به نام خداوند جان و خرد

انتشارات
۷۹۰

تجزیه‌وتحلیل داده‌های RNA-seq

برای دریافت فایل‌های آموزش گام‌به‌گام، نرم‌افزارها و داده‌های مورد نیاز، به پروفایل کتاب در تارنمای انتشارات دانشگاه فردوسی مشهد به نشانی زیر مراجعه فرمایید:

press.um.ac.ir

دکتر علیرضا سیفی
عضو هیئت علمی دانشگاه فردوسی مشهد
محمدرضا رضائی

سرشناسه: سیفی، علیرضا، ۱۳۵۶-
عنوان و نام پدیدآور: تجزیه و تحلیل داده‌های RNA-seq / علیرضا سیفی، محمدرضا رضایی؛ ویراستار علمی امین میرشمسی‌کاخکی؛ ویراستار ادبی هانیه اسدپور فعال مشهد.
مشخصات نشر: مشهد: دانشگاه فردوسی مشهد، ۱۴۰۰.
مشخصات ظاهری: ۱۲۸ ص.: مصوّر، جدول، نمودار.
فروست: انتشارات دانشگاه فردوسی مشهد؛ ۷۹۰.
شابک: ISBN: 978-964-386-483-5
وضعیت فهرست‌نویسی: فاپا.
موضوع: آر. ان. ا. RNA
موضوع: نوکلئوتیدها -- توالی Nucleotide sequence
شناسه افزوده: رضایی، محمدرضا، ۱۳۷۳-
شناسه افزوده: میرشمسی کاخکی، امین، ۱۳۶۵-، ویراستار
شناسه افزوده: دانشگاه فردوسی مشهد، انتشارات.
رده‌بندی کنگره: QP۶۲۳
رده‌بندی دیویی: ۵۷۲/۸۸
شماره کتابشناسی ملی: ۷۶۱۴۲۶۹

تجزیه‌وتحلیل داده‌های RNA-seq

پدیدآورندگان: دکتر علیرضا سیفی؛ محمدرضا رضائی
ویراستار علمی: دکتر امین میرشمسی کاخکی
ویراستار ادبی: هانیه اسدپور فعال مشهد
مشخصات: وزیری، ۱۰۰ نسخه، چاپ دوم، زمستان ۱۴۰۴ (اوّل، ۱۴۰۰)
چاپ و صحافی: همیار
بها : ۲٫۳۰۰٫۰۰۰ ریال

حق چاپ برای انتشارات دانشگاه فردوسی مشهد محفوظ است.

انتشارات
۷۹۰

مراکز پخش:

فروشگاه و نمایشگاه کتاب پردیس: مشهد، میدان آزادی، دانشگاه فردوسی مشهد، جنب سلف یاس تلفن: ۳۸۸۰۲۶۶۶- ۳۸۸۳۳۷۲۷ (۰۵۱)
مؤسّسهٔ کتابیران: تهران، میدان انقلاب، خیابان کارگر جنوبی، بین روانمهر و وحید نظری، بن‌بست گشتاسب، پلاک ۸ تلفن: ۶۶۴۸۴۷۱۵ (۰۲۱)
مؤسّسهٔ دانشیران: تهران، خیابان انقلاب، خیابان منیری جاوید (اردیبهشت) نبش خیابان نظری، شمارهٔ ۱۴۲ تلفکس: ۶۶۴۰۰۲۲۰ – ۶۶۴۰۰۱۴۴ (۰۲۱)

http://press.um.ac.ir Email: press@um.ac.ir

فهرست مطالب

پیشگفتار

فنّاوری‌های نوین توالی‌یابی DNA که با عنوان متداول NGS شناخته می‌شوند، تحول شگرفی در رشته‌های مختلف علوم زیستی ایجاد کرده و امکان مطالعهٔ سریع‌تر و جامع‌تر اساس ژنتیکی و مولکولی پدیده‌های زیستی را فراهم آورده‌اند. هم‌راستا با موفقیت‌های چشمگیر سخت‌افزاری در فنّاوری‌های توالی‌یابی DNA، پیشرفت‌های اساسی نیز در تولید ابزارهای تجزیه‌وتحلیل داده‌های حاصل از این فنّاوری‌ها حاصل شده است.

چالش اصلی در تجزیه‌وتحلیل داده‌های NGS، حجم بالا و پیچیدگی این داده‌هاست که بهره‌برداری مناسب از آن ها نیازمند استفاده از ابزارهای بیوانفورماتیکی کارا و دقیق است. این ابزارهای بیوانفورماتیکی معمولاً توسط متخصّصان علوم ریاضی، کامپیوتر و بیوانفورماتیک طراحی می‌شوند و در غالب موارد بهره‌برداری از آن‌ها برای متخصّصان علوم زیستی که اطلاعات کافی از علوم کامپیوتری ندارند، چالش‌برانگیز است. معمولاً در مراکز تحقیقاتی پویا در دنیا یک یا چند متخصّص بیوانفورماتیک عهده‌دار تجزیه‌وتحلیل داده‌های NGS تولیدشده توسط محقّقان زیست‌شناسی هستند و بنابراین تمام محقّقان نیازمند فراگیری این روش‌های تجزیه‌وتحلیل نیستند. لیکن به‌دلیل اینکه هنوز توسعه‌یافتگی مطلوبی در تشکیل تیم‌های تحقیقاتی با تخصص‌های گوناگون در مراکز تحقیقاتی ایران رخ نداده است ، محقّقان علوم زیستی نیاز دارند که اشراف نسبی بر روش‌های تجزیه‌وتحلیل NGS داشته باشند.

مطالعهٔ ترنسکریپتوم (کل محتوای RNA سلول) با استفاده از روش‌های توالی‌یابی RNA (که با عنوان RNA-seq شناخته می‌شود) یکی از قدرتمندترین روش‌های مطالعهٔ پدیده‌های زیستی برای شناسایی مکانیسم‌های مولکولی و ژنتیکی کنترل‌کنندهٔ این پدیده‌هاست. ابزارهای بیوانفورماتیکی بسیار متنوعی برای تجزیه‌وتحلیل داده‌های RNA-seq ارائه شده است و ابزارهای جدید نیز با سرعت درحال توسعه و ارائه است. هدف از گردآوری کتاب حاضر فراهم‌آوردن مجموعه‌ای است برای محقّقان علوم زیستی که الزاماً اطلاعات گسترده‌ای از بیوانفوماتیک ندارند، ولی نیازمند تجزیه‌وتحلیل داده‌های RNA-seq هستند.

در این کتاب مباحث نظری موردنیاز درمورد RNA-seq و اصول پایه و مقدماتی کامپیوتر توضیح داده می‌شود. سپس روش‌های استاندارد تجزیه‌وتحلیل داده‌های RNA-seq که درحال‌حاضر بیشترین کاربرد را در بین پژوهشگران دارند، معرفی می‌شوند. با استفاده از داده‌های واقعی کل مسیرهای تجزیه‌وتحلیل به‌صورت گام‌به‌گام توضیح داده می‌شوند، به‌نحوی‌که مخاطب با دراختیارداشتن این کتاب بتواند کل مسیر تجزیه‌وتحلیل RNA-seq را روی کامپیوتر شخصی خود تمرین کند. کلّیهٔ نرم‌افزارها، داده‌های مورداستفاده

و خروجی‌های موردانتظار از اجرای هرکدام از مراحل تجزیه‌وتحلیل در قالب یک لوح فشرده همراه کتاب در اختیار خوانندگان قرار می‌گیرد. تلاش شده است که مطالب کتاب و لوح فشرده همراه کتاب به‌نحوی تنظیم شود که علاقه‌مندان به فراگیری روش‌های تجزیه‌وتحلیل RNA-seq با صرف حداقل زمان، اصول این روش‌ها را دریابند و بتوانند از آن‌ها در پروژه‌های پژوهشی خود استفاده کنند.

مزیت عمدهٔ این کتاب این است که ابزارهایی که در آن معرفی می‌شوند ابزارهای استاندارد، رایگان و با متن‌باز هستند. استفاده از این ابزارها به پرداخت هزینه و یا عدول از اصول اخلاقی مربوط به حقوق مالکیت معنوی نیاز ندارد. نکتهٔ مهم‌تر اینکه، چنانچه تجزیه وتحلیل با این ابزارها فراگرفته شود، استفاده از ابزارهای جدیدتری که به‌سرعت درحال توسعه و توزیع در جامعهٔ علمی هستند، بسیار ساده‌تر خواهد بود و پژوهشگر همواره به جدیدترین ابزارهای تجزیه‌وتحلیل RNA-seq دسترسی خواهد داشت.

هرچند تلاش زیادی شده است که اشکالات نگارشی و علمی در متن وجود نداشته باشد، بااین‌حال سپاسگزار خواهم بود چنانچه خوانندگان محترم اشکالات احتمالی و یا پیشنهادهای سازندهٔ خود را برای ارتقای کیفی کتاب به‌نحو مقتضی به این‌جانب منعکس کنند.

علیرضا سیفی
زمستان ۱۳۹۹

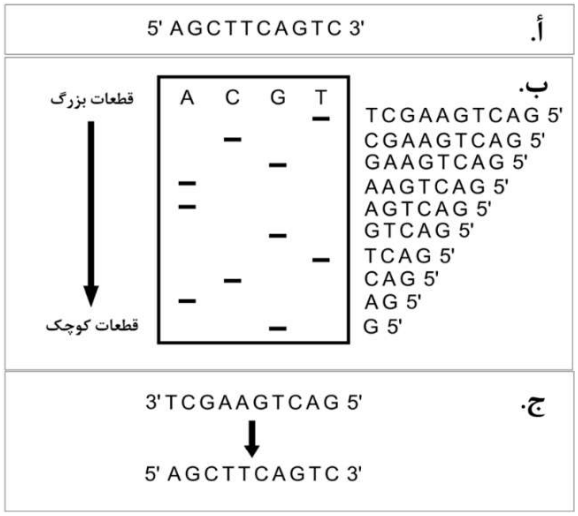
۱

تاریخچهٔ پیشرفت‌ها در ژنتیک، بیولوژی مولکولی و بیوانفورماتیک

ساختار دورشته‌ای DNA در سال ۱۹۵۳ کشف و مشخص شد که مادّهٔ وراثتی مارپیچی است دوگانه که از جفت شدن بازهای G با C و A با T ازطریق ایجاد پیوندهای هیدروژنی تشکیل شده است (۱، ۲). در سال ۱۹۷۳ اوّلین تلاش‌ها برای توالی‌یابی DNA به‌ثمر نشست و والتر گیلبرت[1] و آلن ماکسام[2] توانستند ۲۴ جفت باز از اپرون lac در باکتری *E. coli* را توالی‌یابی کنند (۳). در مدت چهار سال بعد روش کاراتری برای توالی‌یابی توسط فردریک سنگر[3] ارائه شد که به‌نام توالی‌یابی روش سنگر[4] شناخته می‌شود (۴). به‌دلیل اهمیت ابداع روش توالی‌یابی، نیمی از جایزهٔ نوبل در شیمی در سال ۱۹۸۰ به آقای سنگر و گیلبرت اهدا شد (نیمهٔ دیگر به آقای پاول برگ[5] برای معرفی فنّاوری DNA نوترکیب اهدا شد). توالی‌یابی روش سنگر نسل اوّل روش‌های توالی‌یابی درنظر گرفته می‌شود. توضیحات این روش در شکل ۱-۱ و انیمیشن ۱-۱ آمده است.

رشد چشمگیر پیشرفت‌ها در دنیای کامپیوتر نیز از دههٔ ۷۰ میلادی آغاز شد. تا قبل از سال ۱۹۷۰ کامپیوترها ابعاد و وزنی معادل یک یخچال خانگی داشتند و کار با آن‌ها دشوار بود. در سال ۱۹۷۷ اوّلین نسل از کامپیوترهای شخصی به بازار عرضه شد (Commodore PET، Apple II و TRS-80) که کار با آن‌ها برای کاربران بسیار آسان‌تر بود (۶). در سال ۱۹۷۹ با ابداع اوّلین نرم‌افزار آنالیز توالی‌یابی، برای اوّلین بار کامپیوتر در تحلیل داده‌های زیستی استفاده شد (۷).

1. Walter Gilbert
2. Allen Maxam
3. Fredrick Sanger
4. Sanger sequencing
5. Paul Berg



شکل ۱-۱ روش توالی‌یابی سنگر. أ. قطعهٔ DNA فرضی که نیاز است توالی‌یابی شود. ب. نوکلئوتیدهای خاتمه‌دهنده یا دای‌دئوکسی‌نوکلئوتیدها (ddNTPs) که با مادّهٔ فلوئورسنت یا رادیواکتیو علامت‌گذاری شده‌اند، درصورتی‌که به زنجیرهٔ درحال گسترش DNA اضافه شوند از ادامهٔ تکثیر قطعه جلوگیری می‌کنند. چهار واکنش تکثیر DNA هرکدام شامل یکی از چهار ddNTP مختلف تهیه می‌شود. در هرکدام از چهار واکنش، تکثیر قطعه تا زمانی‌که آنزیم DNA پلیمراز ddNTP را به توالی اضافه کند، ادامه می‌یابد. هر چهار واکنش روی ژل پلی‌آکریل‌آمید الکتروفورز می‌شوند و براساس اندازهٔ باندها، توالی به‌دست می‌آید. کوچک‌ترین باند که در پایین‌ترین نقطه قرار گرفته است، به‌عنوان اوّلین باز درنظر گرفته می‌شود و به‌همین‌ترتیب از باند کوچک به بزرگ توالی کامل به‌دست می‌آید. ج. توالی به‌دست‌آمده درواقع توالی دای‌دئوکسی‌نوکلئوتیدها می‌باشد که هنگام تکثیر، روبه‌روی رشتهٔ اصلی قرار گرفته‌اند. درنتیجه رشتهٔ مکمّل آن، توالی مدنظر ماست (۵).

با توجه به سرعت روزافزون تحقیقات ژنتیک مولکولی در آن سال‌ها، آزمایشگاه بیولوژی مولکولی اروپایی در سال ۱۹۷۴ پایه‌گذاری شد که به‌اختصار EMBL[1] خوانده می‌شود. مدتی بعد در ایالات‌متحده نیز اقداماتی برای ایجاد پایگاه اطلاعات ژنتیکی با حمایت مؤسسهٔ بهداشت ملّی ایالات‌متحده (NIH) انجام شد و در سال ۱۹۸۲ GenBank ایجاد گردید. تحول بسیار مهم دیگر ابداع الگوریتم مقایسهٔ توالی‌های DNA با کل توالی‌های موجود در GenBank بود. این مهم به‌همّت جان ویلبور[2] و دیوید لیپمن[3] محقّق شد و در سال ۱۹۸۳ الگوریتم BLAST[4] ارائه شد. ازآن پس این امکان برای محقّقان به وجود آمد که هر توالی DNA موردنظر را با کل توالی‌های موجود در بانک ژن مقایسه کنند.

یکی از نقاط عطف تاریخچهٔ تحولات بیولوژی مولکولی کشف واکنش زنجیره‌ای پلیمراز بود. در سال ۱۹۸۳ کری مولیس[5] پی برد که می‌توان یک قطعهٔ DNA را در آزمایشگاه به میلیون‌ها نسخه تکثیر کرد. این کشف بسیار مهم جایزهٔ نوبل سال ۱۹۹۳ را برای وی به ارمغان آورد.

1. European Molecular Biology Laboratories
2. John Wilbur
3. David Lipman
4. Basic Local Alignment Search Tool
5. Kary B. Mullis

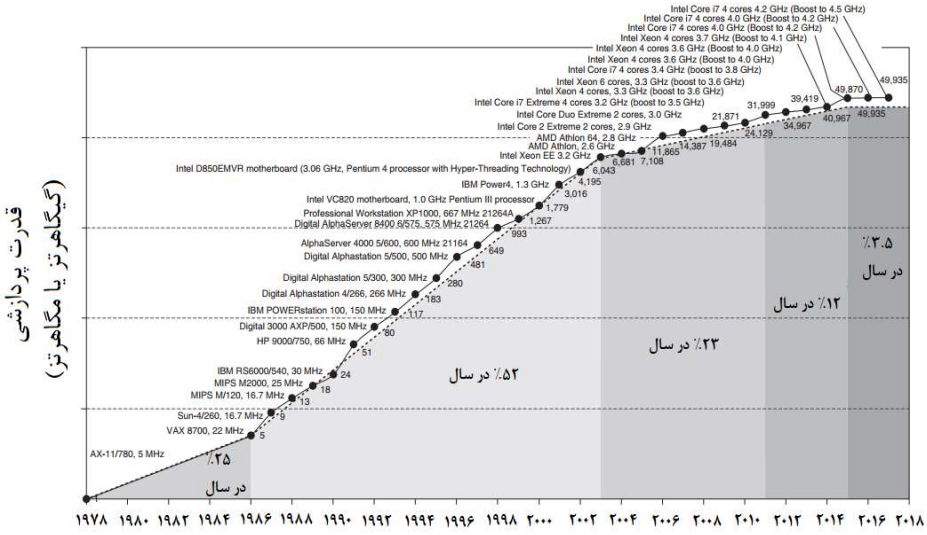
در سال ۱۹۸۴ گروه کامپیوتر و ژنتیک دانشگاه ویسکانسین[1] مجموعهٔ نرم‌افزاری به‌نام GCG را ارائه کردند که ۱۳۰ قابلیت مختلف برای آنالیز توالی‌های DNA، RNA و پروتئین داشت (۸). در همین سال‌ها بود که ریچارد استالمن[2] بیانیهٔ نرم‌افزار آزاد را منتشر کرد. فلسفهٔ نرم‌افزار آزاد ، منشأ چند ابتکار کلیدی در بیوانفورماتیک شد. ازجمله الهام‌بخش مجموعهٔ نرم‌افزاری متن‌باز زیست‌شناسی مولکولی اروپا شد که در ۱۹۹۶ توسه یافته و جایگزین رایگانی برای GCG گردید. این پیشرفت در گسترش نرم‌افزارها، بدون توسعهٔ زبان‌های برنامه‌نویسی مناسب میسر نبود. در میانهٔ دههٔ ۱۹۸۰، زبان‌های برنامه‌نویسی متعددی هم‌زمان با گسترش روزافزون کامپیوتر های رومیزی، منتشر شدند که امروزه نیز هنوز در دنیای بیوانفورماتیک مطرح هستند. با ابداع زبان‌های برنامه‌نویسی جدیدتر ازجمله پرل[3] و پایتون[4] به‌طور مؤثری پیچیدگی‌ها و مشکلات زبان‌های برنامه‌نویسی قبلی رفع شد و روند توسعهٔ نرم‌افزارهای بیوانفورماتیک تسریع شد (۶). استفاده از کامپیوتر برای مدل‌سازی و شبیه‌سازی فرایندهای زیستی تحت‌عنوان زیست‌شناسی مبتنی‌بر کامپیوتر[5] شناخته می‌شود که زمینهٔ تحقیقاتی متفاوت از بیوانفورماتیک است. لازم به ذکر است که آمار زیستی[6] نیز مبحث متفاوتی است که در آن با استفاده از روش‌ها و ابزارهای آماری، آزمایشات زیستی طراحی شده و داده‌های حاصل از این آزمایشات تجزیه‌وتحلیل می‌شود.

در سال ۱۹۸۸ مرکز ملّی اطلاعات زیست‌فنّاوری[7] (NCBI) پایه‌گذاری شد و در سال ۱۹۹۲ بانک ژن به NCBI انتقال یافت. با گسترش اینترنت در دههٔ ۱۹۹۰ امکان استفاده از پایگاه‌های اطلاعاتی و ثبت توالی‌های جدید در این پایگاه‌ها تسریع شد.

هرچند در دههٔ ۱۹۹۰ نرم‌افزارهای زیادی برای آنالیزهای بیوانفورماتیک ارائه شدند، ولی به‌دلیل سرعت پردازش پایین کامپیوترها در آن زمان، آنالیزهای بیوانفورماتیکی با صرف وقت زیاد قابل انجام بود. در فاصلهٔ سال‌های ۱۹۸۶ تا ۲۰۰۳ به‌طور متوسط قدرت پردازش سیستم‌های کامپیوتری ۵۲ درصد در سال ارتقا پیدا کردند (۹) (شکل ۱-۲) و درنتیجه آنالیزهای بیوانفورماتیکی تسهیل و تسریع شدند.

در سال ۱۹۹۵ کرگ ونتر[8] و همیلتون اسمیت[9] توانستند برای اوّلین بار ژنوم کامل یک موجودِ زنده، یعنی باکتری *Haemophilus influenza*، را توالی‌یابی کنند (۱۰). یک سال بعد ژنوم یک موجود یوکاریوت تک‌سلولی، *Saccharomyces cerevisiae* (۱۱)، و سه سال بعد ژنوم یک موجود زندهٔ پرسلولی، نماتد *Caenorhabditis elegans*، به‌طور کامل توالی‌یابی شد (۱۲). اوّلین نسخه از ژنوم انسان که توالی‌یابی آن از سال ۱۹۹۰ شروع شده بود، در سال ۲۰۰۰ ارائه شد.

1. Wisconsin
2. Richard Stallman
3. Perl
4. Python
5. Computational Biology
6. Biostatistics
7. National Center for Biotechnology Information
8. Craig Venter
9. Hamilton Smith



شکل ۱–۲ روند افزایش قدرت پردازش کامپیوترها در ۴۰ سال اخیر. قدرت پردازشی کامپیوترهای شخصی از ۵ مگاهرتز در سال ۱۹۸۷ به ۴/۵ گیگاهرتز در سال ۲۰۱۷ رسیده است. نمودار از منبع (۹) اقتباس شده است.

نسل دوم روش‌های توالی‌یابی

از نیمهٔ دههٔ ۲۰۰۰ م. موج دیگری از تحولات در دنیای فنّاوری ایجاد شد و روش‌های توالی‌یابی نسل دوم ابداع شدند. در این روش‌ها تعداد بسیار زیادی خوانش از هر توالی تولید می‌شود. روش توالی‌یابی 454 که مبتنی‌بر توالی‌یابی براساس تشخیص آزاد شدن گروه پایروفسفات[1] است، در سال ۲۰۰۵ ارائه شد. فنّاوری Solexa/Illumina در سال ۲۰۰۶، SOLiD در ۲۰۰۷ و فنّاوری Ion Torrent در سال ۲۰۱۱ به بازار فنّاوری معرفی شدند (۱۳). این روش‌های توالی‌یابی به‌طورکلی به‌عنوان روش‌های توالی‌یابی نسل دوم شناخته می‌شوند. در این روش‌ها قطعهٔ DNA اوّلیه به روش PCR در امولسیون[2] و یا PCR روی یک صفحه[3] تکثیر می‌شود و سپس قطعات تولیدشده با روش توالی‌یابی براساس سنتز رشته (SBS)[4] یا توالی‌یابی براساس اتصال (SBL)[5] توالی‌یابی می‌شوند.

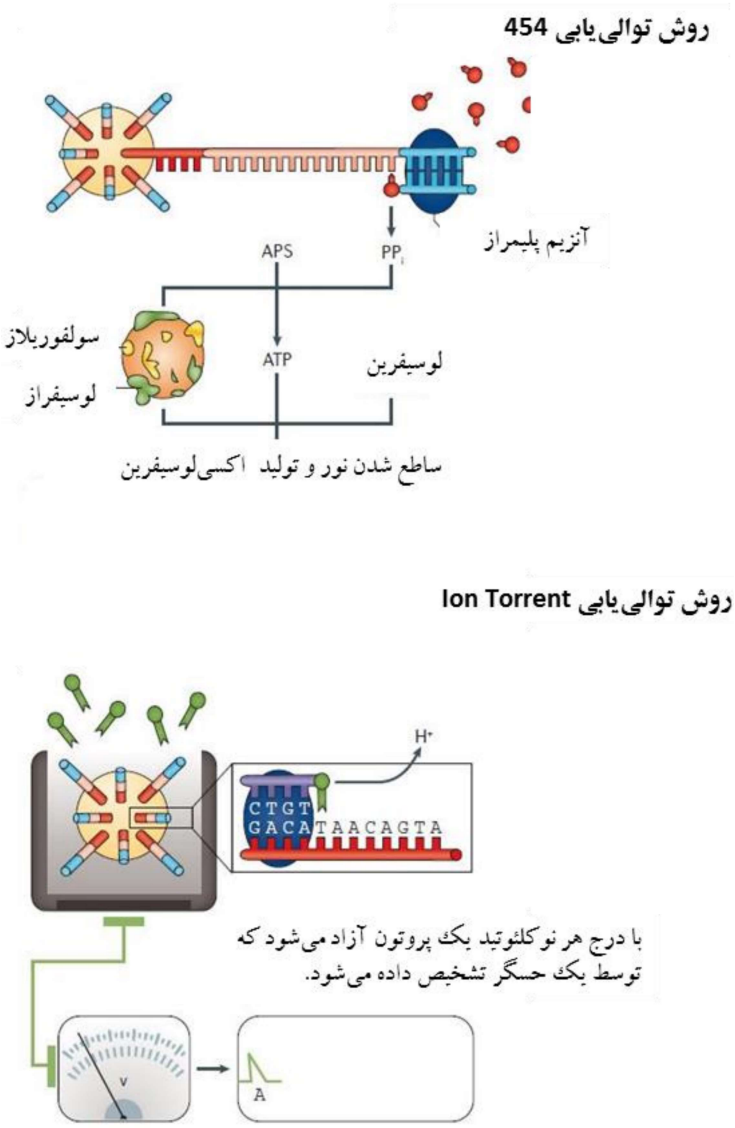
اصول بیوشیمیایی روش‌های توالی‌یابی نسل دوم با هم متفاوت است. برای مثال در روش 454 آزاد شدن پایروفسفات و در روش Ion Torrent آزاد شدن پروتون پایش می‌شود. یادآوری می‌شود که در هنگام همانندسازی DNA دو نوکلئوتید به یکدیگر متصل و پیوند فسفودی‌استر ایجاد می‌شود. در این

1. Pyrophosphate
2. Emulsion PCR
3. Bridge PCR
4. Sequencing by synthesis
5. Sequencing by ligation

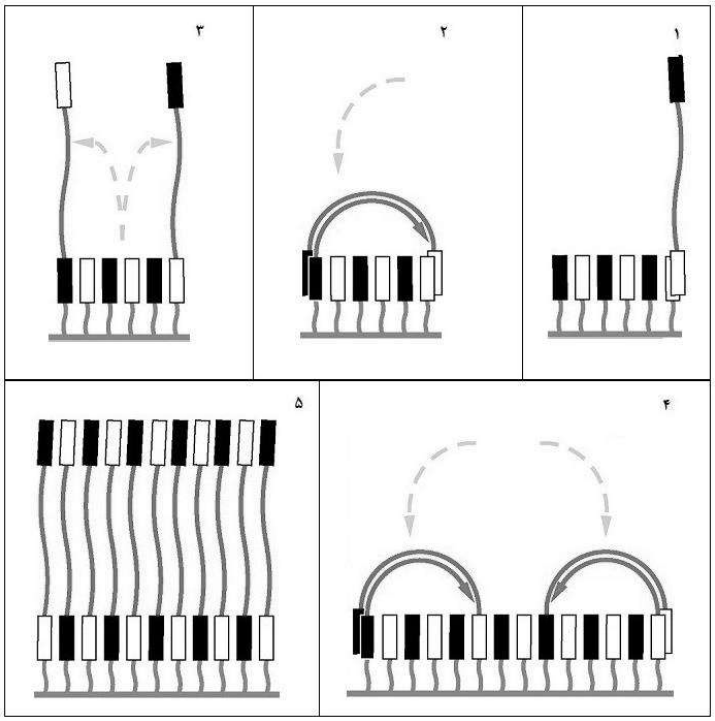
فرایند دو مولکول فسفر (پایروفسفات، PPi) و یک پروتون نیز آزاد می‌شود. روش توالی‌یابی 454 بر تشخیص آزاد شدن PPi استوار است. روش کار به‌این‌صورت است که مخلوطی از آنزیم‌های سولفوریلاز، لوسیفراز و مادّهٔ لوسیفرین در واکنش توالی‌یابی وارد می‌شود. مولکول PPi آزادشده توسط آنزیم سولفوریلاز به ATP تبدیل می‌شود. آنزیم لوسیفراز با استفاده از این مولکول ATP لوسیفرین را به اکسی‌لوسیفرین تبدیل می‌کند. اکسی‌لوسیفرین نوری ساطع می‌کند که می‌تواند توسط یک دوربین تشخیص داده شود (۱۴). روش Ion Torrent نیز مشابه روش 454 است با این تفاوت که در این روش پروتون آزادشده تشخیص داده می‌شود (شکل ۱-۳). درواقع در این روش از یک pH متر با حساسیت خیلی بالا استفاده می‌شود که می‌تواند تغییرات جزئی pH ناشی از آزاد شدن پروتون را آشکار کند.

در روش توالی‌یابی ایلومینا[1]، آداپتورهای ویژه‌ای به دو انتهای مولکول‌های DNA قطعه‌قطعه‌شده متصل می‌شوند. سپس این مولکول‌ها به صفحه‌های ویژه‌ای که الیگونوکلئوتیدهای مکمّل برای آداپتورها روی آن‌ها متصل شده است، اضافه می‌شوند. سپس یک مرحلهٔ PCR باعث ایجاد تعداد زیادی کپی از هر مولکول DNA متصل‌شده در هر نقطه از صفحهٔ ایلومینا می‌شود که به‌عنوان خوشهٔ DNA[2] شناخته می‌شود. این فرایند تکثیر روی صفحه یا تکثیر به‌واسطهٔ ایجاد پل[3] نامیده می‌شود. این نام‌گذاری به‌این‌دلیل است که در هر دور تکثیر، DNA به‌سمت الیگونوکلئوتید مکمّل آداپتور آزاد خود روی صفحه خم و به آن متصل می‌شود. درنتیجه دو الیگونوکلئوتید متصل به صفحه به‌عنوان پرایمر عمل می‌کنند و تکثیر قطعهٔ جدید، روی قطعه‌ای که از یک الیگونوکلئوتید به الیگونوکلئوتید دیگر صفحه پل زده است، انجام می‌شود. سپس توالی‌یابی به روش SBS با استفاده از نوکلئوتیدهای خاتمه‌دهندهٔ برگشت‌پذیرِ فلوئورسنت[4] انجام می‌شود. این نوکلئوتیدها به‌دلیل اینکه ترکیب شیمیایی فلوئورسنت آن‌ها موقعیت '۳ را اشغال می‌کند، وقتی در رشته قرار می‌گیرند از افزودن نوکلئوتیدهای دیگر جلوگیری می‌کنند، درنتیجه قبل از اینکه تکثیر ادامه یابد باید این مانع فلوئورسنت برداشته شود تا توالی‌یابی ادامه یابد. در هر سیکل نور فلوئورسنت ساطع‌شده از نوکلئوتید خاتمه‌دهنده، قبل از حذف این نوکلئوتید که برای ادامهٔ تکثیر در چرخهٔ تکثیر بعدی ضروری است، تشخیص داده می‌شود. به‌این ترتیب با افزودن هریک از چهار نوکلئوتید فلوئورسنت برگشت‌پذیر در هر سیکل، برخی از خوشه‌های DNA، فلوئورسنت را در انتهای خود نشان می‌دهند که نشان از حضور آن نوکلئوتید در توالی آن‌هاست (۵). روش SBS در شکل ۴-۱ و انیمیشن ۱-۲ نمایش داده شده است.

1. Illumina
2. DNA cluster
3. Bridge amplification
4. Fluorescent reversible-terminator dNTPs



شکل ۱-۳ مقایسهٔ اساس بیوشیمیایی دو روش توالی‌یابی نسل دوم. درنتیجه اضافه شدن یک نوکلئوتید به زنجیرهٔ درحال گسترش DNA، پروتون و گروه پایروفسفات آزاد می‌شود. در روش توالی‌یابی 454 گروه پایروفسفات و در Ion Torrent پروتون تشخیص داده می‌شود و براین‌اساس توالی‌یابی صورت می‌گیرد. پایروفسفات توسط آنزیم سولفوریلاز به ATP تبدیل می‌شود. مولکول ATP تولیدشده توسط آنزیم لوسیفراز استفاده می‌شود و لوسیفرین را به اکسی‌لوسیفرین که نور قابل‌آشکارسازی ساطع می‌کند، تبدیل می‌کند. در Ion Torrent یک pH متر با حساسیت بالا تغییرات غلظت پروتون را آشکار می‌کند و به‌این‌ترتیب نوکلئوتید درج‌شده در توالی DNA مشخص می‌شود (۱۵).



شکل ۱-۴ روش توالی‌یابی ایلومینا. در مرحلهٔ ۱،DNA به صفحه متصل می‌شود. در مرحلهٔ ۲ DNA به سمت الیگونوکلئوتید مکمّل خود خم می‌شود تا به آن متصل شود. سپس تکثیر انجام می‌شود. در محلهٔ ۳ دو رشته ساخته شده است. در مرحلهٔ ۴ از روی دو رشتهٔ موجود، دو رشتهٔ دیگر به روش تکثیر پل ساخته می‌شود. در مرحلهٔ ۵ خوشه‌های کلونی از قطعات DNA متصل به صفحه دیده می‌شود (۵).

نسل سوم روش‌های توالی‌یابی

در سال ۲۰۰۹ شرکت فنّاوری Helicos روش جدید توالی یابی معرفی کرد که وجه تمایز آن با روش‌های نسل دوم این بود که در آن نیاز به تکثیر DNA اوّلیه نبود و در تئوری امکان توالی‌یابی یک تک‌مولکول را فراهم می‌آورد. در سال ۲۰۱۱ شرکت Pacific Bioscience نیز فنّاوری مشابهی را روانهٔ بازار کرد. در این روش که به روش PacBio شناخته می‌شود، مولکول DNA پلیمراز در سطح یک روزنه مستقر می‌شود و با عبور DNA از آن روزنه پلیمراز شروع به سنتز رشتهٔ مکمّل آن می‌کند و به‌این ترتیب توالی مولکول DNA مشخص می‌شود. این روش توالی‌یابی در انیمیشن ۱-۳ نمایش داده شده است.

جدول ۱-۱ مقایسهٔ طول خوانش و تعداد خوانش در روش‌های مختلف توالی‌یابی (۱۶)

فنّاوری	طول خوانش (جفت باز)	بازده (تعداد خوانش در هر اجرا)
Roche 454	۷۰۰	۷۰۰ هزار
Illumina HiSeq	۳۰۰	۳۰۰ میلیارد
SOLiD	۱۰۰	۱۰۰ میلیارد
Ion Torrent	۲۰۰	۶۰ میلیارد
PacBio RS II	۱۴۰۰۰	۴۷ هزار

فنّاوری جدیدتر و متفاوت‌تر توسط شرکت‌های Oxford Nanopore و IBM Transistor ابداع شد. اساس کار در این روش این است که مولکول DNA از روزنهٔ بسیار باریکی که در آن شدت جریان الکتریکی مشخصی برقرار است، عبور داده می‌شود و تغییرات در شدت جریان که به نوع نوکلئوتید عبوری بستگی دارد، ثبت می‌شود. براساس تغییرات در شدت جریان، توالی DNA مشخص می‌شود. این روش در انیمیشن ۱-۴ نمایش داده شده است.

روش‌های توالی‌یابی نوین که سرعت و بازده بالایی دارند، انقلابی در علوم زیستی، به‌ویژه در پزشکی و کشاورزی به‌وجود آورده‌اند. این فنّاوری‌ها حجم بسیار زیادی دادهٔ ژنتیکی فراهم می‌آوردند که مطالعهٔ سیستم‌های زنده را در سطح تک‌سلول، بافت، موجود زنده و حتّی یک اکوسیستم فراهم می‌آورد. در پزشکی روش‌های تشخیصی مبتنی‌بر داده‌های NGS در دنیا رایج است و خوشبختانه در ایران نیز درحال متداول شدن است. در علوم کشاورزی با اتکا به این فنّاوری‌ها نقشه‌یابی و شناسایی ژن‌های با ارزش که کیفیت و کمّیت محصولات گیاهی و دامی را افزایش می‌دهند، تسهیل و تسریع شده است. ازطرف‌دیگر، این فنّاوری‌ها چالش‌های جدید و بزرگ‌تری را ازنظر آنالیز داده‌های حاصل از این فنّاوری‌ها به‌همراه آورده است. قابل‌توجه اینکه ابزارهای بیوانفورماتیکی که در سال ۲۰۱۷ منتشر شده‌اند پنج برابر بیشتر از ابزارهای منتشرشده بین سال‌های ۱۹۹۰ تا ۲۰۰۰ بوده است (۱۷) که نشان‌دهندهٔ لزوم یافتن روش‌ها و ابزارهای کارآمدتر برای آنالیز داده‌های NGS است.

توالی‌یابی نوین برای اهداف متنوعی ازجمله توالی‌یابی کل ژنوم[1]، توالی‌یابی مجدّد ژنوم[2]، توالی‌یابی ترنسکریپتوم[3]، توالی‌یابی متاژنوم[4] و متاترنسکریپتوم[5] استفاده می‌شوند. در همهٔ این کاربردها مسیر کلی انجام کار تقریباً مشترک است (شکل ۱-۵).

ابزارهای آنالیز داده‌های NGS

متناسب با سرعت پیشرفت‌ها در توسعهٔ روش‌های نوین توالی‌یابی، محقّقان علوم کامپیوتر، ریاضی و

1. Whole Genome Sequencing
2. Genome Re-sequencing
3. Transcriptome
4. Metagenome
5. Metatranscriptome