کارآیی نامعقول ریاضیات در زیست‌شناسی مولکولی•

آرتر لسک*
ترجمهٔ افرا علیشاهی

عنوان مقالهٔ من اقتباسی است از عنوان مقالهٔ مشهور ویگنر، «کارآیی نامعقول ریاضیات در علوم طبیعی [۱]». البته این عنوان در فیزیک و در زیست‌شناسی مولکولی، از دو بابت مخالفْ توجه‌برانگیز است. در فیزیک، بدیهی است که ریاضیات کارآمد است ــ بسیاری از غولهایی که فیزیکدانها بر شانه‌های آنها ایستاده‌اند ریاضیدان‌اند ــ و مایهٔ شگفتی است که ویگنر این را نامعقول می‌داند. در زیست‌شناسی مولکولی، نقش واقعی ریاضیات بدیهی نیست، و بیم آن می‌رود که انتظار کارآیی از ریاضیات نامعقول باشد، بیمی که در این مورد از فیزیک بسیار موجه‌تر است. البته، بسیاری از ابزارهای متداول در زیست‌شناسی مولکولی محاسباتی ــ مثلاً جستجو در پایگاه داده‌ها برای یافتن دنباله‌های شبیه به یک دنبالهٔ یافت‌شده ــ مسلماً مبتنی بر ریاضیات و علوم کامپیوتر هستند. اما اینکه آیا درک غایی ما از فرآیندهای حیاتی به زبان ریاضیات بیان شود ــ آن‌چنان‌که، مثلاً، مفاهیم تقارن زمینه‌ساز بیان قوانین فیزیک هستند ــ یا به زبان سنتیِ توصیفیِ «روایی» زیست‌شناسی، هنوز محل بحث است.

چرا تردید در کارآیی ریاضیات در زیست‌شناسی می‌تواند معقول باشد؟ خصوصیات مشاهده‌شدهٔ سیستمهای زنده به‌وسیلهٔ ترکیبی از اینها تعیین می‌شوند:

- قوانین فیزیک و شیمی
- سازوکار تکامل
- تصادف تاریخی

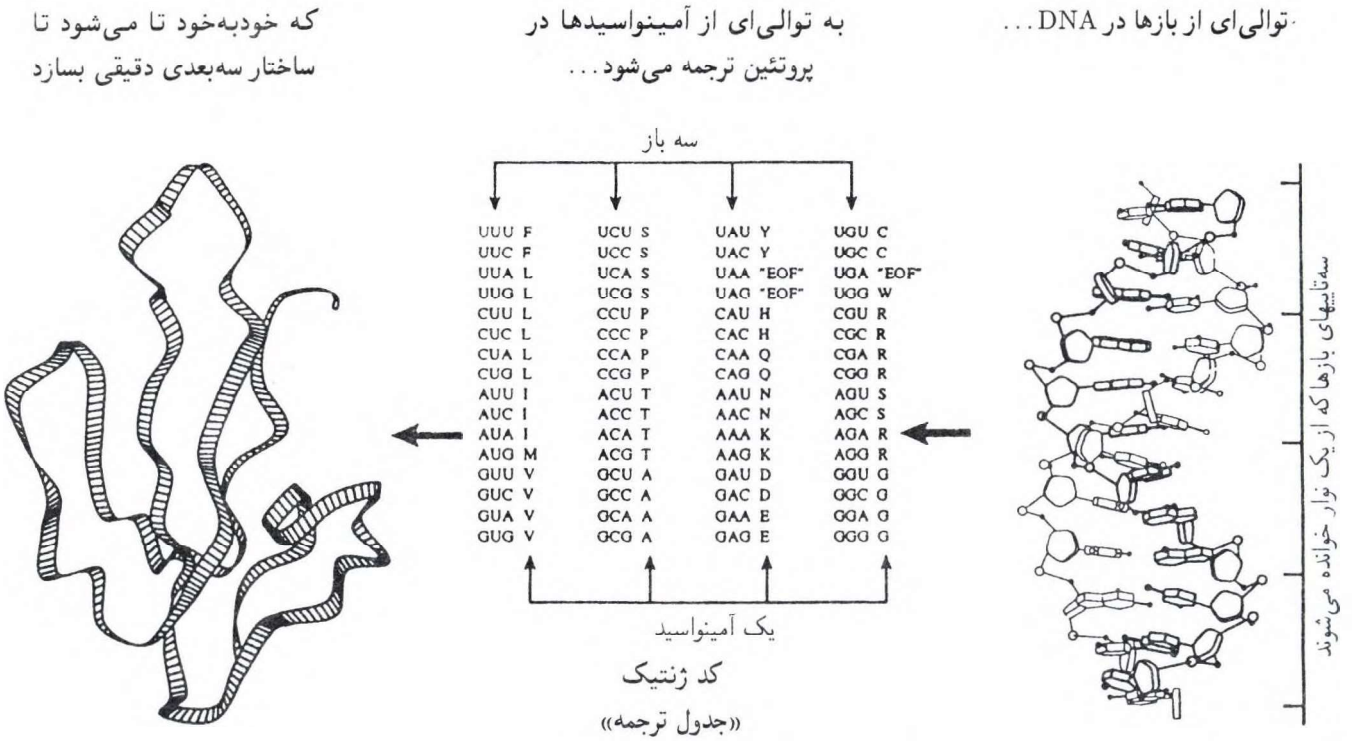
تفکیک تأثیر اینها از هم دشوار است، و کشاکش سازنده‌ای که بین آنهاست، تحقیقات ما را تحت تأثیر قرار می‌دهد. بسیاری از قوانین فیزیک، جهان طبیعی را ــ که شامل سیستمهای زنده هم می‌شود ــ با تعیین روابط بین شرایط اولیه و نهایی توصیف می‌کنند. در زیست‌شناسیْ پیچیدگی مجموعهٔ شرایط اولیهٔ ممکن، دشواریهایی پدید می‌آورد. نقش عظیم تصادف تاریخی ما را به تردید می‌اندازد و به زانو درمی‌آورد: حتی اگر قوانین بنیادی فیزیک و شیمی نتایج ساده‌ای به دنبال داشته باشد که توصیفات مبسوطی از فرآیندهای حیات در اختیار ما قرار دهد، ما ممکن است قادر به کشف آنها نباشیم، زیرا پدیده‌های مورد مطالعهٔ ما پیچیده‌ترند، در برابر ایده‌آل‌سازیهای ساده‌کننده مقاومت می‌کنند، و خصوصیاتی از خود نشان می‌دهند که تحت تأثیر شدید انتخاب شرایط اولیه از بین مجموعهٔ بسیار عظیم و پراکنده‌ای از شرایط ممکن است. سیبها در زیست‌شناسی نقشی بسیار مهمتر از افتادن بر سر مردم ایفا می‌کنند.

موضوع بحث در زیست‌شناسی مولکولی

اشیاء مورد مطالعهٔ ما حداقل شکلی دارند که می‌توانیم سعی کنیم ریاضیات را در مورد آنها به‌کار گیریم. این اشیاء عبارت‌اند از:

- توالیهای ژنها در DNA
- توالیهای آمینواسیدها در پروتئینها
- ساختار پروتئینها
- کارکرد پروتئینها

احتمالاً خوانندگان در بارهٔ پروژهٔ ژنوم شنیده‌اند؛ هدف این پروژه تعیین توالی کامل DNA در موجودات زنده است: یعنی مجموعهٔ نقشه‌ها. توالیهای DNA در ژنومها حاوی همهٔ اطلاعاتی هستند که یک موجود زنده برای تولد، بزرگ شدن و رشد، و مرگ به آنها نیاز دارد. با تکمیل تعیین توالی ژنوم مخمر در سال ۱۹۹۶، ما همانقدر در بارهٔ یک سلول مخمر می‌دانیم که خود سلول مخمر می‌داند. این عبارت آن‌قدر که به‌نظر می‌رسد متکبرانه نیست. ما واقعاً همهٔ اطلاعات را داریم. باید پذیرفت که ما نمی‌توانیم این اطلاعات را به همان کارآمدی تفسیر کنیم که یک سلول مخمر می‌تواند، اما ما مجموعهٔ کامل نقشه‌ها را در اختیار داریم. ولی این نقشه‌ها تنها یک توصیف ایستا از ساختار و عملکرد بالقوه به‌دست می‌دهند؛ می‌ماند اینکه مشاهدات خود را تا مرحلهٔ یکپارچه‌سازی توصیف و کارکرد پروتئین در زمان و فضا درون یک موجود



شکل ۱ جریان اطلاعات در خلال «خوانش» از روی یک ژن. ژنها، یا نقشه‌های بنیادی موجودات زنده، در ساختار DNA گنجانده شده‌اند (راست). در سمت راست شکل، مارپیچ مضاعف DNA را می‌بینیم، که حاوی دو نوار به‌هم پیچیده است که یکی با خطوط باریک و دیگری با خطوط پهن رسم شده است. «اثر پلکانی» توسط مجموعه‌ای از زیرواحدهای شیمیایی به‌نام «باز» پدید می‌آید. در هر نقطه روی هر نوار یکی از چهار باز A، T، G یا C ممکن است قرارگیرند. خوانندهٔ تیزبین می‌تواند ببیند که این بارها ــ «لبه‌های قائم» پله‌های پلکان مارپیچی ــ به شکلهای مختلف ظاهر می‌شوند. بازهای هم‌سطح هر دو نوار با هم ارتباط برقرار می‌کنند، و این ارتباط مستلزم مکمل بودن مطلق است: حضور یک A روی یک نوار مستلزم حضور یک T روی نوار دیگر است، و یک G روی یک نوار به یک C روی نوار دیگر نیاز دارد، و برعکس، یک T روی یک نوار مکمل یک A و یک C مکمل یک G روی نوار دیگر است. به این ترتیب، هر نوار حاوی اطلاعات کافی برای ساختن نوار همراه خود است. به لحاظ منطقی، راه تکثیر DNA آن است که نوارها را از هم جدا کنیم و مکمل هر کدام از نوارهای مجزا را دوباره بسازیم. توالیهای بازها در ژنها، به‌کمک یک جدول ترجمهٔ مستقیم به‌نام «کد ژنتیک» توالیهای آمینواسید در پروتئینها را رمزگشایی می‌کنند (وسط). پیام ژنتیکی با الفبای چهار حرفی A، T، G و C نوشته می‌شود. پروتئینها هم پلیمرهایی هستند که توالی‌ای از سازه‌های شیمیایی را در خود دارند، چنانکه در هر نقطه یکی از بیست آمینواسید ممکن قرار می‌گیرد، که با نمادهای A، C، D، E، F، G، H، J، K، L، M، N، P، Q، R، S، T، V، W، Y مشخص می‌شوند. برای مشخص کردن مجموعهٔ ۲۰ آمینواسید، هر کدام به بیش از دو باز نیاز دارد؛ درواقع از روی توالی DNA در یک زمان سه باز خوانده می‌شود، با افزونگی‌ای که برای تکامل مطلقاً ضروری است (سه سه‌تایی از بازها برای علامتهای انتهایی «پایان حیات» نگهداری شده‌اند). پروتئینها خود به خود تا می‌شوند تا ساختارهای طبیعی فعال سه‌بعدی بسازند (چپ). این نقطه‌ای است که در آن طبیعت از توالیهای یک‌بعدی ژنها به جهان سه‌بعدی‌ای که ما در آن زندگی می‌کنیم جهش می‌کند. مثال فوق یک توکسین از یک مار آبی است، یکی از ساختارهای پروتئین متعددی که به‌وسیلهٔ بلورنگاری با اشعهٔ X تعیین شده است. هر ژن توالی‌ای از بازها دارد، که ابتدا توالی آمینواسیدهای یک پروتئین، و سپس ساختار سه‌بعدی آن، و سپس کارکرد آن را تعیین می‌کند.

زنده بسط دهیم. جمع‌آوری این اطلاعات به «پروژهٔ پروتئوم» معروف است. پروژه‌ای که در حال جمع‌آوری قواست تا در عصر پس از ژنوم ایفای نقش کند.

میزان اندازه‌گیری توالیهای ژنها بسیار زیاد و فزاینده است. در سال ۱۹۹۸ توالی کامل DNAی یک نوع کرم موسوم به Caenorhabditis elegans تکمیل شد ($۹٫۷ \times ۱۰^۷$ باز). محتمل است که در سالهای ۱۹۹۹ و ۲۰۰۰ به‌ترتیب شاهد تکمیل توالی DNAی مگس سرکه ($۱٫۸ \times ۱۰^۸$ باز) و ژنوم انسان ($۳٫۴ \times ۱۰^۹$ باز) و بسیاری موجودات زندهٔ کوچک و بزرگ دیگر باشیم. لویی پانزدهم می‌توانست بگوید: «بعد از من، طوفان»، اما نوح نمی‌توانست؛ ما هم نمی‌توانیم.

توالیها و ساختارهایی که ما مطالعه می‌کنیم ارتباطات مهمی با هم دارند. در سطح مولکول، توالیهای ژنهای DNA، توالیهای آمینواسیدهای پروتئینها را رمزگشایی می‌کنند. سپس توالیهای آمینواسید پروتئینها ساختار سه‌بعدی پروتئینها را تعیین می‌کنند. آنگاه ساختار پروتئینها، کارکرد آنها را تعیین می‌کنند (شکل ۱). برای تعیین کارکرد پروتئین، یک ساختار سه‌بعدی دقیق ضروری است، زیرا تعاملهای مورد نیاز به‌کنار هم قرارگرفتن بخشهای مختلف مولکولها در روابط فضایی دقیق بستگی دارند. نهایتاً پسخورد کارکرد پروتئین به دنبالهٔ ژنها ــ از طریق تکامل به روش انتخاب طبیعی ــ این حلقه را کامل می‌کند.

توالیهای DNA درکامپیوترهای ما، رشته‌های حرفی هستند، یعنی اشیاء یک‌بعدی. ژنها یا زیررشته‌هایی از توالیهای ژنوم، براساس یک کدنویسی تقریباً جهان‌شمول به توالیهای آمینواسید پروتئینها ترجمه می‌شوند. این توالیها هم به‌وسیلهٔ رشته‌های حرفی یک‌بعدی بازنمایی می‌شوند. سپس پروتئینها خود به خود تا می‌شوند تا ساختارهای سه‌بعدی «طبیعی» یکتا بسازند. (شاهد این مدعا آن است که می‌توان آنها را به‌وسیلهٔ حرارت تخریب کرد ــ ساختار سه‌بعدی را ویران کرد ــ و وقتی سرد شوند مجدداً شکل اولیهٔ خود را پیدا می‌کنند، مانند آلیاژهای شکل‌ـحافظه[1]). تاخوردگی خود به‌خودِ پروتئینها

1. shape-memory

همان نقطه‌ای است که طبیعت موجب می‌شود توالیهای یک‌بعدی ژنها به جهان سه‌بعدی‌ای که ما در آن زندگی می‌کنیم جهش کنند.

اهداف زیست‌شناسی مولکولی محاسباتی

اهداف ما از پرداختن به این مبحث چیست؟ نخست آنکه بتوانیم مشابهتها و تفاوتهای موجود بین توالیها و بین ساختارها را به‌سادگی توصیف و طبقه‌بندی کنیم. توپولوژی فضای توالی، فضای ساختار، و فضای کارکرد پروتئین چیست؟ نگاشتهای بین این فضاها چگونه‌اند؟ ما می‌خواهیم که با استفاده از تکامل به‌عنوان اصل سازماندهی، بتوانیم روابط بین توالی، ساختار و کارکرد پروتئین را توصیف و پیشگویی کنیم.

از کجا باید کار را شروع کرد؟ سیدنی برنر یک بار با افسوس گفت: «مسأله در زیست‌شناسی این است که هیچ نوسانگر همسازی وجود ندارد.» منظور او از این گفته این بود که در زیست‌شناسی، برخلاف فیزیک، راه گریزی از پیچیدگی نیست، حتی از طریق ایده‌آل‌سازی. در فیزیک، نوسانگر همساز یک مسألهٔ ساده است که می‌توان به روشهای متعددی آن‌را به دقت حل کرد؛ این مسأله در مورد برخی از پدیده‌ها دقیقاً قابل اعمال است، و برای سایر پدیده‌ها تخمین مؤثری است. نوسانگر همساز در فیزیک یک بستر آزمون سنتی برای روشهای جدید است. در حقیقت، زیست‌شناسی مولکولی محاسباتی دو «نوسانگر همساز» به تعبیر برنر دارد: هم‌ردیف‌سازی توالیها، و برهم‌نهی ساختارها. این اعمال، که می‌توان آنها را با دقت و کارآیی انجام داد، پایهٔ بسیاری از تحلیلهای روابط توالی-ساختار در زیست‌شناسی مولکولی هستند. اکنون دانستن این نکته که جهان واقعی اغلب ناهمساز است شگفت‌انگیز نخواهد بود. با این‌همه، بسیاری از این ابزارهای ارزشمند به‌کمک این موارد ساده ساخته شده‌اند.

اما، ابزارها پاسخ می‌سازند نه سؤال. پژوهش در این حوزه همچنان بر تعامل بین دانشمندان و داده‌ها، با کمک روشهای ریاضیاتی و محاسباتی، متکی است.

توالیها و هم‌ردیف‌سازی آنها

توالیهای ژن و پروتئین شکل رشته‌های حرفی به‌خود می‌گیرند. برای توالیهای ژن، حروف از مجموعهٔ چهار عضوی $\{A, T, G, C\}$ انتخاب می‌شوند که نشان‌دهندهٔ نوکلئوتیدهای آدنین، تایمین، گوانین و کایتوزین هستند. برای توالیهای پروتئین، حروف از یک مجموعهٔ بیست عضوی انتخاب می‌شوند که نشان‌دهندهٔ بیست آمینواسید معمولی هستند.

هم‌ردیف‌سازی

هم‌ردیف‌سازی دو رشتهٔ حرفی به معنی تعیین یک تناظر معنی‌دار بین مؤلفه‌های آنهاست. برای دو رشتهٔ حرفی زیر

c t a t a a t c و g c t g a a c

دو هم‌ردیف‌سازی ممکن عبارت‌اند از

g c t g – a a – c g c t g a – a – – c

و

– c t a t a a t c – c t – a t a a t c

چگونه می‌توان تعیین کرد که کدام‌یک از این دو، یا کدام‌یک از بسیاری از هم‌ردیف‌سازیهای ممکن دیگر، بهترین هم‌ردیف‌سازی است؟ آیا ما می‌توانیم متریکی برای رشته‌های حرفی طرح کنیم و فاصلهٔ بین آنها را تعریف کنیم؟ شاخصهای عدم شباهت بین رشته‌های حرفی عبارت‌اند از:

(۱) فاصلهٔ همینگ، که بین دو رشتهٔ هم‌طول تعریف می‌شود، یعنی تعداد مکانهایی که حاوی حروف نامشابه هستند.

(۲) فاصلهٔ لونشتاین[1] بین دو رشته که الزاماً هم‌طول نیستند، یعنی کمترین تعداد «عملیات ویرایش» مورد نیاز برای تبدیل یک رشته به رشتهٔ دیگر، که در اینجا عمل ویرایش عبارت است از حذف، درج، یا تغییر یک حرف در هر توالی. هر توالی داده شده از عملیات ویرایش، یک هم‌ردیف‌سازی یکتا را القا می‌کند. اما عکس این مطلب صحیح نیست.

در زیست‌شناسی مولکولی، می‌دانیم که درج و حذف در توالیهای ژن و پروتئین رخ داده‌اند. بنابراین فاصلهٔ همینگ به‌اندازهٔ کافی عام نیست. به‌علاوه، شواهدی وجود دارد که نشان می‌دهد وقوع برخی تغییرات محتملتر از تغییرات دیگر است. بنابراین حتی فاصلهٔ لونشتاین نیز باید تعمیم داده شود تا براساس مدل تکاملی زیربنایی ما، وزنهای متفاوتی برای عملیات مختلف ویرایش در آن منظور گردد. برای مثال، به‌نظر می‌آید که جهشها محافظه‌کارانه رخ می‌دهند؛ جایگزینی یک آمینواسید در یک پروتئین با آمینواسید دیگری که اندازه یا خواص فیزیکی-شیمیایی یکسان دارد محتملتر است تا جایگزینی آن با آمینواسید دیگری که خصوصیات آن بیشتر متفاوت است. برای انعکاس این مطلب، به‌جای شمارش گسستهٔ عملیات ویرایش، به هر تغییر در توالی یک «هزینه» (متعلق به $\mathbb{R}$) نسبت می‌دهیم. همچنین شواهدی وجود دارد که نشان می‌دهد هزینهٔ یک فاصلهٔ خالی، مانند مدل لونشتاین، با طول آن متناسب نیست؛ اگرچه انتخاب مناسب وزنهای فواصل خالی به شکل تابعی از طول دقت زیادی می‌خواهد، در بسیاری از طرحها از یک تابع خطی با یک پارامتر α برای مقدار اولیه دادن به فاصلهٔ خالی و یک پارامتر کوچکتر β برای بسط فاصله، به‌منظور محاسبهٔ هزینهٔ فاصلهٔ خالی به شکل

$$\alpha + \beta \times (\text{طول فاصلهٔ خالی} - ۱)$$

استفاده می‌شود. الگوریتمهایی وجود دارند که با کمینه کردن مجموع هزینه‌های عملیات ویرایشی که یک رشته را به رشتهٔ دیگر تبدیل می‌کند، بهترین هم ردیف‌سازی را تعیین می‌کنند.

مسألهٔ هم‌ردیف‌سازی بهینهٔ توالی را به شکل صوری می‌توان چنین بیان کرد. دو رشتهٔ حرفی $A = a_۱a_۲\dots a_n$ و $B = b_۱b_۲\dots b_m$ داده شده‌اند، که هر a_i و b_j عضوی از یک مجموعهٔ الفبای $\mathcal{A}$ است. فرض کنید $\mathcal{A}^+ = \mathcal{A} \cup \{\emptyset\}$. یک توالی از عملیات ویرایش، مجموعه‌ای از زوجهای مرتب (x, y) است، که $x, y \in \mathcal{A}^+$. عملیات ویرایش منفرد عبارت‌اند از:

جایگزینی b_j با a_i، که با (a_i, b_j) نشان داده می‌شود،

حذف a_i از رشتهٔ A، که با (a_i, ϕ) نشان داده می‌شود،

حذف b_j از رشتهٔ B، که با (ϕ, b_j) نشان داده می‌شود.

تابع هزینهٔ d بر عملیات ویرایش به شکل زیر تعریف می‌شود:

$$\text{هزینهٔ یک جهش} = d(a_i, b_j)$$

1. Levenshtein

$d(a_i, \emptyset)$ یا $d(\emptyset, b_j)$ = هزینهٔ یک حذف یا درج

و کمترین فاصلهٔ وزن‌دار بین رشته‌های A و B عبارت است از

$$D(A,B) = \min_{A \to B} \sum d(x,y)$$

که در آن $x, y \in \mathcal{A}^+$ و مقدار مینیمم به‌ازای همهٔ توالیهای عملیات ویرایش که A را به B تبدیل می‌کنند محاسبه می‌شود. اگر $d(x,y)$ یک متریک بر $\mathcal{A}^+$ باشد، $D(A,B)$ متریکی بر رشته‌های حروف از $\mathcal{A}^+$ خواهد بود. (در این بیان از مسأله فرض می‌شود که هزینهٔ فاصله‌های خالی مستقل از طول آنهاست؛ طرحهای واقع‌بینانه‌تر که به فاصله‌های خالی وزن می‌دهند، تعمیمی از این طرح هستند.)

مسألهٔ یافتن $D(A,B)$ است و یک یا چند هم‌ردیف‌سازی که با آن متناظرند. الگوریتمی که این مسأله را در زمان $\mathcal{O}(mn)$ حل می‌کند زمان درازی است که شناخته شده است، و در بسیاری از مسائل از قبیل ویرایش متن، تشخیص گفتار، و تحلیل آواز پرندگان به‌کار گرفته شده است [۲]. این الگوریتم توسط مقاله تأثیرگذار نیدلمن و وونش [۳]، به زیست‌شناسی وارد شد.

چند خصوصیت این الگوریتم قابل توجه‌اند.

• این الگوریتم یک بهینهٔ مطلق به‌دست می‌دهد: توجه داشته باشید که این یکی از دو «نوسانگر همساز» زیست‌شناسی مولکولی محاسباتی است. ما روشی در اختیار داریم که مطمئنیم که در مینیممهای موضعی به دام نخواهد افتاد.

• این خبرِ خوب بود. خبر بد اینکه تعبیر نتایج چندان سرراست نیست. اگرچه توالی‌ای از عملیات ویرایش که از یک هم‌ردیف‌سازی بهینه ناشی شده است ممکن است با یک مسیر تکاملی واقعی متناظر باشد، اما اثبات اینکه چنین است ممکن نیست. هر چه فاصلهٔ ویرایش بیشتر باشد، تعداد مسیرهای تکاملی معقول بیشتر است. نه تنها هم‌ردیف‌سازی‌های بهینه ممکن است یکتا نباشند، بلکه ممکن است هم‌ردیف‌سازیهای زیربهینهٔ بسیاری وجود داشته باشند که ارزش آنها کاملاً به مقدار بهینه نزدیک باشد. مثلاً، فیچ و اسمیت ژنهای جوجه را برای هموگلوبینهای α و β آزمایش کردند [۴]. آنها ۱۷ هم‌ردیف‌سازی بهینه یافتند، که یکی از آنها با هم‌ردیف‌سازی مبتنی بر ساختارهای هموگلوبین شناخته شده مطابقت داشت، و بیش از هزار هم‌ردیف‌سازی یافت شد که با مقدار بهینه کمتر از ۵٪ اختلاف داشت.

مشکلات ناشی از هم‌ردیف‌سازیِ جفت‌جفتِ توالیها

مشاهده شده است که با تکامل پروتئینها، توالیهای آمینواسیدها بسیار سریعتر از ساختار واگرا می‌شوند. در بسیاری از موارد می‌توان یک رابطهٔ تکاملی بین دو ساختار پروتئین یافت، حتی اگر هیچ شباهت قابل درکی بین توالیهای ژنها یا توالیهای آمینواسیدها وجود نداشته باشد. آنچه اتفاق می‌افتد این است: ژنها فضای توالیهای DNA را می‌کاوند، اما انتخاب طبیعی به مثابهٔ ترمزی در برابر تغییر ساختار عمل می‌کند تا کارکرد را حفظ کند. افزونگی[1] در کد ژنتیک ــ این واقعیت که تعداد زیادی از بازهای سه‌گانه یک آمینواسید یکسان را کدنویسی می‌کنند، و بسیاری از تغییرات تک‌بازی منجر به حصول آمینواسیدهایی با

1. redundancy

```
TYLWEFLLKLLQDR.EYCPRFIKWTNREKGVFKLV..DSKAVSRLWGMHKN.KPD
VQLWQFLLEILTD..CEHTDVIEWVG.TEGEFKLT..DPDRVARLWGEKKN.KPA
IQLWQFLLELLTD..KDARDCISWVG.DEGEFKLN..QPELVAQKWGQRKN.KPT
IQLWQFLLELLSD..SSNSSCITWEG.TNGEFKMT..DPDEVARRWGERKS.KPN
IQLWQFLLELLTD..KSCQSFISWTG.DGWEFKLS..DPDEVARRWGKRKN.KPK
IQLWQFLLELLQD..GARSSCIRWTG.NSREFQLC..DPKEVARLWGERKR.KPG
IQLWHFILELLQK..EEFRHVIAWQQGEYGEFVIK..DPDEVARLWGRRKC.KPQ
VTLWQFLLQLLRE..QGNGHIISWTSRDGGEFKLV..DAEEVARLWGLRKN.KTN
ITLWQFLLHLLLD..QKHEHLICWTS.NDGEFKLL..KAEEVAKLWGLRKN.KTN
LQLWQFLVALLDD..PTNAHFIAWTG.RGMEFKLI..EPEEVARLWGIQKN.RPA
IHLWQFLKELLASP.QVNGTAIRWIDRSKGIFKIE..DSVRVAKLWGRRKN.RPA
RLLWDFLQQLLNDRNQKYSDLIAWKCRDTGVFKIV..DPAGLAKLWGIQKN.HLS
RLLWDYVYQLLSD..SRYENFIRWEDKESKIFRIV..DPNGLARLWGNHKN.RTN
IRLYQFLLDLLRS..GDMKDSIWWVDKDKGTFQFSSKHKEALAHRWGIQKGNRKK
LRLYQFLLGLLTR..GDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKR

  L  fl  lL          i W       F          a  WG  K
```

شکل۲ هم‌ردیف‌سازی چندگانهٔ توالیهای ناتمام یک خانواده از پروتئینها که دامنه‌های ETS نامیده می‌شوند. هر خط با توالی آمینواسیدهای یک پروتئین متناظر است، که با توالی‌ای از حروفی که هرکدام نمایندهٔ یک آمینواسید است مشخص می‌شود. با نگاه به هر ستون می‌توان آمینواسیدهایی را که در آن مکان در هر یک از پروتئینهای خانواده پدیدار می‌شوند مشاهده کرد. به این طریق الگوهای اولویت‌دار قابل رؤیت می‌شوند. مثلاً مکان سوم حاوی یک لوسین یا L در هر توالی است ــ این امر دلالت بر این دارد که برخی محدودیتهای ساختاری یا کارکردی، تکامل را از تغییر وضعیت این مکان بازداشته‌اند. حروف زیر جدول مکان سازه‌های نامتغیر (حروف بزرگ) یا نامتغیر با یک استثنا (حروف کوچک) را مشخص می‌کنند. به توزیع نابرابر تغییر در ستونهای مختلف توجه کنید. تناوب سازه‌های ماندگار (۳، ۴ یا ۸) از وجود مارپیچها در پروتئین خبر می‌دهد، که خبر صحیحی است. الگوهای دیگر عمیقتر پنهان شده‌اند، و تعیین آنها نیاز به تحلیل محاسباتی دارد. چنین الگوهایی ممکن است حاوی همبستگیهایی میان توزیع آمینواسیدها در مکانهای مختلف باشند. مثلاً، در ستون چهارم از سمت چپ آمینواسید تیروزین یا Y، تنها در دو توالی آخر پدیدار می‌شود، سایر توالیها دارای تریپتوفان یا W هستند. یک همبستگی تقریبی در الگوی تغییر بین این ستون و ستونهای چهارم و پنجم از راست وجود دارد. بسیاری عقیده دارند (یا حداقل امیدوارند) که همبستگیهای الگوهای تغییر در مکانهای مختلف چنین جدولی از توالیها، سرنخهایی در بارهٔ محلهایی که در فضای سه‌بعدی بر هم اثر می‌گذارند به ما بدهد. متأسفانه قرائن بسیار ضعیف است.

خواص فیزیکی-شیمیایی یکسان می‌شود ــ نتایج ساختاری تغییرات توالی را تعدیل می‌کند.

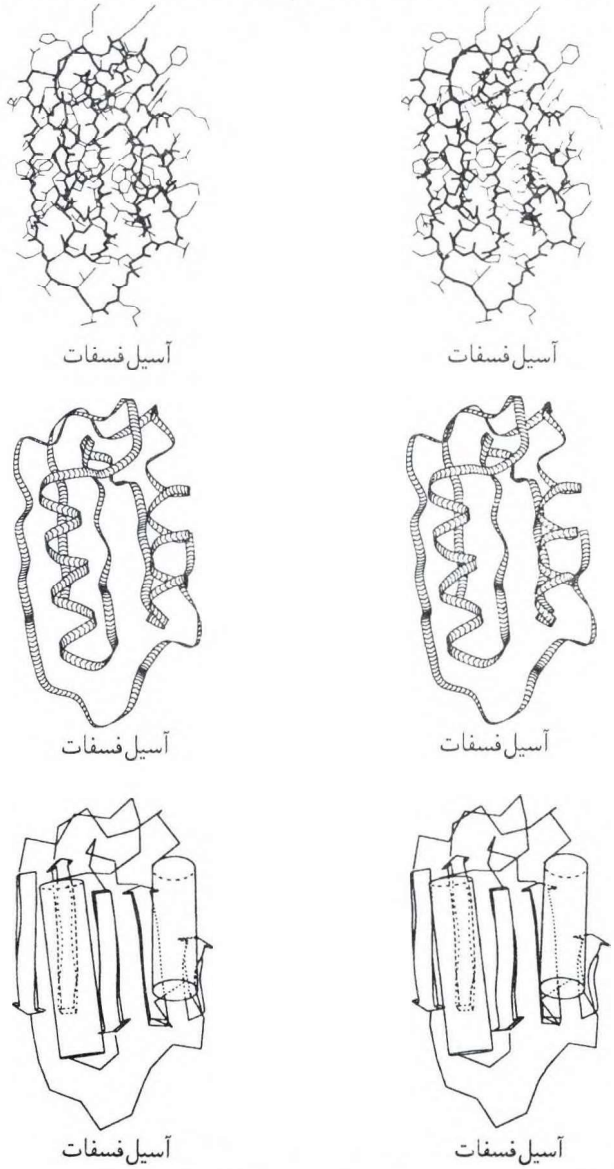
حتی اگر مشابهت در سطح توالی قابل تشخیص باشد، برای پروتئینهایی که ارتباط دوری با هم دارند، به کمک مقایسهٔ ساختارها که آخرین راه چاره است، می‌توان فهمید که هم‌ردیف‌سازی جفت‌جفتِ بهینهٔ توالیها اغلب منجر به جواب غلط می‌شود.

اما اگر تعداد زیادی توالی مرتبط در دسترس باشند، هم‌ردیف‌سازی چندگانهٔ توالیها نتایج با ارزشتر و دقیقتری نسبت به هم‌ردیف‌سازی جفت‌جفتِ توالیها به‌دست می‌دهد. چرا هم‌ردیف‌سازیهای چندگانه اطلاعات توالی را گسترش می‌بخشند؟ از اینجا الگوهای ماندگاری ظاهر می‌شوند. گستره و طبیعت تغییر در هر یک از مکانها راهنمای مهمی برای تعیین نقش ساختاری یا کارکرد ناحیه‌های مختلف توالی است (شکل ۲). برای مثال، سازه[1]هایی که در یک خانوادهٔ کامل از پروتئینها ماندگار [=بی‌تغییر] بوده‌اند معمولاً در کارکرد دخیل‌اند، یا دست‌کم اغلب نقشی اساسی در ساختار ایفا می‌کنند. برعکس، مناطقی که عملیات درج و حذف در آنها زیاد صورت می‌گیرد معمولاً با

1. residue

مناطق حاشیه‌ای متناظرند. (یک توالی در قبال ساختار خود منفعل است، یک جفت توالی هم‌ردیف‌شده، ساختار خود را نجوا می‌کنند، سه یا چند توالی ساختار خود را به صدای بلند فریاد می‌کنند).

اگر توالیها تنها به‌صورت غیرمستقیم به ساختار اشاره می‌کنند، چرا مستقیماً به سراغ ساختار نرویم؟ علت این است که مقدار داده‌ها در مورد توالیهای شناخته‌شده بسیار بیشتر از مقدار داده‌های ساختاری است. برای حدود ۲۰ موجود زنده، توالی همهٔ ژنوم تعیین شده است، که توالیهای کامل



شکل۳ پروتئینها چنان ساختارهای پیچیده‌ای هستند که لازم آمده است تا ابزارهای ویژه‌ای برای نمایش آنها ساخته شود. این شکل یک پروتئین نسبتاً کوچک به نام آسیل فسفات را با سه درجهٔ مختلف ساده‌سازی نشان می‌دهد. بالا: مدل کلی کامل ؛ زنجیرهٔ اصلی پررنگتر از شاخه‌های جنبی است. وسط: مسیر زنجیر، توسط یک خم درون‌یابی‌شدهٔ هموار بازنمایی شده است، فلشها جهت زنجیره را تعیین می‌کنند. پایین: نمودار اجمالی، که در آن استوانه‌ها نمایانگر مارپیچها و فلشها نمایانگر نوارها هستند. با تغییر خطوطی که از پشت اشیاء صلب می‌گذرند به خطوط شکسته، این اشیاء به‌صورت «نیمه‌شفاف» بازنمایی شده‌اند. برای برهم‌نهی بازنماییهای مجاور، صفحه را ۹۰° بچرخانید و به صورت سه‌بعدی نگاه کنید (ولی نه خیلی طولانی!).

همهٔ ژنها را در اختیار ما می‌گذارد. تنها برای اقلیت کوچکی از این ژنها ساختار سه‌بعدی پروتئینهای متناظر را در اختیار داریم.

تحلیل ساختار پروتئینها

اولین مشکل تحلیل ساختار مولکولهایی به پیچیدگیِ پروتئینها، مشکل بازنمایی است. تکنیکهای گرافیک کامپیوتری برای رسم بازنماییهای ساده‌شده‌ای از پروتئینها ارائه شده‌اند. شکل ۳ روشن می‌کند که برای یک مولکول کوچک پروتئین، تعبیر یک بازنمایی دقیق با جزئیات کامل چقدر دشوار است، و نوع تصاویر ساده‌شده‌ای را که برنامه‌ها تولید می‌کنند تا امکان دستیابی بصری به‌موضوع را به ما بدهند نشان می‌دهد. آزمایشگاههای کوچک متعدد، تعداد زیادی بازنمایی متفاوت تولید کرده‌اند؛ یعنی افراد بسیاری بازنماییهای ساده‌شدهٔ مختلفی را پیشنهاد کرده‌اند، و این پیشنهادات در بسته‌های گرافیکی عامی جمع‌آوری شده است. یک ترسیمگر مولکول ماهر آنها را با هم ترکیب خواهد کرد تا جنبه‌های مختلف یک ساختار را با درجات دقت قابل تنظیم نمایش دهد. چنین تصاویری، که به‌صورت تمام رنگی درآمده‌اند و در آنها از جلوه‌های سایه‌زنی فانتزی (اما غیرواقعی، با توجه به اندازهٔ مولکولها نسبت به طول موج نور مرئی) استفاده شده است، مجلات، پوسترها، و حتی لباسها و فنجانها را می‌آرایند.

ما اکنون ساختار ۱۰۰۰۰ پروتئین را می‌شناسیم، و طیف وسیعی از الگوهای فضایی را در آنها مشاهده می‌کنیم. در پاسخ به حرف راترفورد که معتقد است «همهٔ علم یا فیزیک است یا جمع کردن تمبر»، من می‌گویم که مطالعهٔ ساختار پروتئینها بهترین جنبه‌های این دو رشته را درهم ادغام کرده است! ما با تنوع چشمگیری مواجهیم، اما در عین حال به وجود اصول جامع در پس‌زمینه ایمان داریم.

هر پروتئین از یک زنجیرهٔ اصلی پلیمر تکرارشوندهٔ خطی (یعنی بدون شاخه) تشکیل شده است که شاخه‌های جنبی آمینواسید در فواصل منظم به آن آویخته‌اند. پروتئین با رشته‌ای از چراغهای درخت کریسمس قابل مقایسه است، که سیم آن با زنجیرهٔ اصلی تکرارشونده و رشته‌های رنگی چراغها با شاخه‌های جنبی متنوعْ متناظرند.

زنجیرهٔ اصلیْ یک خم فضایی را توصیف می‌کند که به‌وسیلهٔ تعاملهای مطلوب میان زنجیره‌های جنبی که به هم متصل شده‌اند پایدار شده است. چنین خم فضایی را در بخش میانی شکل ۳ به‌راحتی می‌توان دید. دو منطقه در جلو تصویر فرم مارپیچی دارند که محورشان تقریباً عمودی است. این یکی از دو ساختار استانداردی است که مناطق محلی زنجیره به خود می‌گیرند. ساختار استاندارد دیگر، نوارِ تقریباً بازشده است: پروتئین شکل ۳ چهار رشته نوار در خود دارد که در راستای تقریباً عمودی قرار دارند. این نوارها از پهلو بر هم اثر می‌گذارند تا اجتماع خود را پایدار کنند. در چارچوب پایینی شکل ۳، مارپیچها و نوارها، با «نماد»هایی نشان داده شده‌اند، مارپیچها با استوانه و نوارها با فلشهای بزرگ. چارچوب بالایی شکل ۳ بازنمایی ساختار را با بیشترین جزئیات، شامل زنجیرهٔ اصلی و شاخه‌های جنبی نشان می‌دهد؛ تضاد رنگها نشان‌دهندهٔ اهمیت ساده‌سازی در تولید تصویر حتی از یک پروتئین کوچک است، به‌نحوی که تصویر به لحاظ بصری قابل فهم باشد.

گام اولیهٔ تجزیهٔ یک ساختار جدید عبارت است از تعیین مناطق مارپیچ و نوار. این اطلاعاتی است که برای تبدیل بازنمایی چارچوب مرکزی شکل ۳

به چارچوب پایینی مورد نیاز است. متداولترین نوع مارپیچ در ساختارهای پروتئینی، در هر پیچ ۳٫۶ سازه را دربرمی‌گیرد. خصوصیاتی از توالی که این تناوب را نشان می‌دهند، مناطق مارپیچی را تداعی می‌کنند.

برهم‌نهی ساختارها

مانند توالیها، مسألهٔ اساسی در تحلیل ساختارها هم طرح و محاسبهٔ یک شاخص مشابهت است. فرض کنید که مجموعه‌هایی مختصاتی در اختیار داریم که دو ساختار را بازنمایی می‌کنند:

$$p_i = (x_i, y_i, z_i), \qquad i = 1, \ldots N$$

و

$$q_j = (x'_j, y'_j, z'_j), \qquad j = 1, \ldots M$$

درست مانند توالیها، در اینجا هم مسألهٔ هم‌ردیف‌سازی مطرح می‌شود. تقابل بین سه مسألهٔ مرتبط را که در شیمی محاسباتی مطرح‌اند در نظر بگیرید.

(۱) شاخص مشابهت دو مجموعه از اتمها با تناظرهای داده‌شدهٔ

$$p_i \leftrightarrow q_i \qquad i = 1, \ldots, N$$

را تعیین کنید (همتای این شاخص برای توالیها، فاصلهٔ همینگ است.) این مسأله را می‌توان دقیقاً و به نحو کارآیی حل کرد ــ این دومین «نوسانگر همساز» زیست‌شناسی مولکولی محاسباتی است.

(۲) شاخص مشابهت دو مجموعه از اتمها با تناظر نامعلوم را تعیین کنید، اما با این فرض که ساختار مولکولی آنها ــ به‌طور خاص، ترتیب خطی سازه‌ها ــ تناظر را محدود می‌کند. در مورد پروتئینها، هم‌ردیف‌سازی باید ترتیب را در طول زنجیره حفظ کند:

$$p_{i(k)} \leftrightarrow q_{j(k)}, \qquad k = 1, \ldots, K \leq N, M$$

با این قید که $k_1 > k_2 \Rightarrow i(k_1) > i(k_2)$ و $j(k_1) > j(k_2)$. این شاخص را می‌توان متناظر با فاصلهٔ لونشتاین بین رشته‌های حرفی، یا هم‌ردیف‌سازی توالیها با فواصل خالی در نظر گرفت.

(۳) شاخص مشابهت بین دو مجموعه از اتمها با تناظرهای نامعلوم، بدون هیچ محدودیتی بر تناظر، را تعیین کنید:

$$p_{i(k)} \leftrightarrow q_{j(k)}, \qquad k = 1, \ldots, K \leq N, M$$

این مسأله در حالت مهم زیر پیش می‌آید: فرض کنید دو (یا چند) مولکول تأثیرات زیست‌شناختی مشابه دارند، مثلاً دارای خواص داروشناختی یکسان هستند؛ این حالت معمولاً وقتی پیش می‌آید که ساختارها در زیرمجموعه‌ای نسبتاً کوچک از اتمهایشان که مسؤول فعالیت زیستی معینی است، و یک فارماکوفور نامیده می‌شود، مشترک باشند. برای تشخیص فارماکوفور، خوب است بتوانیم زیرمجموعه‌های ماکسیمال اتمهای دو یا چند مولکول را که ساختار مشابه دارند بیابیم.

مسائل (۲) و (۳) نیازمند تعیین نحوهٔ هم‌ردیف‌سازی نقاط هستند. روشهای هم‌ردیف‌سازی که منحصراً بر مختصات (نه بر توالیهای آمینواسید) مبتنی هستند، هم‌ردیف‌سازی ساختاری نامیده می‌شوند. در هم‌ردیف‌سازیهای ساختاری، سازه‌های متناظر یکی‌گرفته می‌شوند زیرا آنها نسبت به کل ساختار مکان مشابهی اختیار می‌کنند. باید به استخراج زیرساختار مشترک ماکسیمال و بنا نهادن هم‌ردیف‌سازی بر آن اندیشید. (برای مثال، زیرساختار مشترک ماکسیمال حروف B و R، حرف P است.) سازه‌های خارج از زیرساختار مشترک ماکسیمال قابل هم‌ردیف شدن نیستند، حقیقتی که با هم‌ردیف‌سازی جفت‌جفتِ توالیها مشخص نمی‌شود؛ این یکی از ضعفهای آن است.

کلیترین رویکرد به این سه مسأله بر حل مسألهٔ (۱)، یعنی حالت تناظر معلوم $p_i \leftrightarrow q_i$ مبتنی است. دو شیء یکسان را می‌توان با انتقال و دوران صلب یکی از آنها، برهم‌نهاد. دو شیء که مشابه هستند از طریق دوران و انتقال به برهم‌نهی تقریبی می‌رسند. اگر اشیاء مجموعه‌های مرتب نقاط باشند، شاخص مشابهت آنها برابر با جذر میانگین مربع انحرافها، Δ، پس از برهم‌نهی بهینه خواهد بود:

$$\Delta^2 = \min_{\mathrm{R,t}} \left\{ \sum_{i=1}^{N} \| \mathbf{R}p_i + \mathbf{t} - q_i \|^2 \right\}$$

که در آن $\mathbf{R}$ ماتریس دوران مناسب ($\det\mathbf{R} = 1$) و $\mathbf{t}$ بردار انتقال است. در برهم‌نهی بهینه، مکانهای میانی (به‌زبان محاوره، «مراکز ثقل») دو مجموعه برهم منطبق می‌شوند. مسألهٔ تعیین جهت نسبی صحیح به‌عنوان «مسأله پروکروتس متعامد» شناخته شده است و راه‌حلهایی مبتنی بر روشهای استاندارد جبر خطی برای آن وجود دارد [۵].

حل مسألهٔ زیرساختار مشترک ماکسیمالْ مبنای تعریف یک متریک را برای ساختارها فراهم می‌کند. بر این مبنا می‌توان مشابهتهای مقطعی و جزئی را پیدا کرد، و یک درخت طبقه‌بندی برای کل مجموعهٔ ساختارهای پروتئین ارائه داد.

رویکردهای موجود به محاسبهٔ زیرساختار مشترک ماکسیمال بر دو بازنمایی از ساختارها متکی بوده‌اند: (۱) به‌صورت فهرستهایی از مختصات $p_i = (x_i, y_i, z_i)$، $i = 1, \ldots, n$، یا (۲) به‌صورت ماتریسهای فاصلهٔ $D_p(i,j) = |p_i - p_j|$. مزیت عمدهٔ ماتریسهای فاصله این است که یک بازنمایی مستقل از مبدأ و راستا برای ساختار ارائه می‌کنند. برحسب ماتریسهای فاصله، مؤلفهٔ ماکسیمال تفاضل ماتریسهای فاصله $\max_{i,j}\{|D_p(i,j) - D_q(i,j)|\}$، شاخصی برای اختلاف ساختاری بین دو مجموعهٔ نقطهٔ هم‌ردیف‌شده به‌دست می‌دهد.

مختصات و ماتریسهای فاصله، بازنماییهایی تقریباً معادل از یک مجموعهٔ نقاط هستند. محاسبهٔ ماتریس فاصله از روی مختصات بدیهی است. اینکه آیا می‌توان مختصات را دقیقاً و مستقیماً از روی ماتریس فاصله بازسازی کرد کمتر واضح است، اما این‌کار را می‌توان به‌کمک قطری‌سازی ماتریس انجام داد [۶]. البته، ماتریس فاصله هم ساختار اولیه و هم تصویر آینه‌ای آن را به‌طور معادل مشخص می‌کند (پس دسته‌های چپ و راست متناظر، دو تصویر آینه‌ای هستند)، اما این ابهام مشکلی جدی برای کاربردهای زیست‌شناسی مولکولی نیست. اطلاعات مربوط به موقعیت و جهت هم مسلماً مفقود می‌شوند.

دشواری اصلی در محاسبات زیرساختار مشترک ماکسیمال از نوع (۲) و (۳)، پیچیدگی ترکیبیاتی ناشی از در نظرگرفتن همهٔ هم‌ردیف‌سازیهای ممکن است. معلوم شده است که الگوریتمهای مبتنی بر ماتریسهای فاصله در مواجهه با این مسأله بسیار کارآتر از الگوریتمهای مبتنی بر مجموعه‌های مختصات عمل می‌کنند. بازنماییهای ماتریسی مرتبط که به‌جای مختصات اتمی، بر مؤلفه‌های ساختاری نظیر مارپیچها و نوارها مبتنی هستند، بازنمایی فشرده‌ای از الگوهای تاخوردگی پروتئینها به‌دست می‌دهند. استخراج زیرماتریسهای مشترک ماکسیمال، بزرگترین زیرساختارهای دارای الگوی تاخوردگی مشترک را آشکار می‌کند. به‌علاوه چنین بازنماییهایی، امکان شمارش همهٔ الگوهای تاخوردگی پروتئین را در اختیار ما می‌گذارند. به‌صورت تجربی برآورد شده است که همهٔ پروتئینهای طبیعی کمتر از حدود ۱۰۰۰ الگوی تاخوردگی دارند. شمارش کامل به ما امکان می‌دهد که انتخابهای طبیعت را آزمایش کنیم، و سعی کنیم تصادف تاریخی و ضرورت معماری را از هم تمیز دهیم.

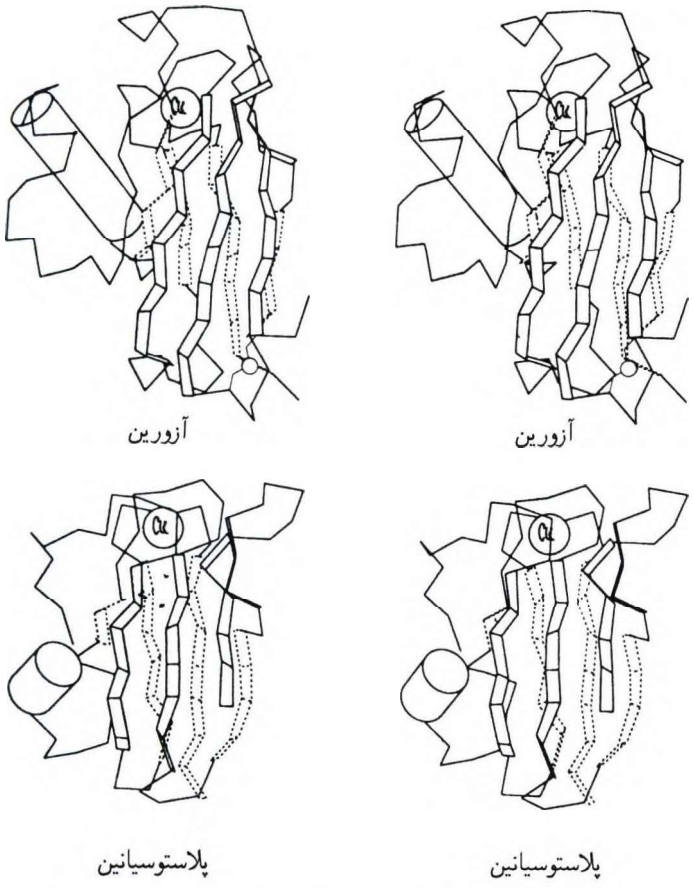
تکامل پروتئین

مطالعهٔ تکامل پروتئین یعنی بررسی اینکه چگونه توالیهای آمینواسید و ساختارهای پروتئین متناظر در گونه‌های وابسته با هم اختلاف پیدا می‌کنند. این تحقیقی از نوع اطلاع‌دهنده است، که به ما در فهم روابط توالی-ساختار کمک می‌کند. زیر اگرچه ما می‌دانیم که یک توالی آمینواسید منفرد همهٔ اطلاعات لازم برای مشخص کردن ساختار پروتئین را در خود دارد، هنوز نمی‌فهمیم که چگونه باید از توالی به ساختار رسید. این را «صورت انتگرالی» مسألهٔ تاخوردگی پروتئین در نظر بگیرید، که مسألهٔ حل نشده‌ای است. در بررسی تکامل پروتئین، مشاهده می‌کنیم که چگونه تغییرات توالیها در تغییرات ساختار منعکس می‌شوند؛ این را آسانتر می‌توان فهمید. این مسأله را «صورت دیفرانسیلی» مسأله تاخوردگی پروتئین در نظر بگیرید.

موضوع:	تاخوردگی پروتئین	تکامل پروتئین
مشاهده:	توالی ← ساختار	تغییر در توالی ← تغییر در ساختار
صورت مسأله:	«صورت انتگرالی»	«صورت دیفرانسیلی»
وضعیت مسأله:	حل‌نشده	حل‌نشده اما باید ساده‌تر باشد

استدلال ساده‌ای بر این دلالت می‌کند که ساختار باید تابعی تقریباً «پیوسته» از توالی باشد، حداقل برای توالیها و ساختارهایی که به‌صورت طبیعی شکل گرفته‌اند. فرض کنید پروتئینی وجود می‌داشت که در آن هر جهشی (هر تغییری در توالی آمینواسید) یک ساختار ناپایدار تولید می‌کرد. در این صورت طبیعت هرگز نمی‌توانسته است با فرآیندهای تکاملی به چنین ساختاری برسد، زیرا هیچ صورت اولیهٔ پایداری نمی‌توانسته برای آن وجود داشته باشد. پس چنین نتیجه می‌شود که ساختارهای طبیعی باید مستحکم باشند. اغلب تغییرات کوچک در توالی باید تغییری در ساختار ندهند. (این شرط برای ساختارهای پروتئینی که به‌روشهای مصنوعی ساخته می‌شوند برقرار نیست.)

درواقع، پروتئینهای طبیعی با توالیهای بسیار شبیه به‌هم، ساختارهای بسیار شبیه به‌هم دارند. پیش از آنکه انسولین مصنوعی انسانی در دسترس قرارگیرد، انسولین خوک درمان بالینی مؤثری برای انسانهای مبتلا به مرض قند بود، با



شکل ۴ در حین تکامل، جهشهای توالیهای ژن بروی هم انباشته می‌شوند، و در نتیجه توالیها و ساختارهای پروتئین از هم فاصله می‌گیرند. این شکل دو پروتئین مرتبط انتقال‌دهندهٔ الکترون، یعنی پلاستوسیانین برگ سپیدار و آزورین باکتریایی را نشان می‌دهد. بخشی از ساختارها که در نیمهٔ راست تصویر واقع است و شامل آرایه‌های پر و نقطه‌چین مناطق «روبانی‌شکل» ــ که نوار نامیده می‌شوند ــ و محل پیوند با مس است، به‌خوبی در حین تکامل ماندگار است، در حالی‌که بخشی از ساختار که در نیمهٔ چپ تصویر واقع است به‌شدت تغییر کرده است.

آنکه توالیهای آمینواسید انسولین خوک و انسان یکسان نیستند. اعتماد به چنین مشابهتهایی روشی برای پیش‌بینی ساختار پروتئینها از روی ساختارهای بسیار نزدیکِ شناخته‌شده به‌دست می‌دهد، که به‌عنوان «مدلسازی مانستگی»[1] شناخته شده است. اما، همچنان‌که تکامل ادامه می‌یابد، توالیها و ساختارها نهایتاً به شکل بنیادیتری واگرا می‌شوند. شکل ۴ دو پروتئین پلاستوسیانین و آزورین را نشان می‌دهد که با هم ارتباط کمی دارند، در منطقهٔ راست شکل، دو نوار وجود دارد که روبه‌رو بسته شده‌اند، و «هستهٔ» ماندگار ساختار را تشکیل می‌دهند، در حالی‌که منطقهٔ حلقوی بلند سمت چپ از شکلی کاملاً متفاوت در دو ساختار برخوردار است.

پیش‌بینی ساختار پروتئین

طبیعت الگوریتمی در اختیار دارد که به کمک آن ساختار سه‌بعدی پروتئین را فقط از روی توالی آمینواسید آن تعیین می‌کند. ما باید قادر به کشف آن

1. homology madelling

باشیم. سپس باید بتوانیم ساختار پروتئینها را که در توالیهای ژن انسان و سایر ژنومها به‌صورت ذاتی قرار دارد پیش‌بینی کنیم، و آنها را در مسائل عملی نظیر طراحی دارو به‌کار گیریم. معلوم شده است که مسألهٔ پیش‌بینی ساختار پروتئین مسألهٔ بسیار دشواری است. رویکردهای بسیاری اختیار شده‌اند، و ادعاهای بسیاری مطرح شده‌اند. اما در حال حاضر هیچ روش محاسباتی که بتواند به‌صورت سازگار حتی یک پیش‌گویی صحیح کیفی از ساختار پروتئین براساس توالی آمینواسید ارائه کند وجود ندارد، مگر آنکه یک پروتئین بسیار مشابه موجود باشد.

فرض کنید که توالی آمینواسیدهای یک پروتئین جدید به شما داده شده بود، و از شما خواسته بودند که ساختار آن را پیش‌بینی کنید. شما باید سعی می‌کردید چه چیزی را پیش‌بینی کنید؟ کاملترین اطلاعاتی که یک پیش‌بینی ممکن است به‌دست دهد مجموعهٔ کاملی از مختصات سه‌بعدی مدل یک پروتئین نهایی است ــ یعنی یک پیش‌بینی سه‌بعدی. یک هدف کمتر بلندپروازانه می‌تواند پیش‌بینی این باشد که مناطق مارپیچ و نوار در کجای توالی پدیدار می‌شوند ــ یعنی یک پیش‌بینی یک‌بعدی. جایی بین این‌دو، پیش‌بینی‌هایی قرار دارند که از پیش‌بینیهای ساختار ثانویهٔ یک‌بعدی فراتر می‌روند، اما تنها برخی اطلاعات کیفی در بارهٔ آرایش فضایی کلی الگوی تاخوردگی ارائه می‌کنند ــ بیایید اینها را پیش‌بینیهای دوبعدی بخوانیم.

از چه نوع اطلاعاتی می‌توان در پیش‌بینی ساختار پروتئین استفاده کرد؟ هدف نهایی، رویکردِ ابتدا به ساکن[1] «محض» است ــ تنها از توالی پروتئین مقصد استفاده کنید و نه هیچ چیز دیگر. باید به‌خاطر داشت که این همان کاری است که طبیعت می‌کند ــ پروتئینها وقتی می‌خواهند تا شوند در پایگاههای دادهٔ وب جستجو نمی‌کنند. اما ما می‌توانیم جستجو کنیم، و موفقیتهایی هم در استفاده از اطلاعات بانکهای داده برای شناسایی تاخوردگی یک پروتئین مقصد از روی ساختارهای شناخته‌شده حاصل شده است. این مسأله به نام تشخیص تاخوردگی معروف شده است. البته این روش تنها در صورتی مؤثر است که ساختار یک یا چند پروتئین با تاخوردگی مشابه پروتئین مقصد در پایگاه دادهٔ شما موجود باشد.

چه کسی را باید متقاعد کرد؟ فهرست زیر به‌گونه‌ای مرتب شده است که تقریباً از بالا به پایین سختگیری کم می‌شود. اغلب دانشمندان پذیرفته‌اند که اقناع «حامیان مالی طرح» از همه مهمتر است!

چه‌کسی را باید متقاعد کرد؟

۱. بلورشناسان
۲. متخصصان طیف‌شناسی NMR
۳. حامیان مالی طرح
۴. داوران مقاله‌ها
۵. همکاران
۶. مادرتان

1. ab initio

درواقع شما چگونه می‌توانید مردم را متقاعد کنید که روش موفقی برای پیش‌بینی ساختار پروتئین در اختیار دارید؟ دو نوع از ادعاها از اساس غیرقابل آزمون هستند. یکی آنکه شما می‌توانید ساختار پروتئینی را که ساختار آن از قبل دانسته شده است پیش‌بینی کنید. دیگر آنکه شما ساختار پروتئینی را پیش‌بینی کرده‌اید که ساختار تجربی آن ناشناخته است و به احتمال زیاد تا مدت طولانی ناشناخته خواهد ماند. باید در حوزهٔ بینابینی میان شناخته‌شده‌ها و آنچه تا زمان طولانی غیرقابل شناسایی است کار کرد، و پیش‌بینی ساختار را با تعیین ساختارهای در دست اقدام هماهنگ نمود.

برای نظم بخشیدن به این فعالیت، برای پاداش دادن به کسانی که موجب پیشرفتهای اصیل شده‌اند، و برای تکذیب ادعاهای آنانی که مصراً ادعا می‌کردند «مسأله پیش‌بینی ساختار پروتئین را حل کرده‌اند»، جان مالت[1] ایدهٔ آزمونهای کورِ سازمان‌یافته را مطرح کرد. ایده این است که دانشمندان در حین فرایند کشف ساختارها، توالیهای آمینواسید را به همه اعلام کنند، اما قول بدهند که ساختار را تا زمان مورد توافقی مخفی نگه دارند. همهٔ کسانی که باور دارند که روشی برای پیش‌بینی ساختار پروتئین در اختیار دارند می‌توانند پیش‌بینیهای خود را تا قبل از تاریخ اعلام ساختار ارائه دهند. پس از اعلام، می‌توان پیش‌بینیها را با تجربه مقایسه کرد ــ که این باعث خرسندی عدهٔ اندکی و تأسف گروه کثیری خواهد شد. این ایده تحت برنامهٔ CASP ــ برآورد انتقادی پیش‌بینی ساختار[2] ــ براساس یک دورهٔ دوساله ارائه شده است.

روشهای پیش‌بینی به دو طبقهٔ کلی تقسیم می‌شوند، استقرایی و استنتاجی. در روشهای استقرایی مستقیماً از بانکهای دادهٔ توالیها و ساختارها استفاده می‌شود. روشهای استنتاجی، رویکردهای واقعاً ابتدا به ساکن هستند ــ حالت جزیرهٔ بی‌آب و علف ــ که هدف آنها پیش‌بینی ساختار پروتئین براساس اصول عام فیزیک، شیمی و زیست‌شناسی، بدون ارجاع صریح به توالیها و ساختارهای شناخته‌شده است. البته پیشرفت روشهای ابتدا به ساکن بستگی به چیزهایی دارد که از بررسی توالیها و ساختارها آموخته‌ایم. تمایز این روشها در این است که درک حاصل از این بررسی به شکل اصول عامی خلاصه شده است که می‌توان آنها را بدون مراجعه به اطلاعات خاص در پایگاههای داده، به‌کار برد.

روشهایِ پیش‌بینی ابتدا به ساکن را می‌توان به دو رویکرد تقسیم کرد، که من آنها را «طبیعی» و «زوری» می‌نامم. رویکرد طبیعی به دنبال درک فرآیند تاخوردگی طبیعی می‌گردد و سپس سعی می‌کند آن‌را شبیه‌سازی کند. رویکرد «زوری» هر فرایندی را که بتواند زنجیره را به شکل [=ترکیب، ساختمان] مناسب برساند مجاز می‌داند، حتی اگر این فرایند در راستای مسیری انجام شود که طبیعی نباشد یا حتی از نظر فیزیکی غیرقابل تحقق باشد.

شواهدی هست که انتخاب طبیعی نه تنها وضعیت طبیعی نهایی پروتئینها، بلکه مسیر تاخودگی آنها را نیز شکل داده است. زیرا نه تنها پروتئینها باید به گونه‌ای تکامل‌یافته باشند که شکل فعال پایداری پیدا کرده باشند، بلکه باید در زمان معقولی، از یک وضعیت تانخوردهٔ اولیه شامل مخلوطی از شکلهای تصادفی به چنین شکلی دست یافته باشند. یک محاسبهٔ ساده براساس

1. John Moult 2. Critical Assessment of Structure Prediction

میزان جهشهای اتمی در محلول نشان می‌دهد که سرعت پیمایش فرسایندهٔ فضای شکلهای ممکن، چند مرتبهٔ بزرگی کمتر از اندازهٔ کافی خواهد بود. (این مطلب گاهی پارادوکس لوینتال[1] نامیده می‌شود.) هیچ مدرکی وجود ندارد که نشان دهد مسیر تاخوردگی واقعاً بر وضعیت نهایی تأثیر می‌گذارد، اگرچه به لحاظ نظری این امر ممکن است. اگر حالات تاخوردگی دیگری ممکن بودند، اما مسیر چنان پیش می‌رفت که به یکی از آنها منجر شود، آنگاه ما پیش‌گویان مجبور بودیم رویکرد طبیعی، نه زوری، را بپذیریم.

مانع پیش‌بینی ساختار کجاست؟ ما فکر می‌کنیم نیروهایی را که شکلهای طبیعی پروتئین را پایدار می‌کنند می‌شناسیم. حتی ممکن است بتوانیم تابع انرژی شکل‌گیری صریحی از مختصات بنویسیم. کاری که باید انجام دهیم کمینه کردن آن است. اما این مهم است که تشخیص دهیم که پروتئینها، به بیان ترمودینامیکی، تنها در حاشیه پایدارند. درواقع انرژی شکل‌گیری یک پروتئین تاخورده برابر با اختلاف بسیار کوچکی بین عبارات متضاد بزرگ است، چیزی که برای تحلیلگران عددی کابوس است.

آیا مشکل این است که نمی‌توانیم تابع انرژی را به‌اندازهٔ کافی دقیق بنویسیم، یا تابع چنان پیچیده است که نمی‌توانیم بهینه‌اش کنیم؟ یک آزمون، کمینه کردن انرژیهای شکل‌گیری پروتئینهاست، با شروع از حالتهای طبیعی شناخته‌شدهٔ آنها. چنین محاسباتی به سمت شکلهای با کمترین انرژی نزدیک به نقطهٔ شروع همگرا می‌شوند، و نشان می‌دهند که توابع انرژی در همسایگی پاسخ صحیح رضایت‌بخش هستند. (این مطلب چندان شگفت‌انگیز نیست، زیرا توابع براساس تنظیم پارامترها تعریف شده‌اند به‌گونه‌ای که حالات طبیعی مشاهده شده را باز تولید کنند.) اما این کافی نیست. تابعی که در همسایگی نقطهٔ مینیمم صحیح است الزاماً مجموعهٔ کاملی از مسیرهای پیموده‌شده در فضای شکلها را ــ که یک برنامه را قادر می‌سازند تا با شروع از یک نقطهٔ دلخواه مینیمم سراسری را بیابد ــ در اختیار ما قرار نمی‌دهد.

دو مسأله وجود دارد. نخست اینکه بسیاری از نیروهایی که پروتئینها را پایدار می‌کنند کوتاه‌برد هستند. حتی اگر تابع انرژی را دقیقاً می‌دانستیم، و اگر کمینه‌سازی را از یک شکل به‌طور تصادفی گسترش‌یافتهٔ نافشرده شروع می‌کردیم، ممکن بود به این نتیجه برسیم که هیچ نیروی بلندبردی وجود ندارد که سیستم را به سمت ساختار صحیح براند. دوم اینکه، حتی اگر با فروریزی سیستم به یک حالت فشرده برسیم، نمای انرژی به‌عنوان تابعی از مختصات چندین مینیمم موضعی را شامل می‌شود که توسط حایلهای بلند از هم جدا شده‌اند. بسیاری از این مینیممهای موضعی کاندیداهایی برای حالت طبیعی خواهند بود. پروتئینهای حقیقی به‌وسیلهٔ ترکیبی از (۱) «پردازش موازی» سنگین که در آن همهٔ سازه‌ها همزمان ابعاد موضعی خود را در فضای شکل می‌کاوند، و (۲) تکامل مسیرهای تاخوردگی که سیستم را به سمت جواب صحیح هدایت می‌کنند، بر این مسائل غلبه کرده‌اند. کامپیوترهای ما نمی‌توانند به پردازش موازی دست یابند، توابع انرژی ما نمی‌توانند مسیرهای تاخوردگی بلندبرد را توجیه کنند، و الگوریتمهای ما نمی‌توانند به‌سادگی مینیمم سراسری یک تابع پیچیدهٔ چندمتغیرهٔ غیرخطی را بیابند. (این نوسانگر همساز، وقتی لازمش داریم کجاست؟!)

دشواری حالت پیشینی، همان‌طور که متوجه شده‌ایم، به پیدایش روشهای تجربی مبتنی بر توالیها و ساختارهای شناخته‌شده منجر شده است. روشهای پیش‌بینی که از بانکهای داده استفاده می‌کنند عبارت‌اند از (۱) روشهایی برای مدلسازی مانستگی ــ پیش‌بینی ساختار مقصد از روی یک پروتئین خیلی مشابه که ساختار آن کاملاً شناخته شده است؛ و (۲) روشهای تشخیص تاخوردگی ــ تخمین‌زدن تطابق‌پذیری توالی آمینواسید با مجموعهٔ الگوهای شناخته‌شدهٔ تاخوردگی پروتئین. این روشها تا حدی (اما نه کاملاً) به‌دلیل رشد بانکهای داده کاملتر شده‌اند. هر چه توالیها و ساختارهای بیشتری شناخته شوند، احتمال آنکه یک پروتئین جدید مشابه پروتئینی باشد که قبلاً شناخته شده است بیشتر می‌شود. برعکس، روشهای ابتدا به ساکن کندتر پیش می‌روند. بعد از یکی از رقابتهای اخیر CASP، یکی از اظهارنظرهای توأم با دلخوری در بارهٔ این روشها این بود که حداقل «شکست از این پس تضمین‌شده نیست [۸].» فرد بدبین ممکن است پیش‌بینی کند که رشد بانکهای داده به این معنی خواهد بود که روشهای مبتنی بر اطلاعات راه‌حلهای عمل‌گرایانه‌ای برای چنان اکثریت بزرگی از سؤالات ارائه خواهند کرد، که اشتیاق نسبت به حمایت از توسعهٔ روشهای ابتدا به ساکن نقصان خواهد یافت. بنابراین مایه شرمساری خواهد بود که زیست‌شناسی محاسباتی، یکی از جالبترین محاسبات زیست‌شناختی را از دست داده باشد!

1. Levinthal

مراجع

1. Wigne, E.P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics* 13, 1-14.
2. Sankoff, D. and Kruskal, J.B., eds. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, Mass.
3. Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.
4. Fitch, W.M. and Smith, T.F. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* 80, 1382-1386.
5. Golub, G. and van Loan, C., *Matrix Computations.* Johns Hopkins Press. Baltimore, 2nd ed. 1989.
6. Young, G. and Householder, A.S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3, 19-22. (For background and history see [7].)
7. Blumenthal, L.M. (1938). *Distance Geometries. A study of the development of abstract metrics.* University of Missouri Studies 13, #2.
8. *New York Times*, March 25, 1997.

******

- Arthur M. Lesk, "The unreasonable effectiveness of mathematics in molecular biology", *Math. Intelligencer*, (2), **22** (2000) 28-37.

* آرتر لسک، دانشگاه کیمبریج، انگستان