندتایی پروتئین مناسب می‌باشد. این تکنیک اساس ترکیب و کروماتوگرافی تمایلی اسپکتروسکوپی جرمی می‌باشد.

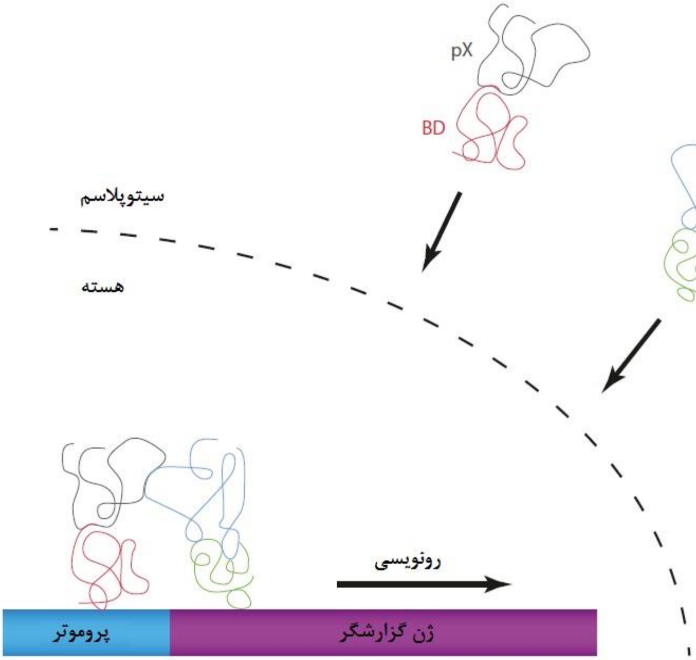
---

[1] Interactomics

[2] Tandem Affinity Purification (TAP)

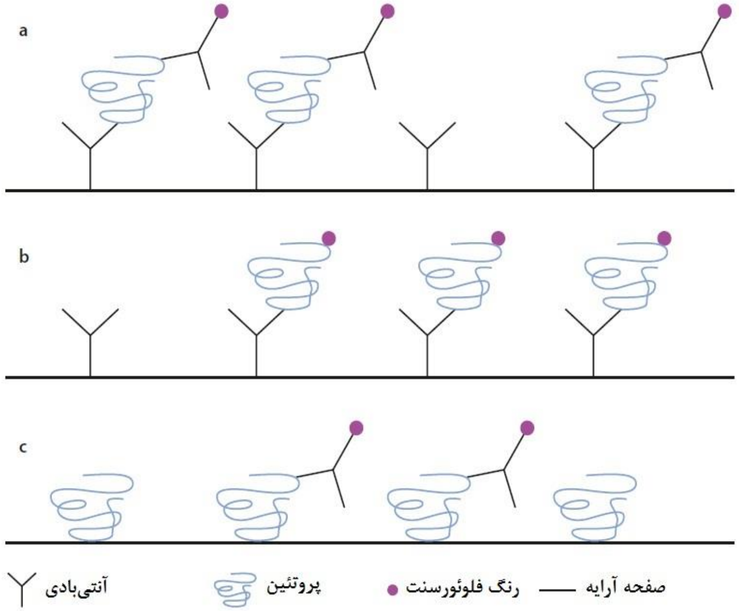


شکل ۸-۶) اثرات دارویی روی شبکه مولکولی.a) شبکه مولکولی متشکل از پروتئین‌ها و لیپیدها در سلامتی بیماران است. بیشتر ارتباطات در رنگ سبز نشان‌گذاری شده است، نشان از همبستگی منفی میان آنالیت‌ها است. b) شبکه مولکولی در گروه بیماران. عمده همبستگی نشان‌گذاری شده با رنگ قرمز معرف تغییر سلامتی شرایط بیمار است. c) شبکه مولکولی بیماران تحت درمان دارویی. بسیاری از خطوط سبزرنگ در سلامتی بیماران که بهبود یافتند دیده شد. هرچند در مسیر دوم، ارتباطات شبکه جدید به نظر می رسد (جعبه آبی) این به خاطر اثرات خارج از هدف (اثرات ناخواسته و جانبی) درمان دارویی است



شکل ۹-۶) شناسایی اثرات متقابل پروتئین-پروتئین با استفاده از سیستم هیبرید دوگانه مخمر. رونویسی یک ژن گزارشگر تنها هنگامی فعال می‌شود که ادغام پروتئین شامل اتصال به دامین DNA یک فاکتور رونویسی (BD) یا پروتئین تصادفی X(pX) با اثر متقابل پروتئین ادغامی دوم محتوی دامین فعال‌سازی فاکتور رونویس شناخته شده (AD) و یک پروتئین تصادفی Y(pY) باشد

ژن هدف به طوری که محصول ژنی با یک توالی کوتاه پپتیدی یا تگ که جداسازی پروتئین برچسب‌دار را از پروتئین دیگر تسهیل می‌کند، اصلاح می‌شود. این روش ملایم است و به طور همزمان آن پروتئین‌های سلولی دارای اثر متقابل که به پروتئین برچسب گذاری شده متصل می‌شوند. کمپلکس‌های چندتایی پروتئین جدا شده توسط ژل الکتروفورز تفکیک می‌شود و اجزای انفرادی بوسیله اسپکتروسکوپی جرمی تجزیه و تحلیل می‌شوند. در این مسیر، کمپلکس‌های چندتایی پروتئین در مخمر نان *Saccharomyces cerevisiae* شناسایی می‌شود. برخی کمپلکس‌های چندتایی پروتئین بیش از ۴۰ ترکیب انفرادی را شامل می‌شوند. علاوه‌براین عملکرد بالقوه برخی پروتئین‌های ناشناخته را بر اساس اثرات متقابل‌شان با پروتئین‌هایی که عملکرد سلولی آن‌ها به خوبی شناخته شده می‌توان مشخص نمود (Gavin *et al.* 2002) همان طور که در مورد هر آزمایش با کارایی بالا به مقدار زیادی از داده‌های تولید شده توسط این تراکتومیکس جهت توسعه پایگاه داده‌های اختصاصی نیاز دارد. مثلاً پایگاه داده اینترکتوم، پایگاه داده اثر متقابل مولکولی IntAct و STRING می‌باشند. اطمینان حاصل شود که تمام داده‌های مربوط به یک آزمایش در پایگاه‌های داده گنجانده شده است، اطلاعات بسیار کمی برای گزارش پروتکل آزمایش اثر متقابل مولکولی ضروری است تا نیازهای جزئی برای ذخیره داده‌های اثر متقابل پروتئین-پروتئین را تنظیم کند (Orchard *et al.* 2007).



شکل ۱۰-۶) آرایه پروتئین. a) ارزیابی ساندویچ، آنتی‌بادی‌هایی که پیوست هستند می‌توانند به صفحات آرایه انتخابی به پروتئین آنتی‌ژنیک بعد از انکوباسیون با پروتئین لیسیت متصل شوند. تشخیص پروتئین کپچر به وسیله آنتی‌بادی ثانویه متصل به جایگاه آنتی‌ژنیک متفاوت روی پروتئین اجرا می‌شود. b) در سنجش کپچر آنتی‌ژن، پروتئین‌های آنتی‌ژنیک به طور مستقیم قبل از انکوباسیون با برچسب آرایه پروتئین نشان‌گذاری می‌شوند. در نتیجه به یک آنتی‌بادی ثانویه برچسب‌گذاری شده برای تشخیص نیاز دارد. c) در آرایه مستقیم یا فاز معکوس، پروتئین‌ها به طور مستقیم به صفحات آرایه متصل شده و با آنتی‌بادی نشاندار شده تشخیص داده می‌شود

۳-۲-۱-٦ آرایه پروتئین

یک روش جایگزین برای تجزیه و تحلیل پروتئوم‌ها استفاده از فناوری آرایه پروتئین[۱] است (Eisenstein 2006). آرایه‌های پروتئین مشابه آرایه‌های DNA ساخته می‌شود. لکه‌هایی با تراکم بالاروی صفحات شیشه‌ای یا غشاء اعمال می‌شوند. این نقاط شامل واکنش‌دهنده‌هایی که میل ترکیبی بالایی برای اتصال به پروتئین دارند (به عنوان مثال آنتی‌بادی) آرایه‌های پروتئین هستند. همچنین برای تولید پروفایل پروتئین مناسب است که در آن سه نوع مختلف از آرایه‌های پروتئینی متمایز می‌شوند(MacBeath 2002).

- نوع اول سنجش ساندویچی[۲] است (شکل ۱۰-۶)، به این صورت که آنتی‌بادی‌ها به آرایه‌های پروتئین متصل می‌شوند. آرایه‌ها سپس با پروتئین لیسیت انکوبه می‌شوند. به شرطی پروتئین در لیسیت وجود دارد که یک آنتی‌بادی روی آرایه لکه‌گذاری کرده باشد و سپس پروتئین به آنتی‌بادی متصل می‌شود. تشخیص این اتصالاست با یک آنتی‌بادی ثانویه انجام می‌شود که در برابر پروتئین مشابه اما یک اپی‌توپ متفاوت نسبت آنتی‌بادی ثانویه کار می‌کند. آنتی‌بادی ثانویه نشانه‌گذاری شده (به عنوان مثال، با یک آنزیم که یک واکنش قابل تشخیص بصری را کاتالیز می‌کند) اجازه تشخیص و میزان اتصال را می‌دهد.
- نوع دوم سنجش آنتی‌ژن کپچر[۳] است (شکل ۱۰-۶)، همانطورکه در بالا ذکر شد، آنتی‌بادی‌های اولیه به طور مستقیم به ماتریکس متصل می‌شوند. این روش متفاوت از سنجش ساندویچی است و درآن پروتئین‌ها در لیزات از قبل برچسب‌گذاری شده‌اند (مانند رنگ‌های فلورسنت). با این سنجش دو نوع سلول لیسیت با کمک پروتئین‌های نشان‌گذاری شده مربوط به لیسیت با رنگ‌های مختلف مقایسه می‌شوند. هر دو لیزات مخلوط و با آرایه پروتئین انکوبه می‌شوند. بسته به میزان پروتئین متصل‌شده نشان‌گذاری شده، یکی از لیسیت‌ها حاوی پروتئین نشان‌گذاری بیشتر است. مفهوم اصلی این روش همانند یک آزمایش پروفایل بیان است.

---

[1] Protein Array

[2] Sandwich Assay

[3] Antigen Capture Assay

- در نوع سوم، سنجش مستقیم یا فاز معکوس[1] است. در این روش پروتئین‌ها و نه آنتی‌بادی‌ها به آرایه‌های پروتئینی متصل هستند که با استفاده آنتی‌بادی‌های نشان‌گذاری شده دنبال می‌شوند. در این روش پروتئین‌هایی که با آنتی‌بادی‌ها اثر متقابل دارند شناسایی می‌شوند (شکل c ۶-۱۰). آرایه‌های پروتئینی همچنین می‌توانند تأثر متقابل پروتئین-پروتئین را شناسایی کنند، همانطور که در بخش پیشین توصیف شد. با این حال، برخلاف سیستم مخمر دوگانه و TAP، این روش یک روش آزمایشگاهی است. اثرات متقابل پروتئین در خارج از سلول و شرایط آزمایشگاهی تجزیه و تحلیل می‌شود که ممکن است در شرایط درون سلول (زنده) منجر به اثرات متقابل نگردد. از سوی دیگر، آرایه‌های پروتئین مزیتی دارند که می‌توان آن‌ها را در مقادیر زیادی تولید نمود، که اجازه تکرار چندگانه آزمایش‌ها و تغییر شرایط آزمایش (pH، دما، غلظت پروتئین، حضور یون‌ها و کوفاکتورها) را می‌دهد. علاوه‌براین، با چنین آرایه‌ها، هزاران پروتئین و حتی کل پروتئوم‌ها را می‌توان هم زمان تجزیه و تحلیل کرد. به عنوان مثال پروتئین‌های مخمر نان *Saccharomycescerevisiae* بطور مطلوبی قادر است با پروتئین متصل به کلسیم کالمودلین[2] اثر متقابل داشته باشد. این آرایه حاوی ۵۸۰۰ پروتئین از ۶۲۰۰ پروتئین ممکن می‌باشد (Zhu *et al.* 2001). ۳۹ پروتئین سهم اثر متقابل بالقوه آن‌ها شناسایی شده است که فقط ششتای آن‌ها از قبل به عنوان پروتئین‌های مرتبط با کالمودلین شناخته شده بودند. مثال نشان می‌دهدکه چگونه آرایه‌های پروتئینی می‌توانند اثرات متقابل جدید پروتئین-پروتئین را تعریف کنند. علاوه‌براین، آرایه‌های پروتئینی همچنین می‌توانند در تشخیص اثرات متقابل پروتئین با گلیکوزیدها، لیپیدها، اسیدهای نوکلوئیک و دیگر لیگاندهای عمومی کمک کنند.

۳-۱-٦ متابولومیکس

مقایسه سلول‌های توموری با سلول‌های طبیعی نشان می‌دهد که چقدر نسبت به سلول‌های اولیه تغییر کردند و موثر بودند و آنزیم‌های متابولیکی بطور مکرر افزایش بیان

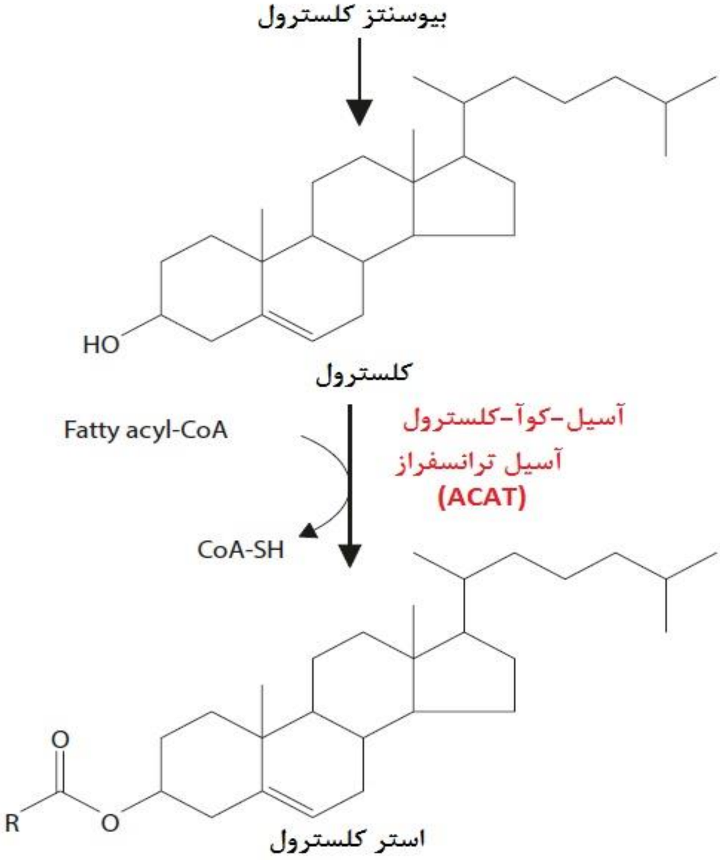
---

[1] Direct or Reverse-Phase Assay
[2] Calcium-Binding Protein Calmodulin

شده‌اند. با این حال نباید تعجب کرد چون سلول‌های سرطانی سریع‌تر رشد می‌کنند و بنابراین نیاز به متابولیت بیشتری دارند. بنابراین تصور می‌شود که با اندازه‌گیری متابولیت‌های سلولی سلول‌ها می‌توانند بطور مشابه در هر دو روش تکنیک‌های ریزآرایه یا پروتئومیکس پروفایل شوند. کل منبع متابولیت سلول، متابولوم نامیده می‌شود و زمینه تحقیقی مرتبط با پروفیل متابولیک، متابولومیکس نامیده می‌شود (شکل ۱-۶). متابولومیکس یک حوزه تحقیقاتی نسبتاً جدید است، اگرچه در سال ۱۹۷۰ و قبل از آن نیز رابینسون و پائولینگ آزمایشاتی را برای شناسایی و اندازه‌گیری متابولیت‌ها در ادرار انسان انجام داده بودند. پایگاه داده متابولیت انسانی (HMDB)[1] حاوی همه متابولیت‌هایی که می‌توانند در بدن انسان یافت شوند یا حداقل باید احتمالاً رخ می‌دهند، می‌باشد. این پایگاه مبتنی بر مسیرهای شناخته شده متابولیسم است، اما شواهد نهایی همچنان در انتظار تأیید است. این پایگاه داده حاوی بیش از ۴۲۰۰۰ مطالب در مورد متابولیت‌ها است که با بیش از ۵۶۰۰ توالی پروتئین مرتبطاست (Wishart *et al.* 2013). این مطالب شامل پپتیدها، لیپیدها، آمینواسیدها، نوکلئوتیدها، کربوهیدرات‌ها، اسیدهای آلی، ویتامین‌ها، موادمعدنی، موادغذایی، مواد دارویی، سموم، آلاینده‌ها و سایر مواد شیمیایی با جرم مولکولی کمتر از ۲۰۰۰ دالتون می‌باشند

---

[1] Human Metabolite Database (HMDB)



شکل ۱۱-۶) سنتز استر کلسترول، با کمک آنزیم آسیل کوانزیم آ کلسترول آسیل ترانسفراز کاتالیز می‌شود

این فهرست نشان می‌دهد که چرا تعریف متابولوم در مقایسه با ژنوم، ترانسکریپتوم یا پروتئوم بسیار دشوار است زیرا تنها به ژنوم بستگی ندارد بلکه به درک ماهیت جذب ماده نسبت به محیط‌زیست نیز وابسته است (به عنوان مثال از طریق غذا یا آلودگی). بنابراین، پایگاه داده شامل متابولیت‌های درون‌زا و برون‌زا[1] می‌باشد.

علیرغم این واقعیت که متابولوم نسبت به ژنوم، ترانسکریپتوم یا پروتئوم نسبتاً کوچک است، تقاضای فنی مورد نیاز برای متابولومیکس‌ها به طور خاص بالا است. علت این حالت در

---

[1] Endogen and Exogen Metabolites

تنوع شدید خواص فیزیکی و فیزیکوشیمیایی متابولیت‌هایی است که اندازه‌گیری می‌شود. برخی متابولیت‌ها کوچک و آبدوست هستند (مانند ویتامینC) و از طرفی دیگر متابولیت‌ها دارای جرم مولکولی بالا و غیرقطبی می‌باشند (مانند استرکلسترول‌ها) (شکل ۱۱-۶). در حال حاضر، هیچ فناوری منحصر بفردی برای شناسایی و تعیین کل متابولیت‌ها به طور همزمان وجود ندارد. با این حال، در چند سال اخیر پیشرفت در فناوری‌ها منتج به روش‌هایی است که می‌توانند تعداد کمی از متابولیت‌ها را به صورت موازی اندازه‌گیری کنند. معمولاً، مقدار نسبی متابولیت‌ها در دو نمونه‌های مختلف همانند روش ریزآرایه DNA با یکدیگر مقایسه می‌شوند. علاوه‌براین، تجهیزات حساس‌تر و استانداردهای مناسب نیز اجازه اندازه‌گیری مطلق متابولیت‌ها در یک آزمایش واحد را می‌دهد.

برای اندازه‌گیری متابولیت‌ها، دو روش اصلی استفاده می‌شود: طیف‌سنجی رزونانس مغناطسی هسته‌ای (NMR)[۱] و طیف‌سنجی جرمی[۲]. طیف‌سنجی NMR می‌تواند داده‌های فیزیکی، شیمیایی، الکترونیکی و به‌ویژه ساختاری از مولکول‌ها و متابولیت‌ها ارائه دهد. هر چند روشی که بیشتر بکارگرفته می‌شود طیف‌سنجی جرمی است. معمولاً در مرحله کروماتوگرافی (مانند کروماتوگرافی گازی GC)[۳] ابتدا جداسازی متابولیت‌ها انجام می‌شود سپس با استفاده از تجهیزات بسیارتخصصی بیش از ۴۰۰۰ داده خام پیک (نمودار)، که مربوط به ۱۸۰۰پیک متابولیت است، می‌توان اندازه‌گیری کرد (Kell 2006). آزمایشات متابولومیک حاوی مقدار زیادی از اطلاعات است که باید تجزیه و تحلیل شوند و تبدیل به دانش زیست‌شناختی مفید گردند.

بسیاری از محققان اظهارنظر کرده‌اند که متابولومیکس ویژگی‌های یک سلول را بهتر از ژنومیکس، ترانسکریپتومیکس یا پروتئومیکس توصیف می‌کند. آن‌ها نظراتشان را با فرآیندهای سلولی که به فرآیند رونویسی ژن و آن به نوبه خود به پروتئین کد شده منجر شده و در نهایت مسئول تولید متابولیت است، اشاره نمودند. بنابراین متابولیت‌ها در انتهای زنجیره اطلاعات قرار دارند و بنابراین به‌طور نزدیکی به نحوه عملکردشان مربوط می‌باشد. استدلال دیگر، تقویت اطلاعات است.آزمایشات نشان می‌دهد که حتی تغییرات جزئی در غلظت تعداد کمی آنزیم منجر به معنی‌داری در غلظت بسیاری از متابولیت‌ها می‌شود(Raamsdonk *et al.* 2001)).

---

[1] Nuclear Magnetic Resonance (NMR)
[2] Mass Spectroscopy
[3] Gas Chromatography (GC)

دلایل این امر این است که سنتز و گردش (تغییر و تبدیل) متابولیت‌ها به طورکلی توسط آنزیم‌های مختلف کاتالیز می‌شود و یک متابولیت می‌تواند در بسیاری از واکنش‌های مختلف دخیل باشد.در این ارتباط صحبت از شبکه‌های متابولیک امکان پذیر به نظر می‌رسد (شکل ۳-۷ و ۴-۷).

قدرت متابولومیکس آن است که امکان ساخت مدل‌های تغییرات کمی در متابولیسم به دلیل ساختار شبکه‌ای آن، فراهم می‌کند. در واقع، در حال حاضر بسیاری از مدل‌ها به‌ویژه برای موجوداتی مانند مخمر نان *Saccharomyces cerevisiae* به خوبی مورد مطالعه قرار گرفته‌اند. به عنوان مثال، با کمک یک مدل متابولیکی که معرف ۷۵۰ ژنو ۱۱۴۹ واکنش در مخمر نان است، تعداد ۴۱۵۴ فنوتیپ رشد پیش‌بینی شده است. مقایسه نتایج آزمایشگاهی نشان داد که مدل در حقیقت توانست ۸۳ درصد فنوتیپ را پیش‌بینی کند (Duarte *et al.* 2004). نسلی از چنین مدل‌های متابولیکی تا حدی با دیگر حوزه‌ها، یعنی سیستم‌های زیست‌شناسی همپوشانی دارند که جزئیات آن در بخش (۶-۲) بیشتر توضیح داده شد.

سازواره‌های الکترونیکی[1] که در حال حاضر به عنوان دستگاه‌های قابل حمل در دسترس هستند، استفاده دیگری در تجزیه و تحلیل متابولیت دارند (Koczulla *et al.* 2011). سنسورهای نانوکامپوزیت از سازواره‌های الکترونیکی برای تشخیص مقدار جزئی از گازهای مولکولی، اسیدها، بازها، و بسیاری از مولکول‌های دیگر ساخته شده‌اند. الگوها برای ترکیب‌های مختلف می‌توانند بازیابی و اصلاح شوند و با روش‌های محاسباتی با استفاده از ترکیبی از سنسورهای مختلف، تجزیه و تحلیل شوند. برای مثال سایرانوز۳۲۰(Cyranose۳۲۰) یک سازواره الکترونیکی از سنسیجنت (Sensigent) است که می‌تواند برای تجزیه و تحلیل میزان هوای تنفسی انسان استفاده شود. در مطالعه ۳۰ بیمار این امکان وجود دارد که بین هوای تنفسی بیماران مبتلا به سرطان ریه غیرسلولی کوچک[2]، بیماری ریوی انسدادی مزمن (COPD)[3]، و بیماران سالم تمایز قائل شود (Dragonieri *et al.* 2009). در مطالعه دیگری سه سویه مختلف باکتریایی، شامل سویه‌های *استافیلوکوکوس آرئوس* مقاوم به متیسیلین

---

[1] Electronic Noses

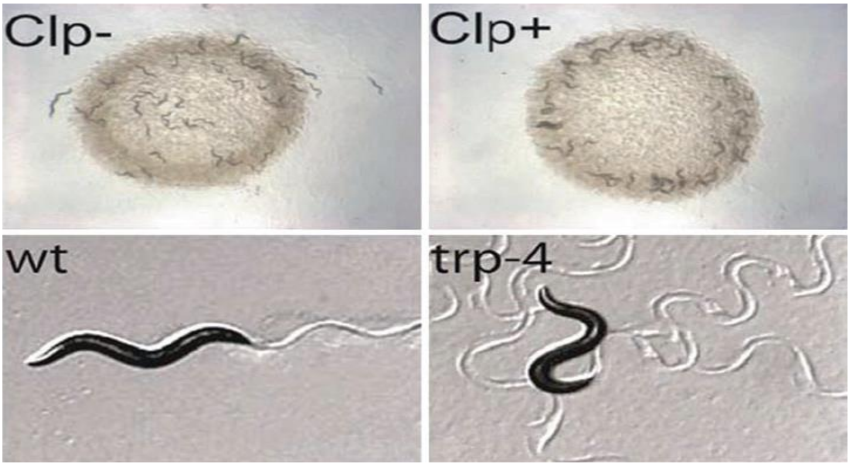
[2] non-Small-Cell Lung Cancer

[3] Chronic Obstructive Pulmonary Disease (COPD)

(MRSA)[۱] و *استافیلوکوکوس آرئوس* حساس به متیسیلین (MSSA)[۲]، تشخیص و متمایز گردیدند (Dutta and Dutta 2006).

٤-١-٦ فنومیکس[۳]

فنوتیپ یا ظاهر فیزیکی، مجموعه‌ای از ویژگی‌های قابل مشاهده بیرونی افراد است (شکل ۶-۱۲) که به هر دو ویژگی مورفولوژیکی و فیزیولوژیکی اشاره دارد. در نتیجه، خواص قابل مشاهده و قابل اندازه‌گیری یک موجود یا سلول بر اساس اثرات متقابل ژنوتیپ با محیط، فنوتیپ را تشکیل می‌دهند (بخش ۶-۱-۲).



شکل ۱۲-۶) فنوتیپ‌ها در کرم حلقوی *Caenorhabditis elegans*. a) بیشترگونه‌ها تغذیه‌کننده انفرادی هستند و یک فنوتیپ چسبنده را نمایش نمی‌دهند (Clp-). b) برخی از گونه‌ها بر روی مرز جمع می‌شوند و به عنوان فنوتیپ چسبیده شناخته می‌شوند (Clp+). فنوتیپ بوسیله یک پدیده طبیعی چندشکلی ژنتیکی در یک تک ژن حاصل می‌شود. c) فنوتیپ کرم متحرک گونه وحشی (فرم طبیعیwt). d) فنوتیپ trp-4 محکوم به مرگ با وضعیت بدنی غیرطبیعی می‌باشد. جهش‌های کانال یونی جنبش فیزیکی با انعطاف‌پذیری بیشتری به همراه دارد

---

[1] Methicillin-Resistant *Staphylococcus aureus* (MRSA)
[2] Methicillin-Susceptible *Staphylococcus aureus* (MSSA)
[3] Phenomics

بنابراین، با این تعریف، متابولومیکس، قابل سنجش است و همچنین معرف فنوتیپ بر اساس اثرات متقابل ژنوتیپ با محیط می‌باشد. بسیاری از روش‌ها در زمینه ژنومیکس عملکردی وجود دارد که عملکرد پروتئین را براساس فنوتیپ‌ها تعریف می‌نمایند. این حوزه تحقیقاتی فنومیکس نامیده می‌شود اگر در یک فرایند با کارایی بالا انجام شود.

در ابتدا، غربالگری ژنتیکی در ژنوم‌هایی استفاده شده بود که دارای جهش‌یافته تصادفی بودند، این جهش‌ها منجربه ثبت فنوتیپ‌ها شد و ژن‌های مسئول به فنوتیپ‌های تغییریافته شناسایی گردیدند. با استفاده از این روش چندین هزار ژن شناسایی و مشخص شدند. ورود کل توالی ژنوم، رویکردهای جایگزین را ارائه می‌دهد تا غربالگری ژنتیکی برای آن ژن‌ها بدون کارکرد توصیف شده اجرا شود. استراتژی که یک ژن مجزا را از به عملکردش مرتبط می‌سازد، ژنتیک برگشتی[1] می‌نامند.

به عنوان تجزیه و تحلیل، آزمایشات ناک‌اوت (حذف یا مرگ)[2] اغلب جایی انجام می‌شود که ژن‌ها جهش‌یافته انتخابی هستند (خاموش هستند[3]) به طوری که هیچ پروتئین عملکردی رمزگذاری نشود. نتیجه می‌تواند یک فنوتیپ تغییریافته شود که بتوان خواص آن را دقیقاً مستند کرد. اگر یک ژن یک پروتئین ضروری را رمزگذاری کند، فنوتیپ حاصل ممکن است کشنده باشد، به عنوان مثال موجود یا سلول بمیرد. چنین آزمایشات ناک‌اوت معمولاً در لاین‌های سلولی یا در موجودات مدل مانند مگس میوه *Drosophila melanogaster* انجام می‌شود. معایب این روش پیچیدگی آزمایشگاهی و وقت‌گیر بودن آن است که منعکس‌کننده این واقعیت است داده‌های موجود ناک‌اوت ژنوم گسترده و کامل تنها برای تعداد کمی موجودات زنده وجود دارد (به عنوان مثال مخمر نان).

مشابه آزمایشات ناک‌اوت، آزمایشات ناک‌این (انتقال ژن به درون)[4] هستند که نقش محصولات ژن را روشن می‌سازد. در این مورد، ژن‌ها به سلول‌ها یا موجودات منتقل می‌شوند سپس مشاهده می‌شود که آیا آن‌ها باعث تغییرات فنوتیپی شده‌اند یا خیر. استراتژی ناک‌این به عنوان اثبات نقش پروتئین اضافه شده، اغلب مورد استفاده قرار می‌گیرد. اگر تغییرات فنوتیپی قبل از ناک‌اوت با کمک آزمایشات ناک‌این برگشت کند، شک و تردید کمتری برای نحوه

---

[1] Reverse Genetics
[2] Knockout
[3] Switched Off
[4] Knockin

عملکرد پروتئین وجود خواهد داشت. به عنوان مثال، باکتری‌هایی که در آن یک پروتئین تاژکدار[۱] خاص ناک‌اوت شده است بی‌حرکت می‌شود. اگر کلنی باکتریایی مشابهی در آزمایشات ناک‌این ژن را بازیابی کند و متعاقباً تحرک بازیابی شود این مدرک محکمی دال بر این است که پروتئین برای عملکرد مناسب تاژک ضروری است.

متاسفانه استراتژی‌های ناک‌اوت و ناک‌این بسیار دشوار و غیرقابل کنترل هستند. کشف و کاربرد آزمایشی RNAمداخله‌گر(RNAi)[۲]منجربه یک انقلاب برای غربالگری ژنتیک معکوس شده است. RNAi یک مکانیزم تکاملی محافظت شده است که شامل سرکوب بیان ژن توسط RNA دورشته‌ای(dRNA)[۳] است (Vanhecke and Janitz 2005). پس از دسترسی به سیتوپلاسم سلول، مولکول‌های RNA دورشته‌ای ابتدا به قطعات با طول ۲۱-۲۵ نوکلئوتید، RNAi یا RNA مداخله‌گر کوچک (siRNA)[۴]نامیده می‌شوند، بوسیله آنزیم دایسر[۵] برش می‌خورند (شکل ۱۳-۶). siRNA تک‌رشته‌ای سپس درون کمپلکس آنزیمی که کمپلکس القاءکننده خاموشی RNA یا (RISC)[۶] نامیده می‌شود بارگذاری می‌شود. کمپلکس آنزیمی فعال‌شده، هدایت‌شده بوسیله رشته siRNA، بطور اختصاصی به mRNA مکمل متصل می‌شود که با فعالیت اندونوکلئازی کمپلکس RISC برش می‌خورد. به این ترتیب، بیان ژن هدف به طور خاص مسدود می‌شود و از ترجمه پروتئین شناخته شده جلوگیری می‌شود. از آنجایی که مهار رونویسی بوسیله RNAi همیشه کامل انجام نشود اصطلاح ناک‌دان[۷] کاربرد دارد.

مزیت فن‌آوری RNAi شامل آزمایشات سریع، ساده، مقرون به صرفه و از طرفی قابل کنترل و کارایی بالا می‌باشد. تعداد زیادی از نشریات ژنوم کامل با استفاده از RNAi را تجزیه و تحلیل کرده‌اند. به عنوان مثال، ۸۶ درصد از تمام ژن‌های نماتد*Caenorhabditis elegans* توسط RNAi مورد بررسی قرارگرفت (Kamath *et al*,2003). تقریباً ۱۰ درصد از ژن‌های هدف منجربه تغییر در فنوتیپ شدند که تقریباً یک سوم از قبل شناسایی شده بودند. در

---

[1] Flagellar Protein
[2] RNA interference (RNAi)
[3] doublestranded RNA (dsRNA)
[4] small interfering RNA (siRNA)
[5] Dicer
[6] RNA-Induced Silencing Complex (RISC)
[7] Knockdown

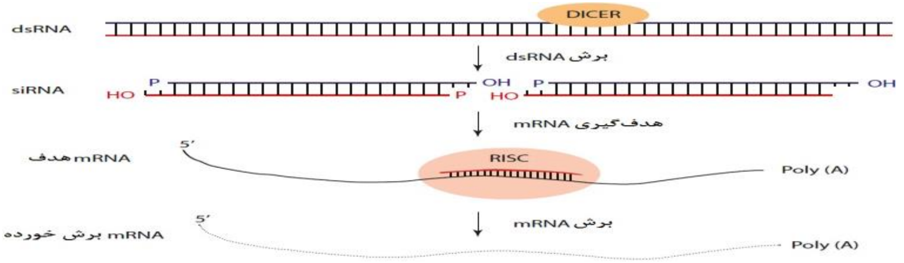
مطالعه دیگری، تعدیل‌کننده‌های جدید p53 که موجب توقف چرخه سلول در سلول‌های انسانی شده برای RNAi مورد جستجو قرار گرفت. از ۸۰۰۰ ژن تجزیه و تحلیل شده، پنج تعدیل‌کننده جدید کشف شد (Berns *et al*. 2004).

متاسفانه، تمام نتایج RNAi کاملاً قابل اعتماد نیستند. به عنوان مثال، مشخص شده است که کارایی RNAi به طورگسترده‌ای وابسته به توالی نوکلئوتیدی متصل است. در برخی موارد، mRNA هدف یا فقط تا حدی تخریب شده است یا اصلاً و در مجموع منجربه نتیجه منفی کاذب می‌شود. آزمایشگر هیچ تغییری در فنوتیپ نخواهد دید و نتیجه‌گیری می‌کند که محصول ژن هیچ عملکرد مهمی ندارد. چنین داده‌هایی باید توسط یک روش مستقل دیگر بررسی شوند مانند RT-PCR، برای تعیین اینکه آیا RNA هدف در واقع تخریب شده است یا خیر. همچنین، RNAi نیز می‌تواند نتایج مثبت کاذب تولید کند. زمانی siRNA توسط دایسرهای نوکلئازی تولید شدند، به بیش از یک مولکول mRNA هدف متصل می‌شوند که منجربه تخریب چندین mRNA می‌شود .بنابراین تغییرات فنوتیپ‌ها را نمی‌توان به طور یک‌طرفه انجام داد و در بدترین حالت ممکن است منجربه پیش‌بینی‌های نادرست عملکرد محصولات ژن شود.

PhnomicDB یک پایگاه اطلاعاتی بسیار جالب و یکپارچه است که در آن فنوتیپ‌های موجودات متفاوتی که توسط روش‌های مختلف تولید شده‌اند (مانند ناک‌اوت، ناک‌این، ناک‌داون)، به یک پایگاه داده که با داده‌های ژنوتیپی تلفیق شده است، جمع‌آوری شده است. علاوه بر این پایگاه داده دیگری مانند: پایگاه داده تنوع ژنوم انسان (HGVDB)[1] ارتباط ژنوتیپی–فنوتیپی انسان را ذخیره می‌کند. (Brookes and Robinson 2015).

---

[1] Human Genome Variation Database (HGVDB)



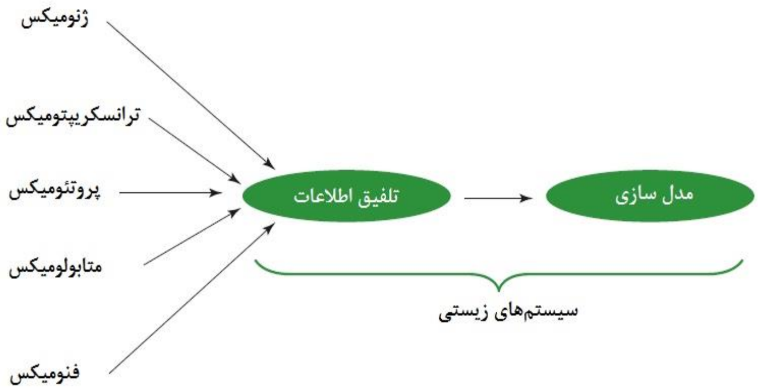
شکل ۱۳-۶) برش اختصاصی mRNA بوسیله RNA مداخله‌گر (RNAi). یک ریبونوکلئاز نوع سه (دایسر) به توالی RNA دورشته‌ای متصل می‌شود و آن را به دوپلکس‌های ۲۱-۲۵ جفت‌بازی می‌شکند و به عنوان RNA مداخله‌گر کوچک (siRNA) معرفی می‌شود. siRNA با یک کمپلکس چند پروتئینی بنام کمپلکس القاءکننده خاموشی RNA یا (RISC) که حاوی یک RNAse است، ترکیب می‌شود. siRNA از RISC باز شده و رشته سنس (Sense) آن آزاد می‌شود و جفت شدن رشته آنتی‌سنس siRNA با رشته مکمل RNA پیام‌رسان (mRNA) را تسهیل می‌کند. فعالیت نوکلئازی RISC با اتصال شروع شده و منجربه برش mRNA هدف می‌گردد. سپس mRNA تخریب می‌شود و نهایتاً باعث کاهش بیان ژن هدف می‌شود

۶-۲ سیستم‌های زیستی

بحث‌های مذکور کارایی بالای ژنومیکس، ترانسکریپتومیکس، پروتئومیکس، متابولومیکس و فنومیکس را به عنوان فناوری‌های مهمی که تعیین عملکرد محصولات ژن را تسهیل می‌کند، فراهم نموده است. با این حال این روش‌ها عاری از ایراد نبوده و نتایج منفی و مثبت کاذبی را نیز تولید می‌کنند. نتایج منفی کاذب می‌تواند منجربه از دست رفتن اطلاعات شود، درحالیکه نتایج مثبت کاذب ممکن است فرد متخصص را در جهت اشتباه هدایت نماید. بنابراین، برای شناسایی نتایج معتبر یک ایده، تمام داده‌های موجود از فناوری‌های ذکرشده را ادغام می‌کنند و آن‌ها را با هم تجزیه و تحلیل می‌کنند (شکل ۱۴-۶). این تلفیق و ادغام داه‌های آزمایشگاهی، بیان فرضیه‌های قابل اعتماد را بهبود می‌بخشد. زمینه تحقیقاتی که بر ادغام داده‌ها متمرکز شده است را به عنوان سیستم‌های زیستی[۱] معرفی می‌کنند، زیرا آن سیستم‌های بیولوژیکی را تجزیه و تحلیل می‌کند. سیستم‌های زیستی با هدف تولید تصویر دقیقی از تمام فرآیندهای تنظیمی در یک سلول یا موجود زنده به همراه تجزیه و تحلیل اثرات متقابل بین اجزای تشکیل‌دهنده بیولوژیکی مانند: مسیرهای متابولیکی، اندامک‌ها، سلول‌ها و بافت‌ها ارائه می‌دهد.

---

[1] Systems Biology



شکل ۱۴-۶) سیستم‌های زیستی داده‌های مشتق شده از فناوری‌های مختلف آزمایشی و مدل‌های محاسباتی را ادغام می‌کند

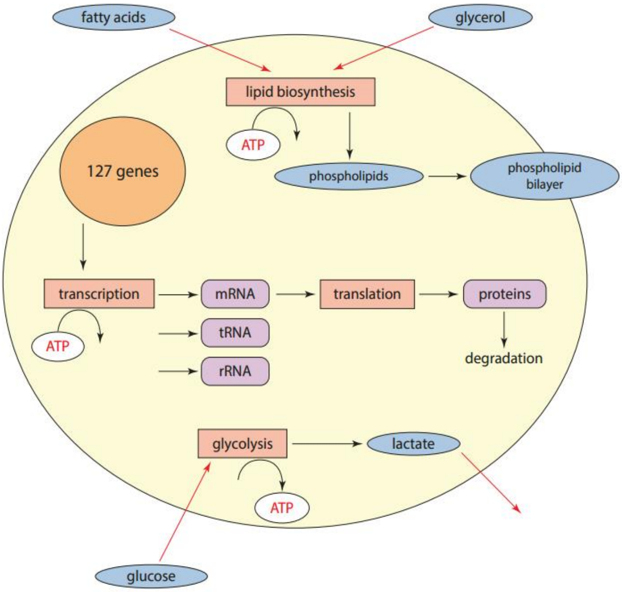
یک مثال از رویکرد سیستم‌های زیستی، تجزیه و تحلیل فاگوزوم‌ها[1] است که در اندامک‌های ویژه سلول‌های فاگوسیت‌کننده یافت می‌شود (مانند ماکروفاژها[2]). پس از فاگوسیتوز، ذراتی مانند باکتری‌ها به فوگوزوم منتقل می‌شوند و در آنجا از بین می‌روند. در مطالعه استوارت و همکاران در سال ۲۰۰۷، فاگوزوم یک سلول مشتق شده از مگس میوه مورد تجزیه و تحلیل قرار گرفت. پروتئین‌های فاگوزوم با استفاده از روش‌های پروتئومیکس کلاسیک شناسایی شده بودند. ساخت یک شبکه اثرات متقابل پروتئین-پروتئین، نتایجی که توسط آزمایشات RNA مداخله‌گر مورد تأیید قرار گرفت را تکمیل نمود. با کمک سیستم‌های زیستی رویکرد مدل دقیق فاگوزوم ساخته شده است و تنظیم پروتئین‌های جدید و مسیرهای مرتبط با فاگوسیتوز شناسایی می‌شود.

با این حال، سیستم‌های زیستی اغلب فراتر از توضیح اطلاعات آزمایشگاهی است. هدف بلند پروازانه این است که مدل‌های کامپیوتری را توسعه دهیم تا سیستم‌های زیستی شبیه‌سازی شده و تغییر پارامتر توالی‌ها پیش‌بینی گردند (مانند تغییر غلظت متابولیت‌های اختصاصی). یکی از اولین مدل‌های ریاضی در زیست‌شناسی در سال ۱۹۵۲ توسط آلن هاجکین و اندرو هاکسلی منتشر گردید که تأثیر بالقوه آن را توضیح می‌داد. از آن زمان به بعد در

---

[1] Phagosome
[2] Macrophages

دسترس بودن داده‌های با کیفیت بالا (کیفی و کمی) و ظرفیت‌های بیشتر کامپیوتر اجازه دادند مدل‌های واقع‌گرایانه‌تر توسعه یابند. به عنوان مثال، یک مدل برای شبیه‌سازی گلیکولیز در مخمرنان تولیدشده است. مقایسه داده‌های آزمایشگاهی، اکثر غلظت متابولیت‌ها به درستی در حدود انحراف معیار۲،پیش‌بینی شده بود. امروزه مدل‌های کامپیوتری بسیاری توسعه یافته‌اند که شبیه‌سازی کامل سلول را اجرا می‌کند (Ishii *et al*. 2004). یک مدل شناخته شده، سیستم E-Cell است که یک باکتری مجازی[1]حاوی۱۲۷ ژن ضروری از ژنوم *Mycoplasma genitalium* می‌باشد (شکل ۱۵-۶). این باکتری دارای کمتر از ۵۰۰ ژن است و بنابراین برای ساخت یک مدل سلولی مناسب می‌باشد.



شکل ۱۵-۶) بررسی متابولیسم در مدل E-cell. سلول مدل دارای مسیری برای گلیکولیز و بیوسنتز فسفولیپید، رونویسی و ترجمه دارد

با استفاده از این مدل انتقال خارج سلولی گلوکز از طریق غشاء سلولی شبیه‌سازی می‌شود و علاوه بر متابولیسم قند تولید ATP را همراهی می‌کند. این مدل شگفتی بزرگی ایجاد نمود

---

[1] Virtual Bacterium

به این صورت که زمانی که غلظت گلوکز خارج سلولی به صفر برسد، این مدل، افزایش موقت در غلظت ATP داخل سلولی را قبل از افت نهایی، پیش‌بینی نمود. برخلاف انتظار این که غلظت ATP بلافاصله پس از تخلیه گلوکز کاهش یابد. پس از حدس و گمان، سرانجام مشخص شد پیش‌بینی مدل صحیح است. در طی گلیکولیز از هر مولکول گلوکز دو مولکول ATP تولید می‌شود. پس از بررسی دقیق‌تر آشکار شد که در قسمت اول گلیکولیز، دو مولکول ATP قبل از این که چهار مولکول ATP در بخش دوم واکنش تولید شود، مصرف شده است. در آن لحظه که غلظت گلوکز به صفر کاهش می‌یابد، قبل از تولید مولکول‌های جدید ATP، مصرف مولکول‌های ATP متوقف می‌شود. بنابراین این مدل، تغییرات کوتاه مدت را شناسایی و به طور دقیق افزایش آنی غلظت ATP را پیش‌بینی می‌کند. در سال ۲۰۱۲ مدل محاسباتی پاتوژن انسانی *Mycoplasma genitalium* ارائه گردید که کل سلول شامل تمام اجزای مولکولی و فعل و انفعالات آن‌ها را شبیه‌سازی نمود (Karr *et al.* 2012). این مدل با توصیف یک چرخه سلولی کامل از یک سلول منفرد و پیش‌بینی رفتار سلولی هدف‌گذاری شده است. بر این اساس ژنوم کامل ۵۲۵ ژن، تجزیه و تحلیل دقیق شدند و نهایتاً تبدیل به بیش از ۹۰۰ منابع داده از قبیل منابع اولیه، کتاب‌ها و پایگاه داده شدند. به طورکلی، مجموعه داده‌های ۱۹۲ نوع وحشی و ۳۰۱۱ سلول ناک‌اوت بر روی یک خوشه با ۱۲۸ گره محاسبه شدند. محاسبات در نهایت با کمک داده‌های آزمایشگاهی که بخشی از توسعه مدل نبودند تأیید گردیدند. بینش عمیق در مورد فرایندهای سلولی که سابق غیر قابل مشاهده بوده با استفاده از این مدل نظیر میزان اثرات متقابل DNA-پروتئین در شرایط زنده ایجاد شده است.

ظهور سیستم‌های زیستی با توسعه یک قالب اختصاصی برای ارائه مدل‌های بیولوژیکی به نام زبان نشانه‌گذاری سیستم‌های زیستی (SBML)[1]، همراه است. SBML یک XML بر مبنای فرمت قابل خواندن کامپیوتر بوده که در شبکه‌های بیولوژیکی دقیقاً شرح داده شده است. ایده اصلی SBML ایجاد یک فرمت استاندارد بود که امکان تبادل ساده داده‌ها را بین بسیاری از نرم‌افزارهای مختلف فراهم می‌کند. بنابراین، هر مدل محاسبه شده را می‌توان در محیط نرم‌افزار مختلف بدون تلاش اضافی آزمایش نمود. هم زمان، پایگاه‌های تخصصی داده‌ها نیز تأسیس شدند که در آن مدل‌های کامپیوتری می‌توانند ذخیره شوند و برای همه دانشمندان

---

[1] Systems Biology Markup Language (SBML)

علاقه‌مند قابل دسترسی باشند. یک مثال پایگاه داده، پایگاه داده مدل زیستی (بیومدل EBI[1]) است.

سایت‌های مفید

agilent. http://www.genomics.agilent.com
affymetrix. http://www.affymetrix.com/
arrayexpress. http://www.ebi.ac.uk/arrayexpress/index.html
bioconductor. https://www.bioconductor.org/
biomodels. https://www.ebi.ac.uk/biomodels-main/
ecell. http://www.e-cell.org
ercc. http://jimb.stanford.edu/ercc/
geo. https://www.ncbi.nlm.nih.gov/geo/
genedisruptionproject. http://www.fruitfly.org/p_disrupt/index.html
genepattern. http://software.broadinstitute.org/cancer/software/genepattern/
hgvdb. http://www.hgvd.genome.med.kyoto-u.ac.jp/
hmdb. http://www.hmdb.ca/
hpp. http://www.thehpp.org/
intact. http://www.ebi.ac.uk/intact/
maqc. http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.
melanie. http://world-2dpage.expasy.org/melanie/
miame. http://fged.org/projects/miame/
sage. http://www.sagenet.org/
sagemap. https://www.ncbi.nlm.nih.gov/projects/SAGE/
sensigent. http://www.sensigent.com/products/cyranose.html
string. http://string-db.org/
swiss2dpage. http://world-2dpage.expasy.org/swiss-2dpage/
tm4. http://www.tm4.org/

---

[1] BioModels Database

منابع

1. Allis CD, Jenuwein T (2016) The molecular hallmarks of epigenetic control. Nat Rev Genet 17:487–500.
2. Berns K, Hijmans EM, Mullenders J et al (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature 428(6981):431–437.
3. Brazma A, Hingamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371.
4. Brookes AJ, Robinson PN (2015) Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16(12):702–715.
5. Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32:490–495.
6. Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14(7):1298–1309.
7. Dutta R, Dutta R (2006) Intelligent Bayes Classifier (IBC) for ENT infection classification in hospital environment.Biomed Eng Online 5:65.
8. Dragonieri S, Annema JT, Schot R et al (2009) An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer 64(2):166–170.
9. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866–5878.
10. Eisenstein M (2006) Protein arrays: growing pains. Nature 444(7121):959–962.
11. Gavin AC, Bosche M, Krause R et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147.
12. Gershon D (2005) DNA microarrays: more than gene expression. Nature 437(7062):1195–1198.
13. Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class.prediction by gene expression monitoring. Science 286:531–537.
14. Griffin TJ, Goodlett DR, Aebersold R (2001) Advances in proteome analysis by mass spectrometry. Current Opinin. Biotech 12:607–612.
15. Holloway AJ, van Laar RK, Tothill RW, Bowtell DL (2002) Options available from start to finish-for obtaining data from DNA microarrays II. Nat Genet 32:481–489.

16. Ishii N, Robert M, Nakayama Y et al (2004) Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 113(1–3):281–294.
17. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast interactome. Proc Natl Acad Sci U S A 98:4569–4574.
18. Ji H, Davis RW (2006) Data quality in genomics and microarrays. Nat Biotechnol 24(9):1112–1113.
19. Kamath RS, Fraser AG, Dong Y et al (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421(6920):231–237.
20. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV et al (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150(2):389–401.
21. Kell DB (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 11(23–24):1085–1092.
22. Koczulla AR, Hattesohl A, Biller H et al (2011) Smelling diseases? A short review on electronic noses. Pneumologie 65(7):401–405.
23. Matsumura H, Bin Nasir KH, Yoshida K et al (2006) SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. Nat Methods 3(6):469–474.
24. MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32:526–532.
25. Orchard S, Salwinski L, Kerrien S et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25(8):894–898.
26. Raamsdonk LM, Teusink B, Broadhurst D et al (2001) A functional genomics strategy that uses Metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19(1):45–50.
27. Rual JF, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062):1173–1178.
28. Quackenbush J (2001) Computational analysis of microarray data. Nature Rev. Genetics 2:418–427
29. Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32:502–508.
30. Stuart LM, Boulais J, Charriere GM (2007) A systems biology analysis of the Drosophila phagosome. Nature 445(7123):95–101.
31. Teusink B, Passarge J, Reijenga CA et al (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267(17):5313–5329.

32. Vanhecke D, Janitz M (2005) Functional genomics using high-throughput RNA interference. Drug Discov Today 10(3):205–212.
33. Wishart DS, Jewison T, Guo AC et al (2013) HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res 41(Database issue):D801–D807.
34. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A et al (2001) Global analysis of protein activities using proteome chips. Science 293:2101–2105.

فصل هفتم

تحلیل مقایسه‌ای ژنوم

۷-۱ عصر توالی‌یابی ژنوم

بیشترین دستاوردهای فوق‌العاده زیست‌شناسی بر پایه ژنوم طی سال‌های اخیر بوسیله پیشرفت در تکنولوژی‌هایی نظیر توالی‌یابی DNA، توسعه سخت‌افزارها و همچنین نرم‌افزارهایی است که امکان ذخیره‌سازی و نام‌گذاری مقادیر عظیمی از داده را فراهم کرده است. در سال ۲۰۱۶ تعداد کل همه نوکلئوتیدهای قابل دسترسی آزاد در پایگاه NCBI، بالغ بر ۲۱۸ میلیارد باز در ۱۹۶ میلیون توالی DNA بوده است. همچنین در این سال نیز تعداد همه توالی‌های پروتئینی در بزرگترین پایگاه داده پروتئینی غیرتکراری UniprotKB[uniprotkb] در EBI حدود ۶۵ میلیون می‌باشد.

اولین ژنوم توالی‌یابی شده کامل، از موجودات میکروبی *Haemophilusinfluenzae* (Fleischmann *et al.* 1995) و *Mycoplasma genitalium* (Fraser *et al.* 1995)، در سال ۱۹۹۵ منتشر شدند. تا سال ۲۰۱۶، تعداد ۱۶۵۱۷۸ ژنوم میکروبی توالی‌یابی شده‌اند یا هم اکنون در حال توالی‌یابی هستند (۱۶۳۳۰۲ از باکتری و ۱۸۷۶ از آرکئی‌باکتر). در بین این موارد ژنوم کامل هر دو سویه باکتری بیماری‌زا[1] و غیربیماری‌زا[2]، که شناسایی عوامل بیماری‌زایی را تسهیل می‌کند نیز وجود دارند. در مجموع طی چند سال آینده، همه موجودات بیماری‌زای مهم انسان، حیوانات و گیاهان توالی‌یابی خواهند شد. این سیل داده به احتمالات جدیدی در تولید عوامل ضدمیکروبی، واکسن‌ها و آزمون‌های تشخیصی منجر می‌شود که همه آن‌ها باید به جنگ علیه بیماری‌های عفونی کمک کند (Selzer *et al.*2000).

علاوه‌براین ژنوم‌های کامل ۲۸۳ موجودات یوکاریوتی نیز شناخته شده‌اند. این موارد شامل *Saccharomycescerevisiae* (مخمر نان)، *Caenorhabditiselegans* (نماتد)، *Drosophila melanogaster* (مگس میوه)، *Arabidopsis thaliana* (آرابیدوپسیس)، *Takifugurubripes* (ببر چین‌دار)، *Homo sapiens* (انسان)، و *Mus musculus* (موش) هستند. به علاوه، از سپتامبر ۲۰۱۶، ۱۳۰۰۰ طرح توالی‌یابی ژنوم یوکاریوتی دیگر نیز در حال انجام است. این داده‌ها در نهایت به رمزگشایی اسرار زیست‌شناسی مربوط می‌شود و بنابراین به غلبه بر بیماری‌های انسان و حیوانات کمک می‌کند.

---

[1] Virulent

[2] Nonvirulent

۷-۲ تحقیقات دارویی بر پروتئین هدف

تحقیق سیستماتیک در مورد ترکیبات فعال به عنوان داروهای جدید به نیمه دوم قرن ۱۹ بر می‌گردد. اولین مثال استیل سالیسیلیک اسید است، که در سال ۱۸۹۷ توسط دو شیمی‌دان فلیکس هافمن و آرتور ایخنگرون[۱] از شرکت بایر[۲] سنتز شد که هم اکنون با نام تجاری آسپرین در کل دنیا مشهور است. هنوز یک سؤال مشاجره برانگیز باقی‌مانده است که کدام دو شیمی‌دان مخترع اصلی سنتز استیل سالیسیلیک اسید بوده‌اند. با این وجود، این ماده اهمیت اقتصادی یا اهمیت علمی‌اش را از دست نداد. بسیاری از آنتی‌بیوتیک‌هاکه استفاده امروزی داشتند در نیمه اول قرن بیستم کشف شدند. اما از حدود دهه ۱۹۶۰، تعداد داروهای جدید به صورت ثابت کاهش یافته است. دلایلی برای این امر وجود دارد، که شامل: کاهش در میزان موفقیت غربال‌گری غیرهدف، هزینه‌های بیشتر برای تحقیق و همچنینافزایش سطح استانداردهای ایمنی لازم می‌باشد. به علاوه، در حوزه بیماری‌های عفونی، با ظهور و انتشار بیشتر مقاومت‌های دارویی شرایط بدتری پدیدار شد. ولی در همان زمان، حوزه جدیدی از تحقیقات مولکولی در سال ۱۹۵۳ با افشای ساختار سه بعدی مارپیچ دورشته‌ایی DNA[۳] توسط جیمز دی. واتسون و فرانسیس اچ. سی کریک[۴] شروع شد.

---

[3] Felix Hoffmann and Arthur Eichengrün

[2] Bayer

[3] DNA double helix

[4] James D. Watson and Francis H.C. Crick

شکل ۱-۷) شباهت یک نماد مسیحی به یک رویکرد مبتنی بر توسعه طراحی اختصاصی دارو.
تصویر سنت جورج را به عنوان یک شکارچی اژدها نشان می دهد. اژدها نماد موجود هدف است که تنها می‌تواند توسط یک ضربه دقیق به قلب (پروتئین هدف) کشته شود.و همه اهداف دیگر بی‌اهمیت‌اند. بر اساس این تصویر، سنت جورج (دانشمند) از اسب خود (ابزار علمی) برای هدایت نیزه (داروهای انتخابی) او به هدف استفاده می‌کند. تصویر اصلی در منطقه Monastery Preveli (کشور یونان)

با شروع توالی‌یابی کامل ژنوم و بررسی اطلاعات زیستی، رویکرد کشف دارو تغییر کرد. بنابراین، رویکرد مبتنی بر هدف (شکل ۱-۷)، اولین گام در شناسایی پروتئین‌هایی است که برای بقا موجود بیماری‌زا ضروری هستند. گام دوم یافتن مواد شیمیایی فعال است که بر

پروتئین هدف در مسیر شیمیایی اثر می‌گذارند. فقط پس از یافتن چنین مواد شیمیایی بهینه شده با طیف فعالیت مطلوب می‌توان از روش‌های درون شیشه‌ای[1] استفاده نمود که پس از آزمون قابل مصرف در سیستم زیستی می‌باشد. برای مثال، برای توسعه یک آنتی‌بیوتیک جدید، یک پیش نیاز ایده‌آل آن است که پروتئین هدف برای بقای باکتری بیماری‌زا لازم باشد و موجود میزبان نیز دارای پروتئین مشابه یا همان پروتئین نباشد تامورد هدف واقع شود، که با این حالت، به صورت بالقوه منجر به سمیت می‌شود. در این روش، تحلیل مقایسه‌ای کل ژنوم[2] برای شناسایی اهداف خاص بیمارگر[3] مناسب بودند. در واقع، این رویکرد توسط هوینن و همکاران (۱۹۹۸) بر روی ژنوم‌های سه باکتری، *Escherichia coli*، *Haemophilus influenza* و *Helicobacter pylori* اتخاذ شدند. علاوه بر پروتئین‌های اختصاصی گونه، پروتئین‌های ارتولوگ یا در هر سه یا در دو تا از سه موجود شناسایی شدند. برای *H. pylori*، عامل بیماری زخم معده یا اثنی‌عشر[4]، محققین پیش‌بینی کردند که ۱۲۳ پروتئین در برهم‌کنش بین بیمارگر و میزبان دخیل هستند، یعنی همان پروتئین‌هایی که اهداف احتمالی برای توسعه آنتی‌بیوتیک هستند.

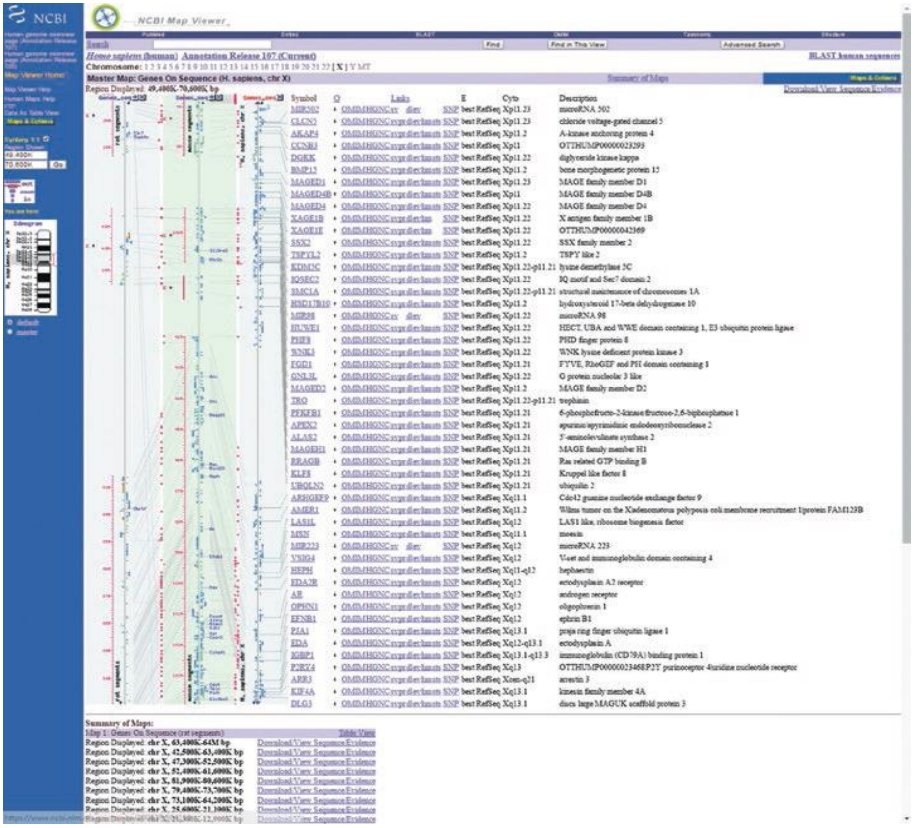
---

[1] *in vitro*

[2] Comparative Whole Genomic Analysis

[3] Pathogen

[4] Gastric and Duodenal ulcers



Homo sapiens (human) Annotation Release 107 (Current)

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 [X] Y MT

Master Map: Genes On Sequence (H. sapiens, chr X)

Region Displayed: 49,400K-70,600K bp

Symbol	O	Links	E	Cyto	Description
MIR502	•	OMIM HGNC sv dl ev SNP	best RefSeq	Xp11.23	microRNA 502
CLCN5	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.23	chloride voltage-gated channel 5
AKAP4	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.2	A-kinase anchoring protein 4
CCNB3	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11	OTTHUMP00000023293
DGKK	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	diglyceride kinase kappa
BMP15	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.2	bone morphogenetic protein 15
MAGED1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.23	MAGE family member D1
MAGED4B	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11	MAGE family member D4B
MAGED4	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	MAGE family member D4
XAGE1B	•	OMIM HGNC sv pr dl ev hm SNP	best RefSeq	Xp11.22	X antigen family member 1B
XAGE1E	•	OMIM HGNC sv pr dl ev hm SNP	best RefSeq	Xp11.22	OTTHUMP00000042369
SSX2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	SSX family member 2
TSPYL2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.2	TSPY like 2
KDM5C	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22-p11.21	lysine demethylase 5C
IQSEC2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	IQ motif and Sec7 domain 2
SMC1A	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22-p11.21	structural maintenance of chromosomes 1A
HSD17B10	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.2	hydroxysteroid 17-beta dehydrogenase 10
MIR98	•	OMIM HGNC sv dl ev SNP	best RefSeq	Xp11.22	microRNA 98
HUWE1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase
PHF8	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	PHD finger protein 8
WNK3	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	WNK lysine deficient protein kinase 3
FGD1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	FYVE, RhoGEF and PH domain containing 1
GNL3L	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22	G protein nucleolar 3 like
MAGED2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.2	MAGE family member D2
TRO	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.22-p11.21	trophinin
PFKFB1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
APEX2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	apurinic/apyrimidinic endodeoxyribonuclease 2
ALAS2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	5'-aminolevulinate synthase 2
MAGEH1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	MAGE family member H1
RRAGB	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	Ras related GTP binding B
KLF8	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	Kruppel like factor 8
UBQLN2	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xp11.21	ubiquilin 2
ARHGEF9	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq11.1	Cdc42 guanine nucleotide exchange factor 9
AMER1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq11.2	Wilms tumor on the X/adenomatous polyposis coli membrane recruitment 1/protein FAM123B
LAS1L	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	LAS1 like, ribosome biogenesis factor
MSN	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq11.1	moesin
MIR223	•	OMIM HGNC sv dl ev SNP	best RefSeq	Xq12	microRNA 223
VSIG4	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	V-set and immunoglobulin domain containing 4
HEPH	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq11-q12	hephaestin
EDA2R	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	ectodysplasin A2 receptor
AR	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	androgen receptor
OPHN1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	oligophrenin 1
EFNB1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12	ephrin B1
PJA1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq13.1	praja ring finger ubiquitin ligase 1
EDA	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq12-q13.1	ectodysplasin A
IGBP1	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq13.1-q13.3	immunoglobulin (CD79A) binding protein 1
P2RY4	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq13	OTTHUMP00000023468/P2Y purinoceptor 4/uridine nucleotide receptor
ARR3	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xcen-q21	arrestin 3
KIF4A	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq13.1	kinesin family member 4A
DLG3	•	OMIM HGNC sv pr dl ev hm ts SNP	best RefSeq	Xq13.1	discs large MAGUK scaffold protein 3

Summary of Maps:

Map 1: Genes On Sequence (rat segments) Table View

Region Displayed: chr X, 63,400K-64M bp Download/View Sequence/Evidence
Region Displayed: chr X, 42,500K-63,400K bp Download/View Sequence/Evidence
Region Displayed: chr X, 47,300K-52,500K bp Download/View Sequence/Evidence
Region Displayed: chr X, 52,400K-61,600K bp Download/View Sequence/Evidence
Region Displayed: chr X, 81,900K-80,600K bp Download/View Sequence/Evidence
Region Displayed: chr X, 79,400K-73,700K bp Download/View Sequence/Evidence
Region Displayed: chr X, 73,100K-64,200K bp Download/View Sequence/Evidence
Region Displayed: chr X, 25,600K-21,100K bp Download/View Sequence/Evidence
Region Displayed: chr X, 21,300K-12,900K bp Download/View Sequence/Evidence

شکل ۲-۷) نقشه کروموزوم X موش صحرایی، موش و انسان برگرفته از پایگاه NCBI
تصویر بیانگر بخشی از نقشه کروموزوم X است. ژن‌های سنتنیک در این منطقه کروموزوم با خطوط خاکستری نشان داده شده است

به دلیل افزایش تعداد ژنوم‌های توالی‌یابی شده باکتریایی، مشخص‌تر شده است که کدام ژن‌ها عموماً در بین باکتری‌ها محافظت شده‌[1] هستند و کدام ژن‌ها برای گونه باکتریایی، اختصاصی عمل می‌کنند. اما این امر همیشه آسان نیست، که بتوان آستانه شباهت توالی‌هایی را مشخص کردکه رویکرد کشف داروی هدف را به دلیل سمیت بالقوه حاصل از برهم‌کنش ناخواسته با پروتئین مشابه انسانی، متوقف کند. برای مثال، دی‌هیدروفولات ردوکتاز[2] باکتریایی

---

[1] Conserved

[2] Dihydrofolate reductase

شباهت ۲۸ درصدی در سطح آمینواسید با پروتئین انسانی دارد، با این وجود داروی ضدباکتریایی تریمتوپریم[۱]، به عنوان یک ممانعت‌کننده انتخابی در توالی‌های ارتولوگ باکتریایی استفاده می‌شود.

۷-۳ تحلیل‌های مقایسه‌ای ژنوم اطلاعاتی را در مورد زیست شناسی موجود فراهم می‌کند

تحلیل‌های مقایسه‌ای ژنوم اغلب "ژنومیکس مقایسه‌ای[۲]"نامیده می‌شوند، به این صورت که دو یا تعداد بیشتری ژنوم با یکدیگر مقایسه می‌شوند (Beckstette *et al.* 2004). هدف از این کار یافتن شباهت‌ها و تفاوت‌های بین این ژنوم‌هاست که اطلاعاتی در مورد زیست‌شناسی موجودات مورد نظر به همراه دارد. هدف مهم دیگر ژنومیکس مقایسه‌ای توصیف ساختار ژنوم و شناسایی نواحی کدکننده[۳] و غیرکدکننده[۴] است (Wei *et al.* 2002).

۷-۳-۱ ساختار ژنوم

تحلیل‌های ساختار یک یا چند ژنوم شامل: مقادیر آماری مانند اندازه و ترکیب نوکلئوتیدی، فراوانی کاربرد کدون و شناسایی نواحی محافظت‌شده بین دو یا چند ژنوم است. درصد و فراوانی محتوای گوانین و سیتوزین (GC) یا محتوای آدنین و تیمین (AT) بین گروه‌های موجودات متفاوت است و به نظر می‌رسد که به صورت قابل توجهی در مسیر تکامل از میکروارگانیسم تا موجودات چند سلولی متفاوت باشد. همچنین، کاربرد کدون برای رمزگذاری آمینواسیدهای یکسان در موجودات مختلف یکسان نیست (فصل ۱ و ۳).

مطالعات مقایسه‌ای فراوانی از ژنوم‌های انسانی و موش نشان داده است که سازماندهی آن‌ها، تا حدود زیادی، مشابه است. این مورد نشان می‌دهد که از زمان آخرین جد مشترک، سازماندهی ساختاری ژنوم، محافظت‌شده است. برای تشریح تشابهات بین قطعات کروموزومی مرتبط از نظر تکاملی بین گونه‌ها، واژه‌های مختلفی تعریف شده‌اند. اگر دو یا چند ژن در

---

[1] Trimethoprim

[2] Comparative Genomics

[3] Coding

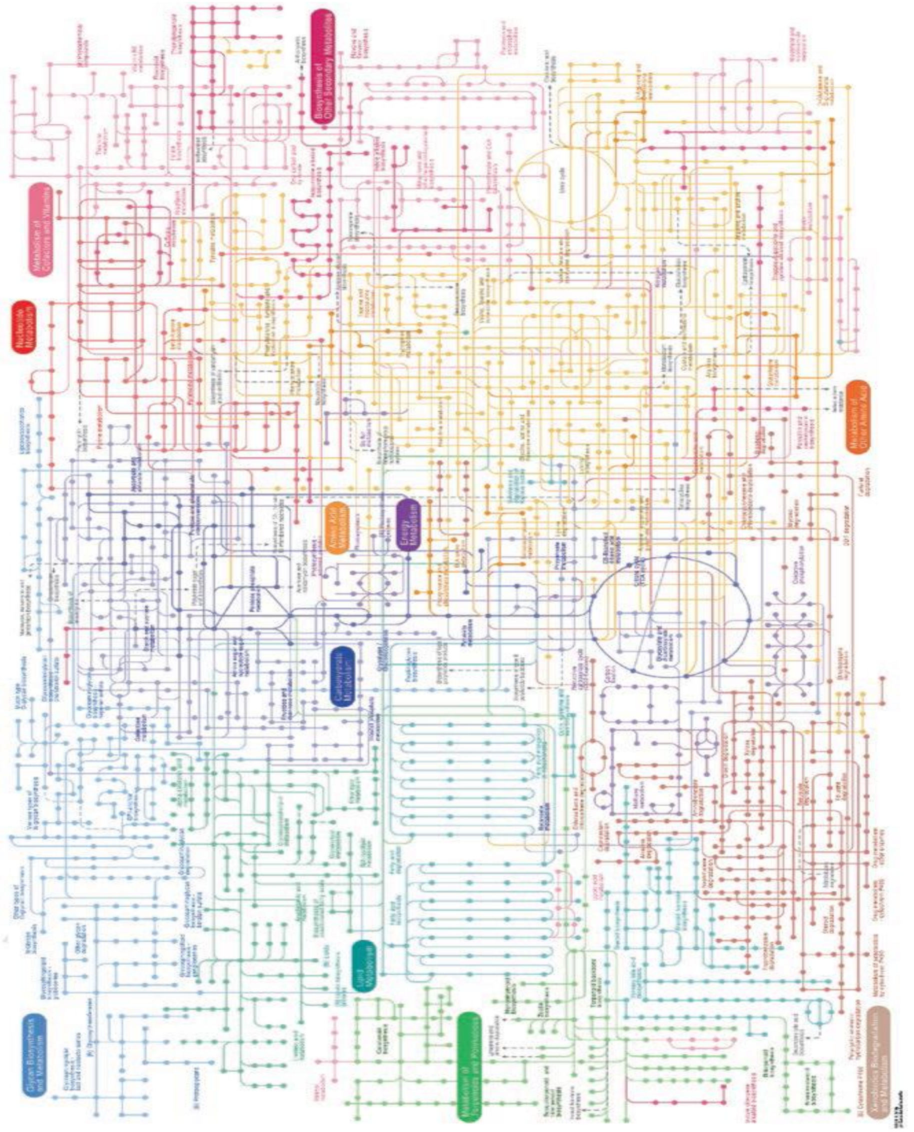
[4] Noncoding

کروموزوم مشابهی قرار گیرد، صحبت از ژن‌های سینتنیک[1] یا سینتنی[2] است. این تعریف فقط در یک گونه به کار می‌رود. اما وقتی ژن‌های سینتنی پروتئین‌های ارتولوگ، روی یک کروموزوم در بین گونه‌ها محافظت شده باشند، اصطلاح سینتنی محافظت‌شده[3] به کار می‌رود که در این حالت ترتیب ژن‌ها روی کروموزوم در نظر گرفته نمی‌شود (شکل ۲-۷). اما، اگر ترتیب ژن‌های روی کروموزوم هم محافظت شده باشد، با عنوان قطعات یا لینکاژهای محافظت‌شده[4] نامیده می‌شوند.

با توجه به تعداد در حال رشد توالی‌های کامل ژنوم یوکاریوتی، مشخص شده است که قطعات محافظت‌شده در همه پستانداران حضور دارند. اگرچه، نواحی سینتنی مختلفی بین گونه‌هایی مانند انسان‌ها و ماهی پف‌دار[5] مشاهده شده است که حدود ۴۵۰ میلیون سال قبل از هم جدا شده‌اند، ولی تاکنون هیچگونه بلوک ژنومی محافظت‌شده برای چنین موجوداتی با خویشاوندی دور، تشریح نشده است (Frazer *et al.* 2003).

---

[1] Syntenic
[2] Synteny
[3] Conserved Synteny
[4] Conserved Linkages or Segments
[5] Puffer Fish



شکل ۳-۷) نقشه مسیرهای متابولیسم در پایگاه داده KEGG

۷-۳-۲ نواحی کدکننده

تحلیل مقایسه‌ای نواحی کدکننده بین ژنوم‌های مختلف نه تنها شناسایی نواحی کدکننده پروتئین رادربرمی‌گیرد می‌شود بلکه مقایسه مستقیم انواع پروتئین‌های ارتولوگ و پارالوگ را نیز شامل می‌شود. شناسایی ژن‌ها در پروکاریوت‌ها به صورت مقایسه‌ای بسیار ساده است زیرا نواحی غیرکدکننده نسبتاً کمی‌وجود دارند. در حالت طبیعی ۸۵ درصد ژنوم باکتریایی، پروتئین‌ها یا RNAهایی را فعال می‌کنند که واحدهای تنظیمی‌یا نواحی غیرکدکننده را رمز می‌کنند. در مقابل، پیش‌بینی ژن‌ها در یوکاریوت‌ها بسیار دشوارتر است زیرا نواحی غیرکدکننده از نظر تعداد نسبت به دوره اولیه تکامل افزایش یافته است. ژنوم‌های یوکاریوتی حاوی تعداد فراوانی از نواحی برون ژنومی‌همچنین تکرارهای غیرکدکننده فراوانی است. به علاوه، ژن‌های یوکاریوتی حاوی اینترون‌ها و اگزون‌ها هستند، و پروتئین‌های مختلف اغلب در نتیجه اتصال متناوب حاصل می‌شوند (فصول ۱ و ۴). برای مثال، ژنوم پروکاریوت *Escherichia coli* حدود ۴۳۰۰ ژن در اندازه ژنوم ۴۶۰۰ کیلوباز (kb) دارد، که به صورت میانگین یک ژن برای هر کیلوباز طول دارد. در مقابل، ژنوم مخمر تک سلولی یوکاریوتی *Saccharomyces cerevisiae* حدود ۶۳۰۰ ژن در اندازه ژنوم ۱۲۰۰۰ کیلوبازی دارد، و ژنوم کرم چند سلولی *Caenorhabditiselegans* حاوی حدود ۱۹۰۰۰ ژن در اندازه ژنوم ۹۷۰۰۰ کیلوباز است. از نظر فیلوژنتیکی، ژنوم انسان بسیار جوان است و تفاوت بزرگی بین تعداد ژن‌ها و اندازه ژنوم نشان می‌دهد که حدود ۱۹۰۰۰ تا ۲۰۰۰۰ ژن در اندازه حدود ۳/۳ گیگاباز است (Ezkurdia *et al.* 2014). ارتباط مشخصی بین اندازه ژنوم و پیچیدگی یک موجود وجود ندارد، که این امر را می‌توان با در نظر گرفتن تعداد مشابه ژن‌ها در ژنوم *C. elegans* و انسان‌ها بیان نمود. تعداد نسبتاً کم ژن‌های کدکننده پروتئین در ژنوم انسان با در نظر گرفتن تغییرات پس از ترجمه مانند اتصال متناوب بیان می‌کند که یک ژن می‌تواند چندین پروتئین را کدکند.

۷-۳-۳ نواحی غیرکدکننده

تحلیل مقایسه‌ای نواحی غیرکدکننده، که در انسان‌ها و سایر پستانداران بیش از ۹۷ درصد ژنوم را شامل می‌شود، هنوز یکی از بزرگترین چالش‌های بیوانفورماتیک بشمار می‌آید. امروزه، تحلیل ژنوم توجه بیشتری رادر شناسایی واحدهای تنظیمی‌ژنوم به خود معطوف کرده است.

برای مثال، در بیوانفورماتیک نشان داده شده است که نواحی غیرکدکننده محافظت‌شده، دارای مناطق اتصالی فاکتورهای رونویسی می‌باشند. به علاوه، احتمال شناسایی چنین نواحی تنظیمی‌در نواحی غیرکدکننده وقتی افزایش می‌یابد که بیش از دو ژنوم موجودات کاملاً نزدیک مقایسه شود. اخیراً نشان داده شده است که نیمی‌از نواحی غیرکدکننده در ژنوم سگ نیز در مقایسه با ژنوم انسان و موش، محافظت شده است.

۷-٤ تحلیل مقایسه‌ای متابولیکی

برای پیش‌بینی ژن[۱]، تأکید خاص بر ژن‌هایی است که پروتئین‌هایی را کد می‌کنند که در متابولیسم دخیل باشند. با استفاده از پیش‌بینی ژن، امکان تعیین مسیرهای متناوب متابولیکی برای تولید انرژی وجود دارد. امروزه برای شناسایی اهداف متابولیکی، مقایسه دو یا چند ژنوم در سطح مسیرهای متابولیکی صورت می‌پذیرد. این مورد درموجودات پروکاریوتی مؤثرتر است زیرا بسیاری از ژنوم‌ها توالی‌یابی شده‌اند. امروزه از نرم‌افزارهایمتعددی استفاده می‌شوند تا متابولوم‌ها را مقایسه کنیم: دایره‌المعارف ژن‌ها و متابولیسم *Escherichia coli* (EcoCyc) [ecocyc]، دایره‌المعارف کیوتو ژن‌ها و ژنوم‌ها (KEGG)[۲][kegg](شکل ۳-۷)، و پایگاه داده Reactome[reactome] در بین بهترین‌ها شناخته شده‌اند.

روش‌های مورد استفاده در پایگاه‌های ذکر شده در بالا شامل تحلیل‌های دستی و نیمه خودکار است. تاکنون، هیچ تحلیل کاملاً خودکار که بتواند همه مسیرهای متابولیکی را محاسبه کند وجود ندارد. به علاوه، چنین پایگاه‌های داده‌ای همیشه کامل نیستند. در حالی که ابتدا پایگاه‌های داده با مسیرهای متابولیکی سر و کار دارند، در طول زمان مکانیزم‌های تنظیمی مختلفی مانند انتقالات غشایی، تنظیم ژن و انتقال پیام نیز در این پایگاه‌ها گنجانده شدند (شکل ۴-۷).

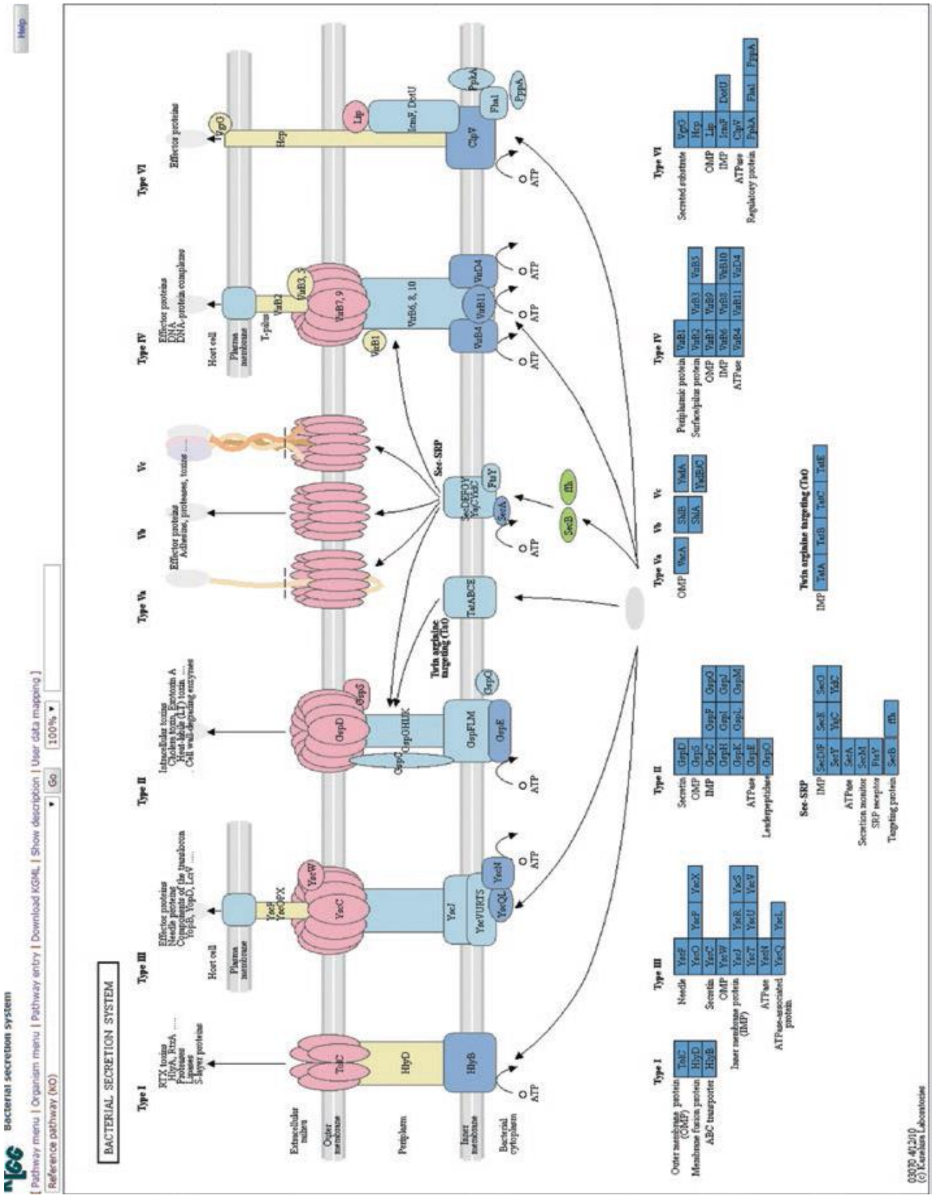
در ژنوم‌ها، ژن‌ها یا پروتئین‌های توالی‌یابی شده را می‌توان به گروه‌های ارتولوگ تقسیم کرد. طبق آن، پروتئین‌هایی که حاضر یا غایب هستند را می‌توان به صورت سیستماتیکی شناسایی کرد و مسیرهای متابولیکی آن‌ها را طراحی نمود. طی تحلیل صورت گرفته در ژنوم باکتری *Helicobacter pylori* متوجه شدند عامل زخم‌معده نه گلیکولیز و نه متابولیسم پنتوز

---

[1] Gene Prediction

[2] Kyoto Encyclopedia of Genes and Genomes (KEGG)

فسفات است بلکه به دلیل غیاب آنزیم‌های لازم عملکردی می‌باشد. زیرا هر دو مسیر متابولیکی پروتون‌ها را تولید می‌کنند، پس pH پایین‌تر، منجر به عملکرد بهتر باکتری در محیط اسیدی شکم یک موجود می‌شود. در مقابل، ژن‌های کدکننده پروتئین‌های متابولیزه‌کننده اسیدهای ارگانیک، مانند گلوکونئوژنز آنابولیک، حضور دارند. بنابراین، تولید انرژی *H. pylori* با تجزیه آمینواسید تأمین می‌شود، و سوبسترای لازم مستقیماً از مسیر گوارشی به دست می‌آیند.



شکل ۴-۷) طرح مسیرهای ترشح باکتریایی در پایگاه KEGG

برای یافتن مسیرهای متابولیکی خاص در پایگاه KEGG، ژنوم باید با ژنوم مرجع[۱] مقایسه شود. اگر در پایگاه، ژنی وجود داشته باشد، با رنگ خاصی مشخص می‌شود. مستطیل‌های رنگی مسیر متابولیکی خاصی را در موجود مورد مطالعه نشان می‌دهند (شکل ۵-۷). برای موفقیت در یافتن مسیرهای متابولیکی باید همه فاکتورهای جایگزین شناخته شوند. در اغلب موارد مسیرهای متابولیکی همه ژن‌ها یا پروتئین‌ها را نشان نمی‌دهد، و ناقص هستند. دلایل این است که همه ژن‌ها یا بعضی از آن‌ها درست پیش‌بینی نشده‌اند یا دانش کنونی در مورد مسیر متابولیکی خاص محدود است. همچنین این امکان هست که یک پروتئین چند عمل را انجام دهد، و بنابراین طیف متابولیکی بزرگتری نسبت به انتظار اولیه داشته باشد. در نتیجه مسیرهای متابولیکی متناوبی که به نتیجه زیستی یکسانی ختم می‌شوند را نمی‌توان حذف نمود.

۷-٤-۱ دایره‌المعارف کیوتو ژن‌ها و ژنوم‌ها

KEGG محصول شبکه ژنوم ژاپنی‌هاست و به صورت گسترده برای تحلیل مسیرهای متابولیکی به کار می‌رود. دو تا از سه پایگاه داده اصلی، PATHWAY و LIGAND، با فرایندهای متابولیکی در سلول‌ها و موجودات کار می‌کنند. سومین پایگاه داده، GENE، حاوی اطلاعات ژن و پروتئین است و با سایر پایگاه‌های داده اولیه قابل مقایسه است (Kanehisa *et al.* 2016). این پایگاه‌های داده با پایگاه BRITE تکمیل می‌شوند، که یک پایگاه داده انتولوژی[۲] برای تشریح روابط زیستی درون مسیرها می‌باشد. به علاوه، KEGG اطلاعاتی را در مورد داده‌های تجربی حاصل از بیان ژن مخمر ارائه می‌دهد (EXPRESSION). پایگاه داده دیگر، SSDB، حاوی اطلاعاتی در مورد گروه‌های پروتئین‌های ارتولوگ است.

جالب‌ترین پایگاه‌های داده بدون شک PATHWAY و LIGAND است. PATHWAY دارای شکل‌های گرافیکی از مسیرهای متابولیکی تعدادی از موجودات مختلف است، که بیشتر مربوط به پروکاریوت‌هاست ولی یوکاریوت‌ها نیزحضور دارند. نمایه‌های مسیرهای متابولیکی مشابه جدول مسیرهای بیوشیمیایی از Boehringer Mannheim است. نقشه‌های جداگانه را می‌توان از فهرست یا جدول تنظیم شده طبق مسیرهای متابولیکی اصلی انتخاب کرد (شکل

---

[1] Reference Genome

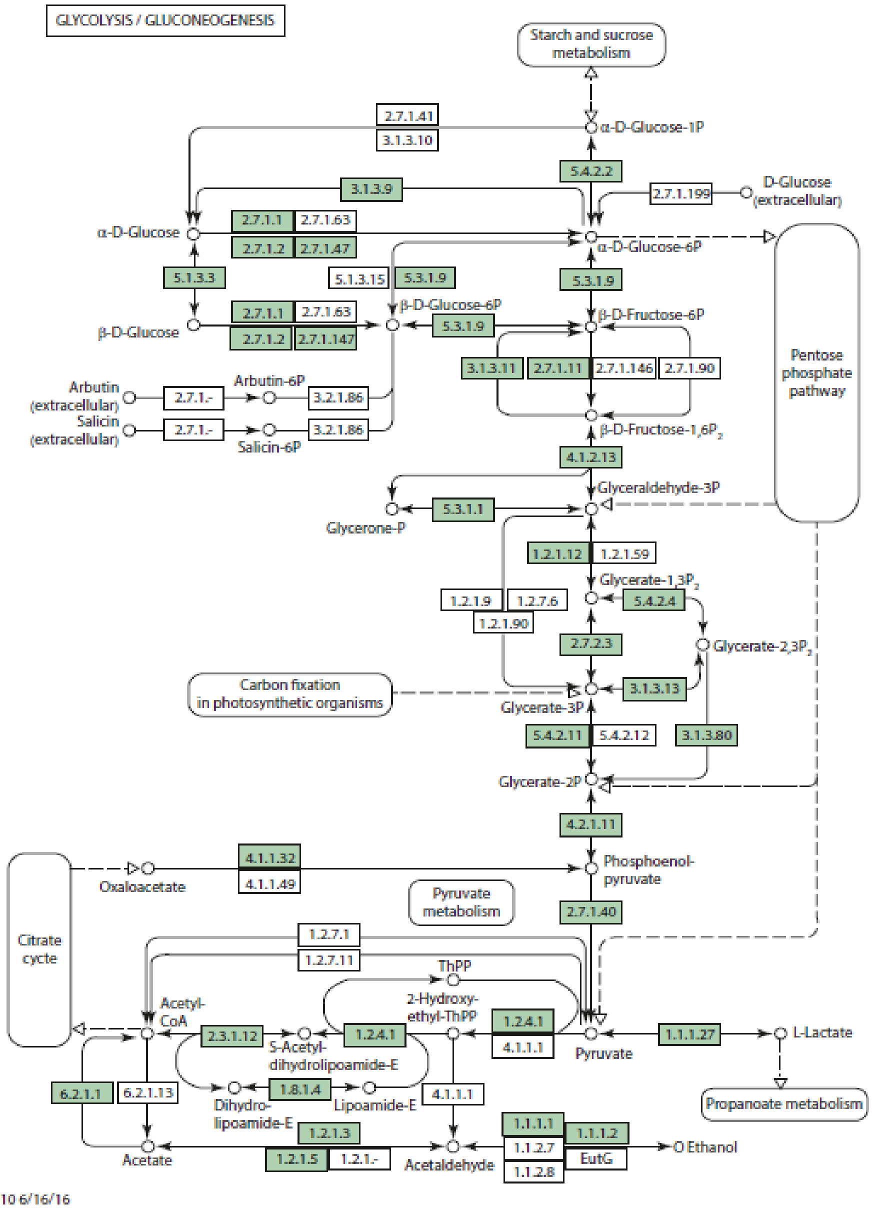
[2] Ontology

۳-۷). آنزیم‌های شناخته شده در مسیرهای مرجع به صورت رنگی سایه زده می‌شود. این ویژگی، مقایسه مسیرهای متابولیکی را بین موجودات ممکن می‌سازد. شکل ۵-۷ مثالی از متابولیسم گلیکولیز یا گلوکونئوژنز را در انسان نشان می‌دهد. آنزیم‌هایی که به رنگ سبز (جعبه‌های کوچک) هستند در ژنوم انسان وجود دارند. جداول متابولیکی منفرد روی سرور KEGG با پایگاه داده LIGAND با یکدیگر مرتبط شده‌اند، که این سرور یک پایگاه داده شیمیایی حاوی مواد، آنزیم‌ها و واکنش‌های مربوطه در مسیر متابولیکی است. برای ارجاع متقابل، جعبه‌های مستطیل کوچک با یک شماره آنزیم (طبقه‌بندی آنزیم، NC-IUBMB 1992) قرار گرفته است. تعداد EC شامل چهار بلوک از شماره‌هاست، که هر کدام با یک دوره از هم جدا شده‌اند. شماره اول یکی از شش گروه عملکردی آنزیمی را تشریح می‌کند (اکسیدوردوکتازها، ترانسفرازها، هیدرولازها، لیازها، ایزومرازها و لیگازها)، که دو بلوک بعدی به زیررده‌های درونی رده اصلی اشاره دارند. آخرین بلوک یک شماره از هر آنزیم در زیررده خاص است. مراجع متقابل دیگری توسط اشکال مدور کنار نام‌ها (مانند β-D- گلوکز) اشاره شده‌اند. مورد اخیر به پایگاه داده LIGAND مربوط نمی‌شود ولی به تشریح کامل مسیرهای متابولیکی مربوطه برمی‌گردد.

با کلیک بر روی دایره *Glycerate-1,3P2* پنجره جدیدی به پایگاه LIGAND باز می‌شود (شکل ۶-۷). علاوه بر شماره ماده منحصر به فرد، نام ماده و فرمول‌های ساده و ساختاری ماده ارائه می‌شود. آنچه در ادامه می‌آید مراجع متقابل به ورودی‌های واکنش‌ها است که در1,3-bisphospho-D-glycerate دخیل است مانند، مسیرهای متابولیکی، و آنزیم‌هایی که با تبدیل 1,3-bisphospho-D-glycerate مرتبط هستند. شماره CAS در بخش *DBLINKS* یک شماره اختصاصی است که به هر ماده شیمیایی توسط سرویس چکیده شیمیایی[1] بعد از اولین انتشار داده می‌شود [cas]. به علاوه، این قسمت با سایر پایگاه‌های داده نیز ارتباط برقرار می‌کند. بخش *Structure* حاوی ارائه گرافیکی ساختار شیمیایی و تعدادی از دکمه‌ها، که به فرد اجازه می‌دهد تا ساختار را به فرمت‌های مختلف بارگیری کنند.

---

[1] Chemical Abstract Service

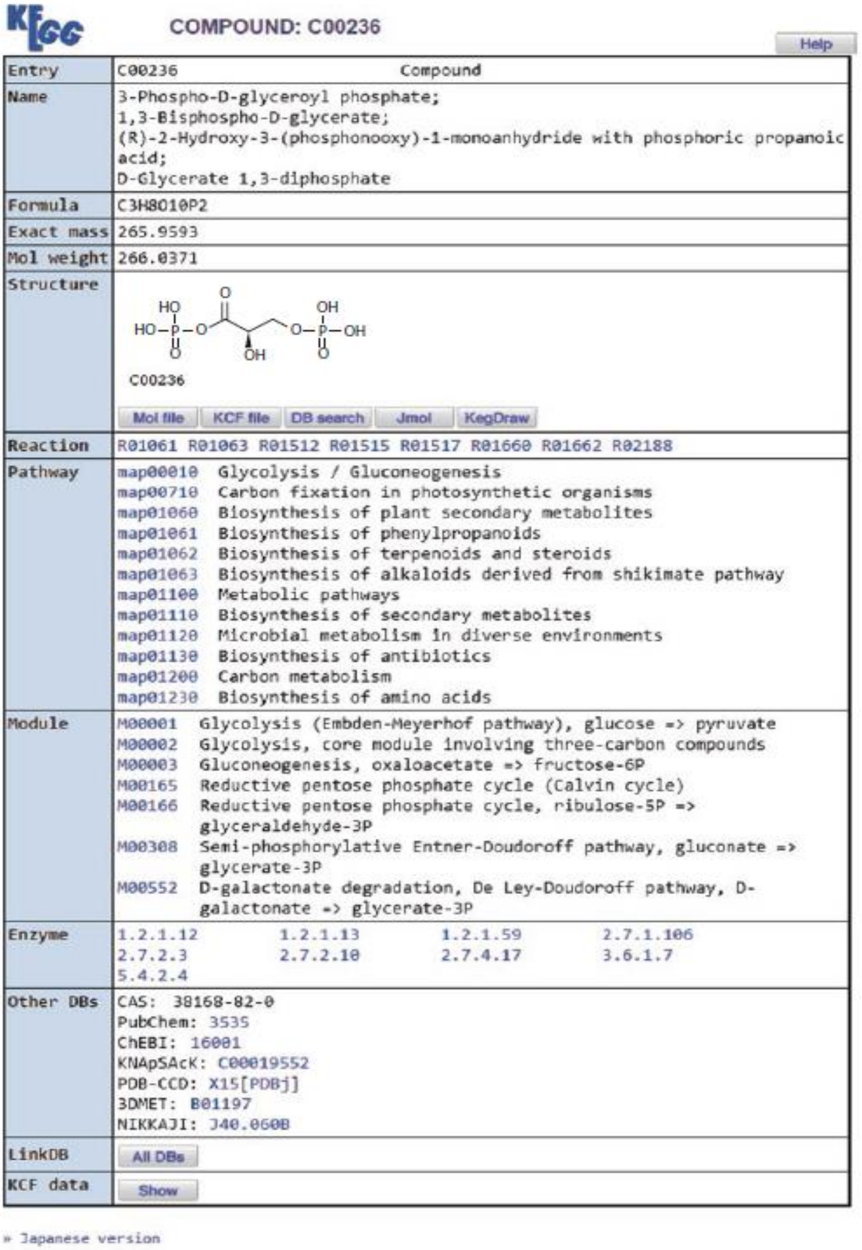


00010 6/16/16
(c) Kanehisa Laboratories

شکل ۵-۷) نقشه متابولیک متابولیسم گلیکولیک / گلوکونوژنیک

آنزیم‌های این مسیر متابولیکی که تا کنون در انسان شناخته شده‌اند، رنگی مشخص شده است

علاوه بر درخواست‌های پایگاه داده از طریق نمایه گرافیکی مسیرهای متابولیکی، پایگاهLIGAND جستجوی متن برای واکنش‌گرها یا آنزیم‌ها و جستجو برای ساختارهای شیمیایی پیچیده‌تر را نیز تسهیل می‌کند.



COMPOUND: C00236

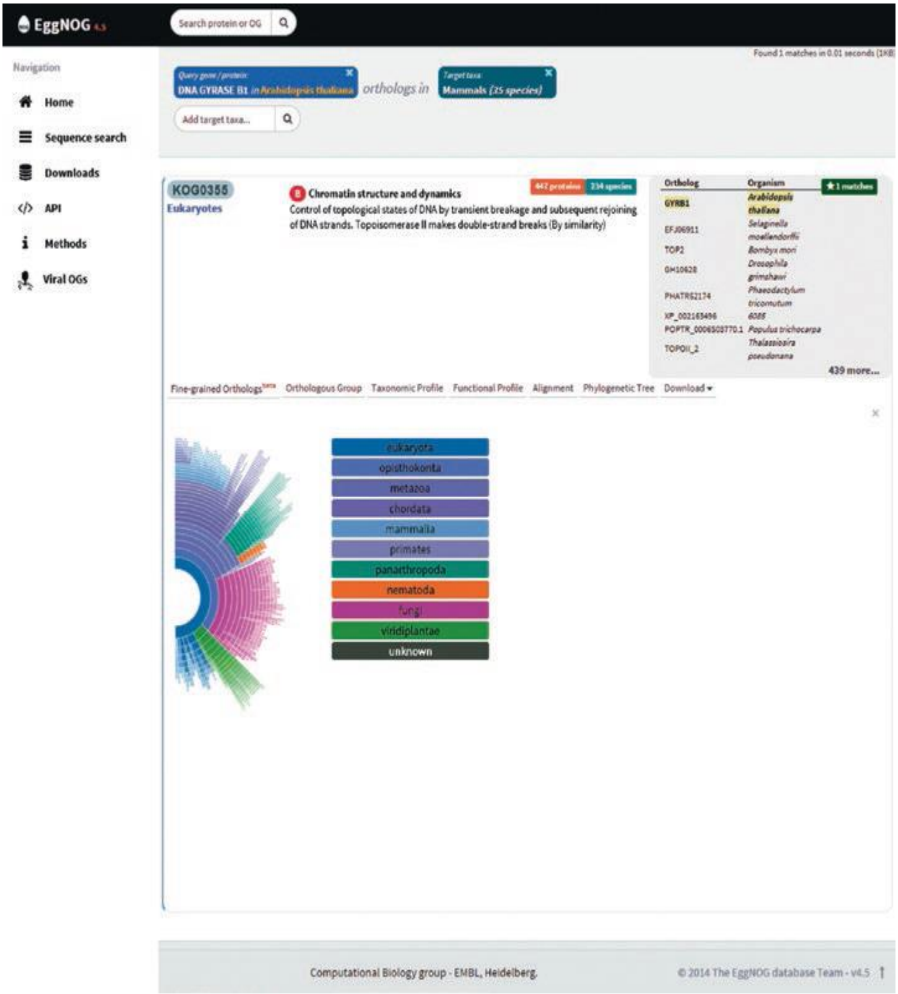
Entry	C00236 Compound
Name	3-Phospho-D-glyceroyl phosphate; 1,3-Bisphospho-D-glycerate; (R)-2-Hydroxy-3-(phosphonooxy)-1-monoanhydride with phosphoric propanoic acid; D-Glycerate 1,3-diphosphate
Formula	C3H8O10P2
Exact mass	265.9593
Mol weight	266.0371
Structure	C00236
Reaction	R01061 R01063 R01512 R01515 R01517 R01660 R01662 R02188
Pathway	map00010 Glycolysis / Gluconeogenesis map00710 Carbon fixation in photosynthetic organisms map01060 Biosynthesis of plant secondary metabolites map01061 Biosynthesis of phenylpropanoids map01062 Biosynthesis of terpenoids and steroids map01063 Biosynthesis of alkaloids derived from shikimate pathway map01100 Metabolic pathways map01110 Biosynthesis of secondary metabolites map01120 Microbial metabolism in diverse environments map01130 Biosynthesis of antibiotics map01200 Carbon metabolism map01230 Biosynthesis of amino acids
Module	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate M00002 Glycolysis, core module involving three-carbon compounds M00003 Gluconeogenesis, oxaloacetate => fructose-6P M00165 Reductive pentose phosphate cycle (Calvin cycle) M00166 Reductive pentose phosphate cycle, ribulose-5P => glyceraldehyde-3P M00308 Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P M00552 D-galactonate degradation, De Ley-Doudoroff pathway, D-galactonate => glycerate-3P
Enzyme	1.2.1.12 1.2.1.13 1.2.1.59 2.7.1.106 2.7.2.3 2.7.2.10 2.7.4.17 3.6.1.7 5.4.2.4
Other DBs	CAS: 38168-82-0 PubChem: 3535 ChEBI: 16001 KNApSAcK: C00019552 PDB-CCD: X15[PDBj] 3DMET: B01197 NIKKAJI: J40.060B
LinkDB	
KCF data	

» Japanese version

شکل ۶-۷) اطلاعات ثبت شده در پایگاه LIGAND برایβ-D-glucose

۵-۷ گروه‌های پروتئین‌های ارتولوگ

پس از تکمیل طرح توالی‌یابی ژنوم، توجه به تحلیل و طبقه‌بندی ژن‌های پیش‌بینی شده و عملکرد محصولاتشان، جلب شده است. ساده‌ترین رویکرد، مقایسه توالی‌های ژن ناشناخته با ژن‌های شناخته شده و استناد یک عملکرد بر اساس شباهت بین آن‌هاست است. به دلیل اینکه مقایسه کل ژنوم یا پروتئوم با روش‌های سنتی بسیار پرکاربرد است، نرم‌افزارهای تجاری ایجاد شده‌اند که مقایسه توالی‌های بزرگ و شناسایی توالی‌های عمومی، برای مثال MUMmer را اجازه می‌دهند (Delcher *et al*. 1999).



Found 1 matches in 0.01 seconds (1KB)

B Chromatin structure and dynamics

Control of topological states of DNA by transient breakage and subsequent rejoining of DNA strands. Topoisomerase II makes double-strand breaks (By similarity)

Ortholog	Organism
GYRB1	*Arabidopsis thaliana*
EFJ06911	*Selaginella moellendorffii*
TOP2	*Bombyx mori*
GH10628	*Drosophila grimshawi*
PHATRS2174	*Phaeodactylum tricornutum*
XP_002163496	*6085*
POPTR_0006S05770.1	*Populus trichocarpa*
TOPOII_2	*Thalassiosira pseudonana*

Computational Biology group - EMBL, Heidelberg.

© 2014 The EggNOG database Team - v4.5

شکل ۷-۷) نتیجه‌جستجوی پایگاه داده eggNOG

نمایه تاکسونومیکی برای ارتولوگ‌ها (برگرفته از پایگاه EMBL)

در مواردی که فواصل فیلوژنتیکی زیادی بین موجودات وجود دارد، مقایسه مستقیم توالی به دلیل شباهت اندک توالی، دشوارمی‌باشد. بنابراین، رویکرد دیگر برای طبقه‌بندی فیلوژنتیکی پروتئین‌ها، مقایسه ژن‌های ارتولوگ و پارالوگ است. ژن‌های ارتولوگ از طریق تشکیل گونه‌هایی خارج از یک جد مشترک ایجاد می‌شوند، ژن‌های پارالوگ از طریق تکثیر ژن ایجاد

می‌شوند. برداشت عمومی‌این است که عمل ژن‌های ارتولوگ نسبت به ژن‌های پارالوگ محافظت شده‌تر است زیرا فشار تکاملی روی ژن‌های پارالوگ بعد از تکثیر ژن کاهش می‌یابد. این مفهوم حدس ارتولو گ[1] نامیده می‌شود. اگرچه این مفهوم موردانتقاد واقع شده است (Studer and Robinson-Rechavi2009; Nehrt *et al.* 2011)، ولی هنوز معتبر است و اساس بیشتر روش‌های نام‌گذاری عملکردی را تشکیل می‌دهد (Huerta-Cepas *et al.* 2016). بنابراین تعیین دقیق ارتولوژی بین پروتئین‌ها، اهمیت زیادی دارد. متأسفانه، پیش‌بینی چنین روابطی هم از نظر تحلیلی و هم از نظر اطلاعاتی بسیار سخت است. این مورد دلایل فراوانیداردکه شامل: تکثیر آشیانه‌ای[2]، بازآرایی ژنوم و انتقال افقی ژن[3] می‌شود که روابط واقعی را می‌پوشاند.

امروزه چند سیستم پیچیده برای طبقه‌بندی پروتئین‌های ارتولوگ ایجاد شده است. یک سیستم شناخته شده پایگاه داده COP در NCBI است (خوشه‌های گروه‌های ارتولوگ) [cog][4] (Wheeler *et al.* 2007). علاوه بر جستجوی بر اساس متن، این امکان وجود دارد که توالی‌ها را در برابر پایگاه‌های داده مقایسه کنیم و عملکرد محصولات ژن را مقایسه کنیم. کیفیت نتایج در این روش به دلیل امکان تغییرات دستی بسیار بالا بود. ولی در پایگاه داده به دلیل اینکه یک سیستم پایدار است، فقط می‌تواند خوشه‌هایی از گونه‌ها را بسازد که قبلاً پیش‌بینی شده بودند و تحت تأثیر کاربر قرار نمی‌گرفتند. پایگاه داده COG در سال ۲۰۱۳ قطع شد.

پایگاه داده حاضر برای گروه‌های پروتئین ارتولوگ، پایگاه داده eggNOG است [eggnog]. این پایگاه داده حاوی خوشه‌های گروه‌های ارتولوگ در سطوح مختلف تاکسونومیکی به مراه نام‌گذاری کاربردی آن‌هاست. به علاوه، ورودی‌های پایگاه داده با ورودی‌های انتولوژی ژن (GO)، مسیرهای متابولیکی KEGG، و اطلاعات دامین‌های SMART/Pfam با یکدیگر مرتبط شده‌اند. در حال حاضر، این پایگاه داده حاوی ۲۰۳۱ موجود یوکاریوتی و پروکاریوتی است (نسخه v4.5, 2015). همچنین، تعداد ۱۶۵۵ موجود پروکاریوتی نیز در برابر این پایگاه داده پیش‌مقایسه شده‌اند. برای فرایند ساخت خوشه، داده‌ها

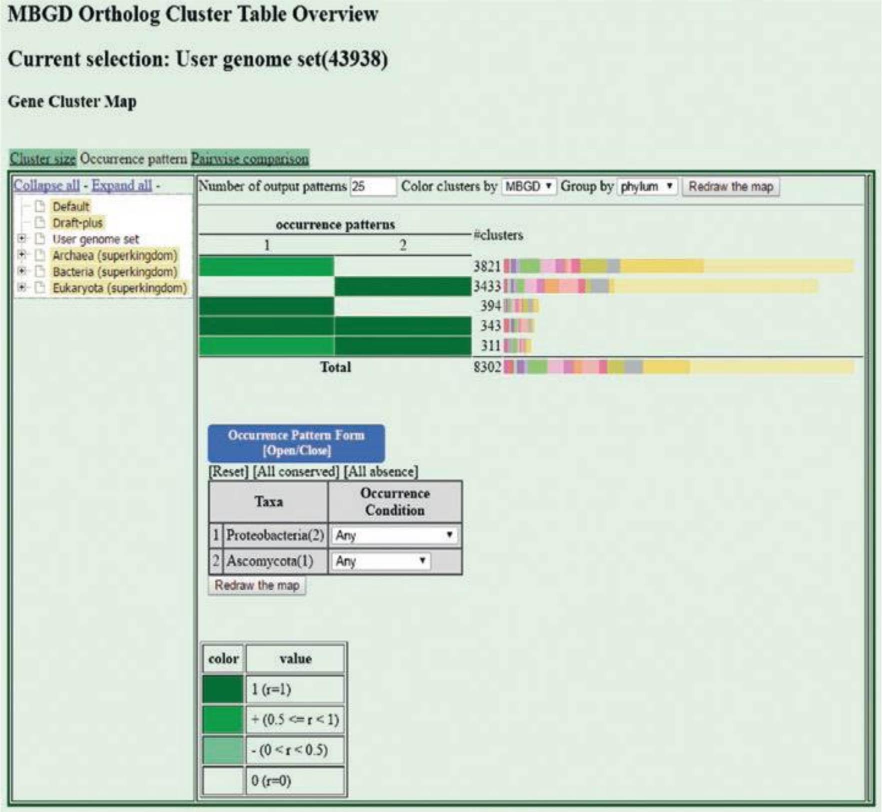
---

[1] Ortholog Conjecture
[2] Nested Duplications
[3] Horizontal Gene Transfers
[4] Clusters of Orthologous Groups

از چندین پایگاه داده اولیه استفاده شده‌اند، و پس از مرحله اطمینان کیفیت، همه توالی‌ها به صورت جفتی با همدیگر با استفاده از همردیفی اسمیت-واترمن[1]مورد مقایسه قرار گرفته‌اند. در این پایگاه خوشه‌هایی با توجه به خصوصیات تاکسونومیکی گروه‌بندی شده‌اند. اساس این خوشه‌بندی این است که گروه‌های ارتولوگ در سطوح تاکسونومیکی تحت بررسی قرار می‌گیرند. برای مثال، دسته‌بندی یک سری از توالی‌های پستانداران با بعضی از توالی‌های مهره‌داران خویشاوند دور در یک گروه واقع شده‌اند. دسته‌بندی eggnog که بر اساس خوشه استفاده شده در پایگاه داده COG برنامه‌ریزی شده است [cog]، سه سلسله یوکاریوت‌ها، باکتری‌ها و آرکئی‌ها را فهرست می‌کند: بخش COG حاوی همه سه سلسله با تمرکز بر پروکاریوت‌هاست، KOG حاوی یوکاریوت‌هاست و arKOG حاوی آرکی‌هاست. بنابراین دسته‌بندی به صورت مستقل برای هر کدام از سطوح تاکسونومی از پیش تعریف شده، انجام می‌شود. در نتیجه، تناقض‌هایی که از پروتئوم ناقص یا از فرض‌های الگوریتم‌ها به دست می‌آید، در گام اطمینان کیفیت حذف می‌شود. در گام آخر، یک روش خودکار استفاده می‌شود تا بهترین نام‌گذاری انطباق را از پایگاه‌های داده نام‌گذاری مختلف انتخاب کند. چنین نام‌گذاری برای انسان قابل فهم است ولی به سختی می‌توان آن‌ها را در یک تحلیل آماری به کار برد. بنابراین، پایگاه داده COG یک طبقه‌بندی تک حرفی را معرفی کرد که در پایگاه eggNOG استفاده می‌شود. هر گروه ارتولوگ برای قرارگیری در طبقه‌بندی از یک ماشین محاسباتی استفاده می‌کنند.

---

[1] Smith–Waterman Alignments


MBGD Ortholog Cluster Table Overview

Current selection: User genome set(43938)

Gene Cluster Map

Number of output patterns 25 Color clusters by MBGD Group by phylum

occurrence patterns 1	occurrence patterns 2	#clusters
🟩		3821
	🟩	3433
🟩		394
🟩	🟩	343
🟩	🟩	311
Total		8302

	Taxa	Occurrence Condition
1	Proteobacteria(2)	Any
2	Ascomycota(1)	Any

color	value
🟩	1 (r=1)
🟩	+ (0.5 <= r < 1)
🟩	- (0 < r < 0.5)
	0 (r=0)

شکل ۸-۷) نتایج تحلیل خوشه ای پایگاه MBGD

پایگاه eggNOG دو نوع سیستم جستجو را ارائه می‌دهد: یک سیستم بر اساس متن و یک سیستم بر اساس توالی می‌باشد. برای جستجوی متن، یک واژه جستجو مانند نام یک پروتئین یا ژن وارد می‌شود. اگر امکانات ورودی داده‌ها برای موجودات مختلف وجود داشته باشد، سیستم موجود هدف را پیدا می‌کند. سپس لیستی از موجودات هدف دیگری نیز ظاهر می‌گردد. این مورد می‌تواند به صورت موجود انفرادی یا همه اعضای دسته باشد. بر اساس فهرست موجود هدف، eggNOG سطح تاکسونومی را انتخاب می‌کند. نتایج حاوی نمایه‌های تصویری مختلف است، برای مثال، ممکن است نمایشی از درخت فیلوژنتیکی توالی‌ها باشد که از نشانگرهای رنگی برای منبع و موجودات هدف همچنین برای گونه‌ها و ژن‌ها استفاده می‌کند(شکل ۷-۷). به علاوه، امکان نمایش همردیفی توالی‌ها یا پروفایل تاکسونومیکی یا

عملکردی نیز وجود دارد. بخش تحلیل پایگاه شامل همه اعضای گروه ارتولوگ می‌شود، ولی گونه‌هایی که هدفجستجو نباشند، پنهان می‌مانند. همچنین، امکان نمایش جزئیات ارتولوگ‌های جفت‌شده نیز فراهم آمده است.

در بخش جستجوی توالی، امکان انتخاب فهرست موجودات هدف نمی‌باشد. در این بخش ابتدا یکی از سه سلسله موجود باید انتخاب شوند. سپس در فهرست نتایج، گروه‌های ارتولوگ حاصل را نشان می‌دهد. همانند جستجوی متن هر کدام از ورودی‌ها زیرلینک مشاهده یکسانی دارند.

سیستم مشابهی با نام پایگاه داده ژنوم میکروبی (MBGD)[1] محاسبه خوشه‌ها را طبق پارامترهای تعریف شده توسط کاربرد ارائه می‌کند [mbgd] (Uchiyama *et al*. 2015) (شکل ۸-۷). در این روش طبقه‌بندی پروتئین‌ها در خوشه‌های ارتولوگ به نوع موجودات بستگی دارد و یک مجموعهاطلاعات غیرقابل تغییر بر نتایج جستجو، اثر می‌گذارند. بنابراین پایگاه MBGD یک سامانه طبقه‌بندی را نسبت به نتایج ثابت طبقه‌بندی شده از قبل، فراهم می‌کند. نتایج محاسبه خوشه به پارامترهای ورودی توسط کاربر بستگی دارد، که از طریق ویژگی‌های ارتولوژی، همولوژی و یا بر اساس جداول شباهت از پیش محاسبه شده همه پروتئین‌ها در پایگاه داده عمل می‌کند. علاوه بر جستجو بر اساس متن، پایگاه MBGD ابزاری را برای ارزیابی و نام‌گذاری توالی توسط خود کاربر ارائه می‌دهد.

سایت‌های مفید

biochem-pathway. http://web.expasy.org/pathways/
cas. http://www.cas.org/
cog. http://www.ncbi.nlm.nih.gov/COG/
ecocyc. http://ecocyc.org/
eggnog. http://eggnog.embl.de/
enzym. http://www.chem.qmw.ac.uk/iubmb/enzyme/
genbank. http://www.ncbi.nlm.nih.gov/Genbank/
gold. https://gold.jgi.doe.gov/
kegg. http://www.kegg.jp/
mbgd. http://mbgd.genome.ad.jp/

---

[1] Microbial Genome Database (MBGD)

mummer. http://mummer.sourceforge.net/

reactome. http://www.reactome.org/

uniprotkb. http://www.uniprot.org/uniprot/

منابع

1. Beckstette M, Mailänder JT, Marhöfer RJ, Sczyrba A, Ohlebusch E, Giegerich R, Selzer PM (2004) Genlight: interactive high-throughput sequence analysis an comparative genomics. J Integr Bioinform Yearbook 2004:79–94.
2. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL (1999) Alignment of whole genomes. Nucleic Acids Res 27:2369–2376.
3. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866–5878.
4. Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.
5. Fraser CM, Gocayne JD, White O et al (1995) The minimal gene complement *Mycoplasma genitalium*. Science 270:397–403.
6. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1–12.
7. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293.
8. Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5.
9. Kanehisa M, Yoto S, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462.
10. NC-IUBMB (1992) Nomenclature Committee of the International Union of Biochemistry and molecular Biology, Enzyme Nomenclature 1992. Academic, Orlando.
11. Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7:e1002073.

12. Selzer PM, Brutsche S, Wiesner P, Schmid P, Müllner H (2000) Target-based drug discovery for the development of novel antiinfectives. Int J Med Microbiol 290:191–201.
13. Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ Trends Genet 25:210–216.
14. Uchiyama I, Mihara M, Nishide H, Chiba H (2015) MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res 43:D270–D276.
15. Wei L, Liu Y, Dubchak I, Shon J, Park J (2002) Comparative genomics approaches to study organism similarities and differences. J Biomed Inform 35:142–150.
16. Wheeler DL, Barrett T, Benson DA, Bryant SH et al (2007) Database resources of the National Center for Biotechnology. Nucleic Acids Res 35:D5–D12.