مجله علوم آماری، پاییز و زمستان ۱۳۹۴

جلد ۹، شماره ۲، ص ۲۵۳-۲۶۷

مدل‌های بیزی در انطباق پروتئین‌ها

سید مرتضی نجیبی[۱]، موسی گل‌علی‌زاده[۲]، محمدرضا فقیهی[۱]

[۱] گروه آمار، دانشگاه شهید بهشتی

[۲] گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۳/۱۰/۲ تاریخ آخرین بازنگری: ۱۳۹۴/۲/۲۸

چکیده : در این مقاله راه‌حل‌های احتمالاتی مسئله انطباق ساختار سوم پروتئین و تفاوت آن با الگوریتم‌های قطعی مطالعه و بررسی می‌شود. برای این منظور دو مدل احتمال بیزی معرفی و راه حلی در خصوص اضافه کردن اطلاعات توالی و نوع اسید آمینه (ساختار اول) به آن ارائه خواهد شد. همچنین نحوه برآورد پارامترهای انطباق به کمک الگوریتم مونت کارلوی زنجیر مارکف و نمونه‌گیری از توزیع پسین معرفی می‌شود. در نهایت نحوه کاربست این روش‌ها در انطباق پروتئین‌ها نشان داده شد و برآورد پارامترها در مدل‌های متفاوت برای یک مجموعه داده واقعی ارزیابی و مقایسه خواهد شد.

واژه‌های کلیدی : آمار شکل، فاصله اندازه–شکل، بیوانفورماتیک، انطباق ساختاری پروتئین، ساختار سوم پروتئین، استنباط بیزی.

آدرس الکترونیک مسئول مقاله: سید مرتضی نجیبی، sm_najibi@sbu.ac.ir

کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۲P۱۰، ۶۵D۱۸

۱ مقدمه

تاکنون الگوریتم‌های بسیاری در حوزه بیوانفورماتیک ساختاری به منظور رده‌بندی ساختار هندسی پروتئین‌ها معرفی شده‌اند که عموماً بر پایه انطباق زوجی شکل هندسی پروتئین‌ها هستند. هدف اصلی این روش‌ها، انطباق پروتئین‌ها بر یکدیگر به گونه‌ای است که بعد از دوران و تغییر مکان، فاصله نقاط جورشده از یکدیگر حداقل و تعداد نقاط جور شده ماکسیمم باشد. از آن‌جا که بررسی تمامی حالت‌های ممکن امکان‌پذیر نیست، تقریبا در تمامی روش‌ها از روش‌های اکتشافی [۱] برای دستیابی به مدل بهینه استفاده شده است. اما در عمل این الگوریتم‌ها لزوماً جواب یکسانی برای یک مسئله ارائه نمی‌دهند. به عنوان مثال میزان شباهت دو روش مرسوم انطباق ماتریسی [۲] ($DALI$) و تعمیم ترکیبی [۳] (CE) حداکثر ۴۰ درصد است (شیندیالو و بورنه، ۲۰۰۰). از این‌رو به نظر می‌رسد مقایسه آماری پروتئین‌ها تنها به کمک فاصله آن‌ها از یکدیگر، اطلاعات واقعی راجع به ساختار هندسی‌شان را حذف کرده و در نتیجه جواب نهایی تحلیل، رضایت‌بخش نیست. لذا انتظار می‌رود در نظر گرفتن ساختار هندسی پروتئین به عنوان منبع اطلاعات ذی‌قیمت آماری بتواند در ارزیابی میزان دقیق مشابهت و تفاوت ساختمان آن‌ها مفید و موثر واقع شود.

برای مقایسه اشکال هندسی در آمار شکل، مدل‌های آماری متنوعی معرفی شده‌اند (درایدن و ماردیا، ۱۹۹۸). این روش‌ها در مقایسه با روش‌های قطعی، از رویکرد احتمالاتی استفاده می‌کنند و در یک دید کلی هدف‌شان مدل‌سازی خطاهای تصادفی است (تیلور و همکاران، ۲۰۰۳). اطلاعات اولیه آماری شکل در این الگوریتم‌ها به کمک نقاط شاخص [۴] به‌دست می‌آید. اما در مسائلی همچون انطباق پروتئین‌ها، تناظر بین نقاط یا برچسب [۵] آن‌ها مشخص نیست. از این‌رو برخی از محققان، مانند کنت و همکاران (۲۰۰۴)، گرین و ماردیا (۲۰۰۶)، درایدن و همکاران

---

۱ Heuristic

۲ Distance alignment matrix

۳ Combinatorial extension

۴ Landmarks

۵ Label

(۲۰۰۷)، کنوبی و درایدن (۲۰۱۲) و ماردیا و همکاران (۲۰۱۳) مطالعاتی را در زمینه تحلیل اشکال بدون برچسب انجام داده‌اند. اما استفاده از این روش‌ها در انطباق ساختاری پروتئین، به راحتی و سرعت الگوریتم‌های قطعی نیست. یکی از راه‌حل‌های بهبود این مدل‌ها، اضافه کردن اطلاعات اضافی پروتئین‌ها در قالب توزیع پیشین آگاهی‌بخش است (فالیزی و همکاران، ۲۰۱۴؛ نجیبی و همکاران، ۲۰۱۵).

در این مقاله دو مدل احتمالاتی انطباق زوجی اشکال بدون برچسب معرفی و نحوه اضافه کردن اطلاعات مربوط به توالی اسیدهای آمینه و نوع آن‌ها به کمک پیشین‌های آگاهی‌بخش تجربی ارائه خواهد شد. همچنین کاربست این مدل‌ها در یک مجموعه داده واقعی و میزان تأثیر اضافه کردن اطلاعات ساختاری پروتئین به مدل‌های قبلی مورد مطالعه قرار گرفته است.

۲ مدل‌های بیزی انطباق پروتئین‌ها

فرض کنید مجموعه $\boldsymbol{X}$ با k نقطه و مجموعه $\boldsymbol{\mu}$ با ℓ نقطه $(k \neq \ell)$ در فضای m بعدی$(m \geq ۲)$ در اختیار باشند. در اصطلاح آمار شکل به این دو مجموعه پیکره یا پیکربندی و به فضایی که مشاهدات در آن قرار می‌گیرند، فضای پیکربندی گویند (درایدن و ماردیا، ۱۹۹۸). از آنجا که برچسب گذاری بین نقاط نامعلوم است، مسئله انطباق پروتئین در واقع یافتن زیرمجموعه‌هایی از دو پیکره $\boldsymbol{X}$ و $\boldsymbol{\mu}$ تعریف می‌شود به گونه‌ای که بعد از انجام تبدیلات انتقال[۶] و دوران، فاصله این دو زیرمجموعه از یکدیگر حداقل شود. این فاصله در واقع اندازه‌ای است که در فضای مشاهدات بعد از انجام تبدیلات انتقال و دوران تعریف می‌شود.

در این مسئله مختصات اتم‌های کربن-آلفا در هر پروتئین، نقاط شاخص در دو پیکره $\boldsymbol{X}$ و $\boldsymbol{\mu}$ هستند. تناظر بین این نقاط از طریق برآورد ماتریسی به‌نام ماتریس انطباق[۷] $(\boldsymbol{\Lambda})$ به‌دست می‌آید که دارای k سطر و $۱ + \ell$ ستون است. حداکثر یکی از عناصر هر سطر این ماتریس یک و مابقی صفر هستند. ستون $۱ + \ell$ این ماتریس

[۶] Translation

[۷] Match matrix

نشان دهنده نقاطی از $\boldsymbol{X}$ است که با هیچ نقطه‌ای از $\boldsymbol{\mu}$ جور نشده‌اند. بنابراین بر اساس این ماتریس، در صورتی که فرض شود تعداد p نقطه از دو پیکربندی با یکدیگر جور شده‌اند، می‌توان ماتریس‌های $\boldsymbol{X}$ و $\boldsymbol{\mu}$ را به‌صورت

$$\boldsymbol{X} = (\boldsymbol{X}^M, \boldsymbol{X}^N), \quad \boldsymbol{\mu} = (\boldsymbol{\mu}^M, \boldsymbol{\mu}^N)$$

افراز کرد، که در آن $\boldsymbol{X}^M_{p\times m}$ و $\boldsymbol{\mu}^M_{p\times m}$ به ترتیب دو زیرمجموعه از نقاط $\boldsymbol{X}$ و $\boldsymbol{\mu}$ هستند که با یکدیگر جور شده‌اند و $\boldsymbol{X}^N_{(k-p)\times m}$ و $\boldsymbol{\mu}^N_{(\ell-p)\times m}$ نیز دو زیرمجموعه از نقاط $\boldsymbol{X}$ و $\boldsymbol{\mu}$ هستند که با یکدیگر جور نشده‌اند.

مدل‌های انطباق بیزی این مسئله بر اساس نحوه برآورد تبدیلات انتقال و دوران، به دو گروه کلی تقسیم‌بندی می‌شوند. در گروه اول، مطابق لی (۱۹۸۸) و درایدن و همکاران (۲۰۰۷)، مدل انطباق در فضای غیر اقلیدسی اندازه-شکل تعریف می‌شود که در آن بردار انتقال γ و ماتریس دوران Γ به کمک تحلیل پروکراستس به‌دست می‌آیند و جزء پارامترهای اصلی مدل محسوب نمی‌شوند. در گروه دوم، مدل‌بندی احتمالاتی انطباق بر اساس ماردیا (۲۰۰۶) و کنوبی و همکاران (۲۰۱۲) به‌طور مستقیم بر روی نقاط شاخص در فضای پیکربندی تعریف می‌شود و بردار انتقال γ و ماتریس دوران Γ جزء پارامترهای اصلی مدل هستند. در ادامه این مدل‌ها معرفی و در بخش سوم نحوه بهبود آنها مورد مطالعه قرار گرفته است.

۱.۲ مدل احتمال در فضای اندازه-شکل

بدون از دست دادن کلیت مسئله فرض کنید پیکربندی $\boldsymbol{X}_{k\times ۳}$ تصادفی و پیکربندی $\boldsymbol{\mu}_{l\times ۳}$ ثابت است به نحوی که p نقطه شاخص از پیکربندی $\boldsymbol{X}$ با p نقطه از $\boldsymbol{\mu}$، نظیر به نظیر جور شده‌اند. به‌علاوه فرض کنید دو پیکربندی $\boldsymbol{X} = C_p\boldsymbol{X}$ و $\boldsymbol{\mu} = C_p\boldsymbol{\mu}$، قبلاً توسط ماتریس مرکزی کننده $C_p = I_p - \mathbf{1}_p\mathbf{1}_p^t/p$ مرکزی شده‌اند، که در آن I_p ماتریس همانی p بعدی و $\mathbf{1}_p$ برداری شامل p مقدار یک است. توجه شود که $\mathbf{1}_p^t\boldsymbol{X} = \mathbf{0}_m$ و $\mathbf{1}_p^t\boldsymbol{\mu} = \mathbf{0}_m$، که در آن $\mathbf{0}_m$ برداری شامل p مقدار صفر است. مجموعه p نقطه در فضای سه بعدی که تحت تبدیلات انتقال و دوران پایا است،

مانند یک نقطه در فضای اندازه–شکل است که با $S\Sigma_3^p$ نمایش داده می‌شود. اگر اندازه مرکزی ماتریس $\boldsymbol{X}$ به‌صورت $S(\boldsymbol{X}) = ||X|| = \sqrt{\boldsymbol{X}^t\boldsymbol{X}}$ در نظر گرفته شود، فاصله اندازه–شکل که در واقع متر ریمانی بین دو نقطه $\boldsymbol{X}$ و $\boldsymbol{\mu}$ در فضای $S\Sigma_3^p$ است (لی، ۱۹۸۸)، به‌صورت

$$d_S^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M) = \inf_{\Gamma\in SO(m),\gamma\in R} ||(\boldsymbol{\mu}^M)^t - \Gamma(\boldsymbol{X}^M)^t - \gamma \times \mathbf{1}_p||^2$$

تعریف می‌شود، که در آن $\Gamma_{m\times m}$ ماتریس دوران ($|\Gamma| = 1$ و $\Gamma\Gamma^t = \Gamma^t\Gamma = I_3$)، γ بردار انتقال و $SO(m)$ فضای ماتریس‌های دوران (تبدیلات متعامد مخصوص) هستند (کندال و همکاران، ۱۹۹۹). جواب کمترین توان‌های دوم این فاصله، همان فاصله پروکراستس خواهد شد. اما از آنجا که در پروتئین‌ها نقاط شاخص در فضای سه‌بعدی قرار دارند، به منظور برآورد این فاصله در فضای اندازه– شکل از تقریب آن در فضای مماس استفاده می‌شود (کنت و ماردیا، ۲۰۰۱). بنابراین می‌توان از توزیع نرمال چند متغیره در فضای مماس بر اندازه–شکل $\boldsymbol{\mu}$، برای پرشیدگی نقاط جور شده به‌صورت (درایدن و همکاران، ۲۰۰۷)

$$f(\boldsymbol{X}^M) \propto (2\pi)^{-\frac{Q}{2}}\tau^{\frac{Q}{2}} \exp\{-\frac{\tau}{2}\hat{d}_S^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M)\}$$

استفاده نمود، که در آن $\hat{d_S}^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M)$ برآورد فاصله اندازه–شکل بین نقاط جورشده در دو پیکربندی در فضای مماس، $\tau = 1/\sigma^2$ پارامتر دقت و $Q = pm - m(m-1)/2 - m$ بعد فضای اندازه–شکل است. در مقابل فرض می‌شود که $k-p$ نقطه جورنشده $\boldsymbol{X}^N$ به‌صورت مستقل دارای تابع چگالی یکنواخت در فضای کراندار $\mathcal{A}$ با حجم $|\mathcal{A}|$ هستند، یعنی

$$f(\boldsymbol{X}^N) \propto |\mathcal{A}|^{k-p}, \quad \boldsymbol{X}^N \in \mathcal{A}^{k-p}.$$

در این صورت تابع درستنمایی پیکربندی $\boldsymbol{X}$ به شرط معلوم بودن $\boldsymbol{\Lambda}$، $\boldsymbol{\mu}$ و τ به‌صورت

$$\begin{aligned} L(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau) &\propto f(\boldsymbol{X}^M|\boldsymbol{\Lambda}, \tau, \boldsymbol{\mu}) f(\boldsymbol{X}^N|\boldsymbol{\Lambda}) \\ &\propto |\mathcal{A}|^{k-p}(2\pi)^{-Q/2}\tau^{Q/2} \exp\{-\frac{\tau}{2}d_S^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M)\} \end{aligned}$$

خواهد بود. دو پارامتر τ و $\boldsymbol{\Lambda}$ پارامترهای اصلی این مدل‌بندی هستند. فرض کنید توزیع پیشین پارامتر τ به‌صورت $\Gamma(\alpha_\circ, \beta_\circ)$ و احتمال پیشین برای هر سطر ماتریس $\boldsymbol{\Lambda}$ به‌صورت

$$P(\lambda_{i(\ell+1)} = 1) = \psi \;\; P(\lambda_{ij} = 1) = \frac{1-\psi}{\ell} \;\; 1 \le j \le \ell,\; i = 1, \ldots, k$$

تعریف شود، که در آن $0 \le \psi \le 1$. بنابراین چگالی پسین توأم τ و $\boldsymbol{\Lambda}$ به شرط مشاهدات $\boldsymbol{X}$ و $\boldsymbol{\mu}$ به‌صورت

$$\pi(\tau, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{\mu}) = \frac{\pi(\tau)\pi(\boldsymbol{\Lambda})L(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau)}{\sum_{\boldsymbol{\Lambda}} \int_\circ^\infty \pi(\tau)\pi(\boldsymbol{\Lambda})L(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau)}$$

خواهد شد. به دلیل عدم وجود صورت بسته برای توزیع پسین و همچنین به‌علت ساختار خاص پارامترها، امکان استنباط مستقیم بر اساس توزیع پسین وجود ندارد. لذا درایدن و همکاران (۲۰۰۷) پیشنهاد دادند که با استفاده از روش‌های شبیه‌سازی $MCMC$ و الگوریتم‌های گیبز و متروپولیس–هستینگز، برآورد ماکسیمم توزیع پسین[۸] (MAP) پارامترها انجام شود. برای این منظور نیاز است توزیع‌های شرطی کامل دو پارامتر τ و $\boldsymbol{\Lambda}$ تعیین شود. به سادگی می‌توان دریافت که توزیع شرطی کامل τ به شرط بقیه پارامترها به‌صورت

$$\tau|(\boldsymbol{X}, \boldsymbol{\Lambda}, \boldsymbol{\mu}) \sim \Gamma\left(\alpha_\circ + \frac{Q}{2}, \beta_\circ + \frac{1}{2} d_S^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M)\right).$$

خواهد بود. بنابراین می‌توان به راحتی با تولید نمونه از توزیع پسین شرطی کامل τ در هر مرحله از شبیه‌سازی با قدم‌های گیبز این پارامتر را به‌روز کرد. به علاوه توزیع شرطی کامل $\boldsymbol{\Lambda}$ به‌صورت

$$\begin{aligned} \pi(\boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{\mu}, \tau) &\propto \pi(\boldsymbol{\Lambda})L(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau) \\ &\propto |\mathcal{A}|^{k-p}(2\pi)^{-Q/2}\tau^{Q/2} \;\; \exp\{-\frac{\tau}{2} d_S^2(\boldsymbol{X}^M, \boldsymbol{\mu}^M)\} \end{aligned}$$

است. اما به دلیل عدم وجود صورت بسته برای توزیع این پارامتر، شبیه‌سازی از آن به کمک الگوریتم متروپولیس هستینگز با توزیع نامزد مناسب انجام می‌شود. درایدن

[۸] Maximum A Posteriori

و همکاران (۲۰۰۷) پیشنهاد دادند که در هر مرحله از الگوریتم، سطر iام از بین k سطر ماتریس انطباق به تصادف انتخاب و سپس ستون j از بین $\ell + ۱$ ستون ماتریس $\boldsymbol{\Lambda}$ با احتمال‌های $(\psi, \frac{۱-\psi}{\ell}, \ldots, \frac{۱-\psi}{\ell})$ انتخاب و مقدار λ_{ij} برابر یک درنظر گرفته شود. سپس ماتریس انطباق جدید با احتمال

$$\alpha_{\boldsymbol{\Lambda}} = \min\{۱, \frac{\pi(\boldsymbol{\Lambda}^*|\boldsymbol{X}, \boldsymbol{\mu}, \tau)q}{\pi(\boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{\mu}, \tau)q^*}\}$$

پذیرفته شود، که در آن

$$\frac{q}{q^*} = \begin{cases} \frac{\psi}{۱/\ell} & \text{اگر نقطه جورنشده به نقطه جورشده تبدیل شود} \\ \frac{۱/\ell}{\psi} & \text{اگر نقطه جورشده به نقطه جورنشده تبدیل شود} \\ ۱ & \text{اگر نقطه جور شده با نقطه دیگری در } \mu \text{ جور شود} \end{cases}$$

بنابراین بر اساس چنین توزیع پیشنهادی می‌توان نمونه‌های متفاوتی را از چگالی شرطی کامل $\boldsymbol{\Lambda}$ تولید نمود. چون در ساختار این الگوریتم، به‌روزرسانی پارامترها با استفاده از قدم‌های گیبز و متروپولیس هستینگز بوده و هر دو مرحله نیازمند رسیدن به همگرایی مناسب هستند، ممکن است تعداد تکرارهای زیادی برای همگرایی نیاز باشد. به‌علاوه تحلیل پروکراستس جزئی در فضای مماس در هر مرحله از الگوریتم به منظور برآورد پارامترهای مکان و دوران انجام می‌شود. بنابراین به‌طور طبیعی حجم محاسبات در هر مرحله زیاد و سرعت این الگوریتم کم است.

برای افزایش سرعت همگرایی این الگوریتم، کنوبی و درایدن (۲۰۱۲) پیشنهاد دادند که از چهار گام اضافی نزدیکی، دوران، انتقال و تلنگر[۹] (پرش‌های بزرگ) به منظور دستیابی به نقطه شروع مناسب، استفاده شود. الگوریتم انطباق در این حالت مشابه الگوریتم قبلی است با این تفاوت که در تکرارهای اولیه، پس از تکرار مشخصی (مثلاً N_{settle}) به تصادف یکی از پرش‌های بزرگ یک بار تکرار می‌شود. با انجام هر پرش بزرگ ماتریس انطباق با تغییر چندگانه نقاط جور شده به ماتریس جدیدی تبدیل خواهد شد و با تکرار الگوریتم به صورت معمول به تعداد N_{settle} تأثیر ناگهانی این پرش بزرگ به تعادل می‌رسد. این روند تا تکرار مشخصی از الگوریتم (مثلاً N_{init}) ادامه می‌یابد. پس از آن روند طبیعی الگوریتم ادامه پیدا

[۹] Flip

می‌کند تا به همگرایی مطلوب دست یابد. این تکنیک در مطالعات شبیه‌سازی و مثال‌های کاربردی کارایی بهتری نسبت به الگوریتم قبلی دارد. اما همچنان به دلیل وابستگی شدید به نقطه شروع، از کارایی لازم در پروتئین‌ها برخوردار نیست. لذا در بخش‌های بعدی راه‌حل‌های مناسبی برای حل این مشکل پیشنهاد می‌شود.

۲.۲ مدل احتمال در فضای پیکربندی

گرین و ماردیا (۲۰۰۶) یک مدل آماری مشابه در فضای پیکربندی بجای فضای اندازه–شکل، برای مسئله انطباق اشکال بدون برچسب ارائه داده‌اند. در مدل آن‌ها مسئله انطباق به یافتن جواب بهینه در رابطه

$$(\boldsymbol{\mu}^M)^t = \Gamma(\boldsymbol{X}^M)^t + \gamma \times \mathbf{1}_p + \Sigma$$

تبدیل شد، که در آن Σ ماتریس خطای $m \times p$ است و دو عنصر انتقال و دوران نیز به‌عنوان پارامترهای مدل محسوب شده و مانند بقیه پارامترها برآورد می‌شوند. در این روش، مدل‌سازی نقاط جور شده و جور نشده با استفاده از فرآیند پواسن و یک مدل بیزی به‌طور مستقیم بر روی داده‌ها انجام می‌شود. به طور دقیق‌تر، فرض می‌شود که هر دو پیکربندی $\boldsymbol{X}_{k\times 3}$ و $\boldsymbol{\mu}_{\ell\times 3}$، از یک فرآیند پواسن با میانگین متغیر پنهان تبعیت می‌کنند و در نهایت با تشکیل تابع درستنمایی برآورد پارامتر های این مدل به همراه ماتریس انطباق انجام می‌شود. با بهره‌گیری از این روش، کنوبی و درایدن (۲۰۱۲) فرض کردند که نقاط جور شده $\boldsymbol{X}^M$ به‌صورت یک پرشیدگی تصادفی نرمال از میانگین $\boldsymbol{\mu}^M$ باشند و نقاط جور نشده $\boldsymbol{X}^N$ از توزیع یکنواخت تبعیت نمایند. بر اساس نتایجی که آنها در مقاله خود گزارش نمودند، این مدل از سرعت کمتری نسبت به مدلی که در فضای اندازه–شکل تعریف می‌شود، برخوردار است. در این نوع مدل‌بندی نیز مشکلات مربوط به همگرایی و وابستگی به نقطه شروع وجود دارد. از این‌رو در بخش ۳ روش‌های بهبود این الگوریتم‌ها معرفی خواهد شد و در بخش ۴ با استفاده از داده‌های واقعی عملکردشان مورد مقایسه قرار خواهد گرفت.

۳ بهبود مدل‌های انطباق

فرض کنید ماتریس‌های Z و ζ نشان‌های مربوط به پیکربندی $\boldsymbol{X}$ و $\boldsymbol{\mu}$ باشند، به طوری که ζ به‌صورت ثابت در نظر گرفته شود و توزیع‌های $\boldsymbol{X}$ و Z به‌صورت شرطی از یکدیگر مستقل باشند. اگر θ پارامتر(های) توزیع مشخصه Z باشد، می‌توان تابع درستنمایی مدل جدید را به‌صورت

$$L(\boldsymbol{X}, Z|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau, \zeta, \theta) = L(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{\mu}, \tau)L(Z|\zeta, \theta)$$

نوشت. اولین بار درایدن و همکاران (۲۰۰۷) این مدل‌بندی را بر اساس توزیع نرمال برای اضافه کردن اطلاعات بار جزئی و شعاع واندروال به مدل احتمال در مسئله انطباق استروئیدها استفاده نمودند. در عمل استفاده از این اطلاعات سبب بهبود نسبی سرعت همگرایی الگوریتم $MCMC$ در انطباق استروئیدها می‌شود. در مسئله انطباق پروتئین نیز می‌توان با استفاده از همین روش خواص شیمیایی و فیزیکی پروتئین‌ها مانند اندازه، قطبیت، بار جزئی، هیدروفوبیسیته را به مدل‌سازی احتمالاتی اضافه نمود. اما از آنجا که ساختار پروتئین در مقایسه با ساختار استروئیدها بسیار بزرگتر و پیچیده‌تر است، در نظر گرفتن یک توزیع احتمال کلی برای نشان‌های نقاط، بهبودی را در روند الگوریتم ایجاد نمی‌کند.

اکثر الگوریتم‌های قطعی از نتایج انطباق ساختار اول، دوم و خواص شیمیایی و فیزیکی پروتئین‌ها در انطباق ساختار سوم بهره می‌برند. در روش‌های احتمالاتی نیز می‌توان از این اطلاعات به منظور افزایش سرعت و دقت الگوریتم استفاده کرد. بر اساس همین انگیزه، فالیزی و همکاران (۲۰۱۴) و نجیبی و همکاران (۲۰۱۵) با تعمیم مدل‌های گرین و ماردیا (۲۰۰۶) و درایدن و همکاران (۲۰۰۷)، مدل‌بندهای جدیدی را بر اساس انطباق اطلاعات ساختار اول سوم به‌صورت همزمان معرفی کرده‌اند. آن‌ها توزیع پیشین مناسبی را بر مبنای انطباق ساختار اول با در نظر گرفتن گپ‌های بین توالی اسید آمینه‌ها ارائه داده‌اند که در عمل سرعت الگوریتم را در رسیدن به همگرایی مطلوب افزایش می‌دهد. ایده‌ی اصلی این روش‌ها انطباق موضعی نقاط به جای انطباق سراسری آن‌ها بر اساس اطلاعات دیگری به جز اطلاعات ساختار سوم است. به‌عنوان مثال در روش نجیبی و همکاران (۲۰۱۵)،

ابتدا یک جورسازی موضعی بر اساس هرم‌های دلاونی ایجاد شده برای هر مشخصه پروتئین انجام می‌شود و سپس این اطلاعات در قالب توزیع پیشین یا توزیع نامزد در الگوریتم متروپولیس هستینگز، استفاده می‌شود. نتایج شبیه‌سازی استفاده از این روش‌ها در انطباق پروتئین‌ها نشان داده است که سرعت و دقت الگوریتم بهبود می‌یابد. دلیل اصلی این بهبود را می‌توان وجود اطلاعات ساختاری موضعی، میزان مشابهت ساختار اول بر اساس هر یک از ماتریس‌های جانشینی و اطلاعات توالی اسیدهای آمینه در این مدل‌ها دانست.

۴ انطباق احتمالاتی پروتئین‌ها

هدف از این بخش انطباق زوجی شکل سه‌بعدی یک مجموعه از پروتئین‌ها شامل ۵ زوج پروتئین است. این مجموعه به تصادف از بین ۱۶ زوج پروتئینی که اولین بار توسط اریتز و همکاران (۲۰۰۲) مورد استفاده قرار گرفته است، انتخاب شده است. داده‌های مربوط به این پروتئین‌ها از بانک اطلاعاتی پروتئین به آدرس http : //www.pdb.org قابل دسترسی هستند. برای آشنایی با نحوه توالی اسیدهای آمینه و ساختار دوم آنها، نمودار سه بعدی دو پروتئین 1ACX با ۱۰۸ اسید آمینه و 1COB با ۱۵۱ اسید آمینه در شکل ۱ رسم شده است. به منظور مقایسه نتایج این مجموعه دو حالت متفاوت برای مدل احتمالاتی درایدن و همکاران (۲۰۰۷) در نظر گرفته شده است. (الف) از اطلاعات توالی و نوع اسید آمینه در انطباق استفاده نشود (ب) بر اساس نجیبی و همکاران (۲۰۱۵)، اطلاعات نوع اسید آمینه و توالی آنها به کمک مثلث‌سازی دلاونی و میانگین وزنی تابع مشابهت ساختار سوم و نوع اسید آمینه به کمک ماتریس بلوکی جانشینی [۱۰] (BLOSUM62) به مدل اضافه شود. در تمامی حالت‌ها به منظور شروع الگوریتم ۶ نقطه به تصادف انتخاب و بعد از ۱۰۰۰۰۰ تکرار، همگرایی الگوریتم بررسی و در صورت عدم همگرایی، الگوریتم مجدداً اجراء شده است. این روش تا همگرایی لازم تکرار و پس از تایید همگرایی، الگوریتم تا ۴۰۰۰۰۰ تکرار ادامه داده و پارامترهای مدل بر اساس نمونه‌های مستقل

[۱۰] BLOcks SUbstitution Matrix

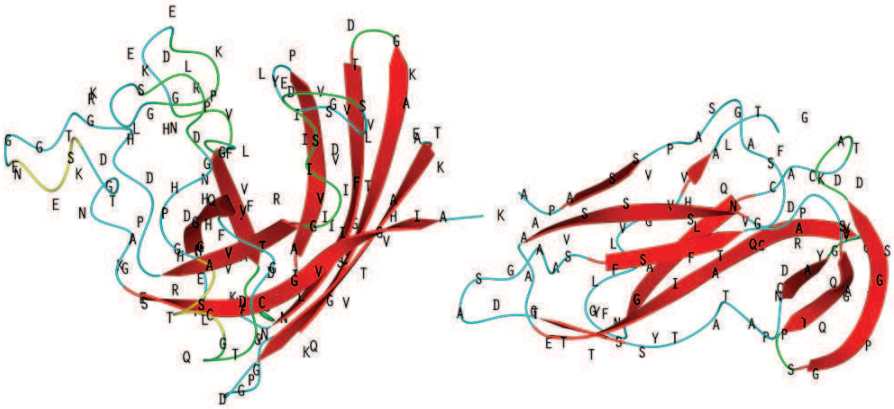
جدول ۱: نتایج مدل انطباق احتمالاتی در فضای اندازه–شکل و روش قطعی CE در ۵ جفت پروتئین به همراه بازه اعتبار پسین ۹۵ درصد پارامترها

پروتئین	k	ℓ	(الف) $\hat{p}$	(الف) $\hat{RMSD}$	(ب) $\hat{p}$	(ب) $\hat{RMSD}$	الگوریتم CE p	الگوریتم CE $RMSD$
1ABA-1DSB	۸۷	۱۸۸	۵۷ (۵۲,۵۹)	۳/۷۸ (۳/۴۵,۴/۲۱)	۵۶ (۵۲,۵۹)	۳/۶۱ (۳/۴۵,۴/۲۱)	۵۴	۴/۵۹
1ACX-1COB	۱۰۸	۱۵۱	۸۹ (۸۳,۹۴)	۳/۹۸ (۳/۸۷,۴/۳۱)	۹۱ (۸۴,۹۵)	۴/۰۹ (۳/۸۴,۴/۲۵)	۹۳	۴/۰۵
1PGB-5TSS	۵۶	۱۹۴	۵۱ (۴۴,۵۵)	۳/۱۶ (۲/۴۱,۳/۶۷)	۵۴ (۴۷,۵۸)	۳/۲۱ (۲/۵۴,۳/۸۰)	۴۹	۳/۱۳
1TNF-1BMV	۱۵۲	۱۸۵	۱۱۰ (۱۰۱,۱۱۵)	۴/۱۳ (۳/۷۳,۴/۲۴)	۱۱۱ (۱۰۲,۱۱۶)	۴/۰۳ (۳/۷۸,۴/۳۱)	۱۱۵	۴/۱۰
2TMV-256B	۱۵۴	۱۰۶	۸۱ (۷۶,۸۵)	۲/۹۸ (۲/۶۵,۳/۴۴)	۷۹ (۷۵,۸۵)	۲/۹۱ (۲/۵۶,۳/۴۴)	۸۵	۳/۵۹

برآورد می‌شود. به منظور مقایسه نتایج الگوریتم‌های احتمالاتی با یکدیگر، برآورد تعداد نقاط جور شده ($\hat{p}$) و جذر میانگین مربع انحرافات[۱۱] ($\hat{RMSD}$) به همراه بازه اعتبار پسین ۹۵ درصد آنها و تعداد نقاط جور شده (p) و $RMSD$ در الگوریتم CE (شیندیالو و بورنه، ۱۹۹۸) در جدول ۱ گزارش شده است.

در اجرای مدل (الف) برای هر یک از ۵ زوج پروتئین، الگوریتم به‌طور میانگین پس از دهمین تکرار (حداقل ۶ و حداکثر ۱۸ تکرار) به همگرایی لازم دست یافت. در حالی که در اجرای مدل (ب) که در آن از اطلاعات اسیدهای آمینه و توالی آنها در مدل‌سازی استفاده شده است، الگوریتم به‌طور میانگین در سومین تکرار (حداقل ۱ و حداکثر ۵ تکرار) به همگرایی لازم دست یافت. تعداد نقاط جور شده در روش احتمالاتی و روش CE در هیچ کدام از پروتئین‌ها به‌طور ۱۰۰ درصد یکسان نبوده است. همان‌طور که ملاحظه می‌شود تعداد نقاط جور شده در روش CE در دو زوج پروتئین‌های 1ABA-1DSB و 2TMV-256B در بازهٔ اعتبار پسین مدل‌های احتمالاتی قرار گرفته است، اما مقدار $RMSD$ هیچ کدام از آن‌ها در بازهٔ اعتبار پسین متناظر

[۱۱] Root Mean Squuare Deviation



شکل ۱: دو پروتئین و توالی کربن‌های آلفا به همراه اطلاعات نوع اسید آمینه و نوع ساختار دوم در فضای سه بعدی الف) پروتئین 1ACX با ۱۰۸ اسید آمینه ب) پروتئین 1C0B با ۱۵۱ اسید آمینه، صفحات مشکلی رنگ نشان دهنده‌ی ساختار دوم صفحات موازی بتا و حروف نمایانگر نوع اسید آمینه است.

قرار نگرفته است. بنابراین می‌توان گفت که در این دو زوج پروتئین با احتمال ۹۵ درصد الگوریتم CE به‌خوبی عمل نکرده است. از آن‌جا که در داده‌های واقعی جواب درست و حقیقی وجود ندارد و مقایسه الگوریتم‌ها به سادگی انجام نمی‌شود، نمی‌توان هیچ کدام از این روش‌ها را به‌طور قطعی پذیرش یا رد کرد. اما با توجه به نتایج فوق، می‌توان نتیجه گرفت که با نادیده گرفتن زمان اجرای روش‌های احتمالاتی، نتایج انطباق آنها منطقی و قابل مقایسه و رقابت با روش‌های قطعی است. از طرفی دیگر، بازهٔ اعتبار پسین و دیگر ابزارهای آماری می‌توانند در مقایسه نتایج الگوریتم‌ها استفاده شوند. در انتها برای ارائه نمایشی از نتایج نهایی انطباق، تصویر پروتئین‌های منطبق شده 1ACX-1COB بر یکدیگر بر اساس ۹۱ نقطه جور شده که از برآورد پارامترها توسط مدل احتمال در فضای اندازه–شکل به همراه اطلاعات اسیدهای آمینه و توالی آنها به‌دست آمده است، در شکل ۲ رسم شده است.

شکل ۲: توالی کربن‌های آلفا پروتئین 1ACX (رنگ مشکی) زمانی که بر پروتئین 1C0B (رنگ خاکستری) با ۹۱ نقطه جور شده است ($\hat{RMSD} = ۴/۰۹$)

بحث و نتیجه‌گیری

در مقاله حاضر روش‌های احتمالاتی مسئله انطباق ساختاری ارائه و با تشکیل تابع درستنمایی، تعیین پیشین‌های مناسب استنباط پسینی انجام شد. با این حال، لزوم تعمیم روش‌های آماری در این حوزه و ارائه راهکارهای مناسب برای افزایش سرعت و دقت الگوریتم‌های پیشنهادی ضروری است. یکی از راه‌حل‌های مقابله با تکرار بسیار زیاد الگوریتم‌ها، استفاده از اطلاعات شیمیایی و فیزیکی پروتئین‌ها مانند خواص هیدروفوبیسیتی، اندازه، قطبیت و ... در انطباق موضعی پروتئین‌ها و در نهایت انطباق سراسری آنها است. علاوه براین، استفاده از نمایش زاویه‌ای بجای نمایش مختصاتی پروتئین و بهره‌گیری از آمار جهت‌دار ممکن است موجب بهبود بهتر روش‌ها شود.

تقدیر و تشکر

نویسندگان مقاله بر خود لازم دانسته از نظرات ارزشمند سردبیر، هیئت تحریریه و سه داور محترم مجله علوم آماری که سبب ارتقاء کیفیت مطالب ارائه شده در این مقاله شدند، تشکر و قدردانی نمایند.

مراجع

Dryden, I., Hirst, J. and Melville, J. (2007), Statistical Analysis of Unlabeled Point Sets: Comparing Molecules in Chemoinformatics, *Biometrics*, **63**, 237-251.

Dryden, I. L. and Mardia K. V. (1998), *Statistical Shape Analysis*, John Wiley, Chichester.

Fallaize, C., Green, P., Mardia, K. and Barber, S. (2014), Bayesian Protein Sequence and Structure Alignment, *arXiv preprint, arXiv:1404.1556*.

Green P. J. and Mardia K. V. (2006), Bayesian Alignment Using Hierarchical Models, with Applications in Protein Bioinformatics, *Biometrika*, **93**, 235-254.

Kendall, D. G., Barden, D., Carne, T. K. and Le, H. (1999), *Shape and Shape Theory*, John Wiley, Chichester.

Kent, J. T. and Mardia, K. V. (2001), Shape, Tangent Projections and Bilateral Symmetry, *Biometrika*, **88**, 469-485.

Kent, J. T., Mardia, K. V. and Taylor C. C. (2004), Matching Problems for Unlabelled Configurations, *Bioinformatics, Images, and Wavelets, Proceedings of LASR 2004*, Ed. R. G. Aykroyd, S. Barber and K. V. Mardia, University of Leeds, Leeds, 33-36.

Kenobi, K. and Dryden, I. L. (2012), Bayesian Matching of Unlabelled Point Sets Using Procrustes and Configuration Models, *Bayesian Analysis*, **7**, 1-20.

Le, H. (1988), *Shape Theory in Flat and Curved Spaces, and Shape Densities with Uniform Generators*, Ph.D thesis, University of Cambridge.

Mardia, K. V., Fallaize, C. J., Barber, S., Jackson, R. M. and Theobald, D. L. (2013), Bayesian Alignment of Similarity Shapes, *The Annals of Applied Statistics*, **7**, 989-1009.

Najibi, S. M., Faghihi, M. R., Golalizadeh, M. and Arab, S. S. (2015), Bayesian Alignment of Proteins via Delaunay Tetrahedralizaton, *Journal of Appllied Statistics*, **42**, 1064-1079.

Ortiz, A. R., Strauss, C. E. M. and Olmea, O. (2002), Mammoth (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison, *Protein Science*, **11**, 2606-2621.

Shindyalov, I. N. and Bourne, P. E. (1998), Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path, *Protein Engineering Journal*, **11**, 739-747.

Shindyalov, I. N. and Bourne, P. E. (2000), An Alternative View of Protein Fold Space, *Proteins: Structure, Function, and Bioinformatics*, **38**, 247-260.

Taylor, C. C., Mardia, K. V. and Kent, J. T. (2003), Matching Unlabelled Configurations Using the EM Algorithm, *Stochastic Geometry, Biological Structure and Images, Proceedings of LASR 2003*, Ed. Aykroyd, R. G., Mardia, K. V. and Langdon, M. J., University of Leeds, 19-21.