پایگاه داده قطعات پپتیدی:
نرم افزار جدید برای کمک به طراحی پروتئین های نوترکیب

فاطمه فضلعلی [*۱]، سید شهریار عرب [۲]

۱- فاطمه فضلعلی fatemeh.fazlali@modares.ac.ir

۲- دکتر سید شهریار عرب sh.arab@modares.ac.ir

چکیده

پروتئین ها یکی از اجزای اصلی سلول ها هستند که معمولا از چندین قطعه تقریبا مجزا تشکیل شده‌اند .این قطعات می توانند در طراحی و ساخت پروتئین های نو ترکیب با عملکرد های خاص استفاده شوند. طراحی پروتئین با ویژگی های مشخص یکی از زیر مجموعه های مهندسی پروتئین است. طراحی پروتئین های نو ترکیب کاربرد های اساسی در پژوهش و صنعت دارد. دستیار طراحی پروتئین یک نرم افزار بیوانفورماتیکی برای طراحی پپتید است. استفاده از روش های بیوانفورماتیک به کاهش هزینه های ساخت و بررسی کامل تمام محیط حل مسئله کمک میکند. برنامه ارائه شده دارای یک رابط کاربری دارد که با اتصال به یک پایگاه داده امکان جستجوی برای یافتن پپتیدی با شرایط مشخص مورد نظر کاربران از میان بیش از چهل میلیون قطعه پروتئینی که از حدود ۱۹۶ هزار ساختار سوم پروتئین تولید شده‌اند را ممکن میسازد. این نرم افزار از طریق آدرس http://bioinf.modares.ac.ir/software/linda برای عموم قابل استفاده است.

کلمات کلیدی: پروتئین های نو ترکیب، مهندسی پروتئین، پایگاه داده قطعات پپتیدی، طراحی پروتئین ، طراحی پپتید، داده های حجیم پروتئینی، بیوانفورماتیک، بیوانفورماتیک ساختاری

۱. مقدمه

زیست شناسی مولکولی در سال های اخیر بروی شناخت بیشتر سلول متمرکز شده است. روزانه حجم وسیع و متنوعی از داده های جدید در اثر بررسی سلول ها تولید و منتشر می شود. پروتئین ها از اصلی ترین اجزای سازنده سلول ها هستند. این درشت مولکول های پیچیده بیشتر عملکرد های اصلی سلول را برعهده دارند. پروتئین ها در ویژگی های ساختاری و عملکردی سلول و در بافت اندام های بدن ها تاثیر گذارند. پروتئین ها مکانیزم های مولکولی ضروری هستند که در تمام بخش های حیات وجود دارند. اجزای سازنده پروتین ها آمینواسید ها هستند. بیست مدل آمینواسید مختلف داریم که با ترکیب کردن آنها می توان زنجیره های پروتئینی متفاوتی تولید کرد.[1]

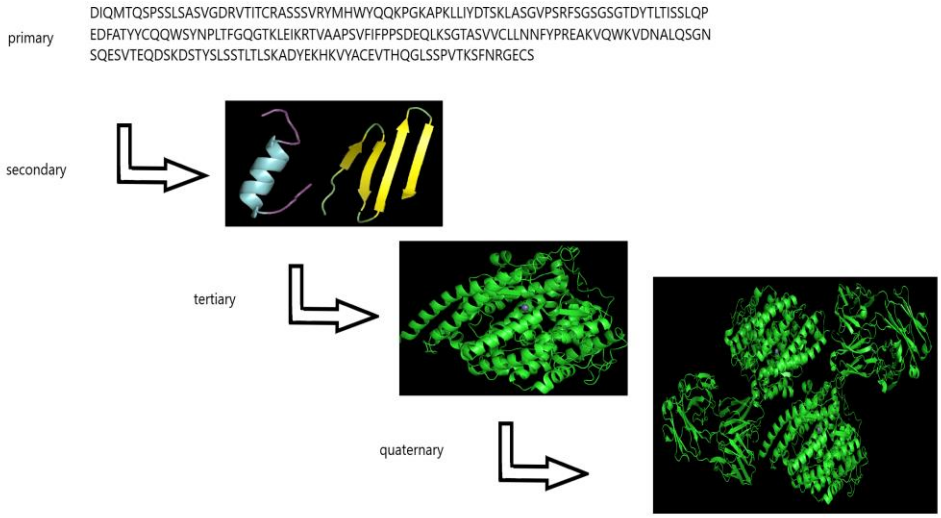
* **Email: fatemeh.fazlali@modares.ac.ir**

عملکرد هر پروتئین به شدت به ویژگی های ساختاری هر پروتئین وابسته است و با ایجاد تغییر در ساختار پروتئین ها می توان عملکرد های مختلفی در آنها ایجاد کرد. به پروتئین های جدیدی که از ترکیب پروتئین های دیگر برای ایجاد یک عملکرد خاص ساخته می شوند پروتئین نو ترکیب گفته می شود. امروزه بازار جهانی پروتئین های نو ترکیب بیش از یک میلیارد دلارارزش گذاری شده است و پیش بینی می شود که تا سال ۲۰۳۰ به بیش از ۵ میلیارد دلار برسد. حجم و میزان رشد این بازار در یک دهه نشان از اهمیت این صنعت دارد.

مهندسی پروتئین دانشی است که به تغییر از پیش برنامه ریزی شده اسیدهای آمینه ها با استفاده از اطلاعات ساختاری و عملکردی پروتئین ها می پردازد. این تغییرات معمولاً بر اساس ساختار سه بعدی شناخته شده یک پروتئین معین و مکانیسم بیوشیمیایی آن است . شاخص ترین مزیت این کار کاهش هزینه های آزمایشگاهی است که با کم کردن فضای جستجو با انجام محاسبات کامپیوتری اتفاق می افتد.[2]

طراحی و ساخت پروتئین ها و آنزیم های جدید با عملکرد های نو بواسطه ایجاد تغییراتی در توالی های آمینواسیدی دارای دو راهکار عمده است . یکی طراحی هدفمند پروتئین ها* و دیگری تکامل جهت دار† در آن ها است. به علاوه برای ایجاد تغییرات در اندازه و ارایش فضایی مولکول ها هم می تواند با روش های دیگر مهندسی پروتین انجام داد.[3]

ساختار پروتئین‡ شکل فضایی است که پروتئین به صورت خود به خودی در محیط های مختلف دارد. [4] پروتئین ها دارای چهار مدل ساختاری هستند که در شکل اول نمایش داده شده‌اند. [5]



شکل ۱ – ساختار های اول و دوم و سوم و چهارم یک پروتئین

مدل ساختاری اول توالی آمینواسیدها است که با پیوند به یکدیگر متصل شده اند و زنجیره پلی پپتید خطی را ایجاد کرده‌اند. ساختار اول ساده ترین و در حال مهم ترین ساختار یک پروتئین است زیرا ساختار فضایی و عملکردی هر پروتئین عموما از روی ساختار اول ان تعیین می‌شود.

* Rational protein design
† Direct evolution
‡ protein structure

مدل ساختاری دوم به چیدمان فضایی آمینواسید ها در توالی های خطی نزدیک به هم می‌پردازد. برخی از این چیدمان ها قاعده مند و تکرار شونده اند. مثل مارپیچ های آلفا و زنجیره های بتا. این الگو ها معمولا با لینکر ها که توالی های کوچک اسید امینه هستند به یک دیگر متصل می شوند .با داشتن یک توالی اسید امینه معمولا می توانیم ساختار دوم این توالی را با دقت خوبی محاسبه کنیم.

مدل ساختار سوم به چیدمان فضایی آمینواسید ها در توالی های خطی دور از هم می‌پردازد. گاها مشخص کردن ساختار دوم و سوم مجزا برای یک پروتئین کوچک ممکن نیست.

مدل ساختار چهارم برای پروتئین هایی است که بیشتر از یک زنجیره پلی پپتیدی دارند. در این پروتئین ها هر زنجیره پلی پپتیدی یک زیر واحد نامیده می‌شود و ساختار چهارم آرایش فضایی و نحوه تماس زیر واحد ها بر هم را نشان می‌دهد. در پروتئین های کوچک که فقط یک زیر واحد دارند ساختار سوم و چهارم یکسان است.

DSSP* عبارتی برای تعریف ساختار دوم پروتئین است. در سال ۱۹۸۳ اولین بار الگوریتمی برای ساختار دوم ارائه شد تا با استفاده از مفاهیم بصری به ساختار دوم برسیم. در واقع این الگوریتم با استفاده از مختصات اتمی سعی در شناخت الگو ها دارد. هر الگو دارای ویژگی های ثابتی است و الگوریتم در کل هشت نوع ساختار دوم برای اسید آمینه ها متصور می شود.[6]

از سال ۱۹۷۱ پایگاه داده پروتئین ها† شروع به آرشیو اطلاعات ساختار سوم پروتئین های موجود کرده‌است و دسترسی به اطلاعات آن رایگان و عمومی است. این پایگاه داده هر هفته توسط مؤسسه بین المللی worldwide Protein Data Bank بروز رسانی می‌شود. [7]

هر رکورد PDB شامل اطلاعات شناسایی مانند عنوان، نام نویسندگان، اطلاعات آزمایش انجام شده (مثل روش و میزان تمایز)، تصویر ثابت و ساختار سه بعدی قابل دست ورزی پروتئینی است.

۲. روش

برای آسان تر کردن فرآیند طراحی پروتئین های نوترکیب یک نرم افزار تحت وب طراحی شده است که به یک پایگاه داده حاوی قطعات پپتیدی متصل است. کاربران برای جستجو در میان قطعات پپتیدی می توان از شش ویژگی استفاده کنند و فضای جستجو را محدود کنند. این شش ویژگی شامل تعداد و توالی آمینو اسید های موجود، توالی ساختار دوم هر قطعه، قطبیت هر توالی و فاصله بین ابتدا تا انتهای قطعه پپتیدی میزان دسترسی پذیری هر آمینواسید ها هستند که در ادامه توضیح داده می شوند. برای محاسبات این ویژگی ها از جدول ۱ استفاده شده است.

تعداد و توالی آمینو اسید های موجود : کاربران می توانند قطعاتی شامل ترکیبات مختلف بیست نوع آمینواسید موجود را برای یافتن لینکر مناسب به رابط گرافیکی بدهند. این عدد می تواند بین سه تا بیست آمینواسید باشد.

* Define Secondary Structure of Protein
† Protein Data Bank archive (PDB)

توالی ساختار دوم هر قطعه : قطعات پپتیدی می توانند هر هشت حالت مختلف در الگوریتم DSSP را برای جستجو در فضای PDB به رابط گرافیکی بدهند.

قطبیت هر توالی : توالی های آمینواسید را می توان از جنبه های مختلفی دسته بندی کرد. یکی از این دسته بندی ها چگونگی میان کنش آن ها با آب است و فهم ای عملکرد برای طراحی پروتئین های جدید ضروری است. به کمک جدول خواص آمینواسید ها (جدول یک) میزان قطبیت هر یک از قطعات پپتیدی از قبل محاسبه شده و در پایگاه داده قرار دارد و کاربران این ویژگی هم برای محدود کردن فضای جستجو می توانند استفاده کنند. هر یک از اینو اسید ها می تواند قطبی* یا غیر قطبی† باشند.

فاصله بین ابتدا تا انتهای قطعه پپتیدی: براساس واحد آنگستروم(Å)‡ می تواند به رابط گرافیکی داده شود تا تمامی قطعاتی با این طول از پایگاه داده پیدا شوند.

میزان دسترسی پذیری§ هر آمینواسید ها : برهم کنش های آبگریز نقش عمده ای در شکل گیری و پایداری پروتئین ها دارند .میزان آبدوست یا آبگریز یا آمینواسید را با میزان در دسترس حلال بودنش می سنجند. با توجه به خواص زنجیره های جانبی و قطبی بودن مولکول آب اکثر امینو اسید ها در آب حل می شوند. میزان دسترسی پذیری از قبل برای آمینواسید ها محاسبه شده و کاربران میتوانند قطعه پپتیدی مورد نظرشان را براساس این که می خواهند چه بخش هایی روی سطح پپتید(در تماس با حلال) و چه بخش هایی در داخل ساختار پپتید (بدون تماس مستقیم با حلال) باشد هم انتخاب کنند.

* Polar
† Nonpolar
‡ Angstrom
§ Accessibility

جدول ۱ : جدول خصوصیات اسید امینه های طبیعی که شامل ستون های بار الکتریکی، قطبیت، پی اچ ایزوالکتریک، شدت قدرت اسید، جرم ملوکولی و نام سه و تک حرفی اسید امینه هاست .

NO	AA_3	AA_1	Charge	Polarity	PI	PKa	PKb	PKx	MW	RW	Max_ACC
1	Ala	A	0	N	6	2.34	9.67	71.08	89.1	89.01	120.56
2	Arg	R	1	P	10.75	2.34	9.04	12.48	174.2	156.19	229.51
3	Asn	N	0	P	5.41	2.17	8.8	0	132.12	114.11	149.85
4	Asp	D	-1	P	2.77	8.8	9.6	3.65	133.11	115.09	157.04
5	Cys	C	0	P	5.07	1.88	10.28	8.18	121.16	103.15	143.79
6	Glu	E	-1	P	3.22	1.96	9.6	4.25	147.13	129.12	188.42
7	Gln	Q	0	P	5.65	2.19	9.13	0	146.15	128.13	186.83
8	Gly	G	0	P	5.97	2.17	9.6	0	75.07	57.05	89.41
9	His	H	1	P	7.59	2.34	9.17	6	155.16	137.14	200.14
10	Ile	I	0	N	6.02	1.82	9.6	0	131.18	113.16	196.42
11	Leu	L	0	N	5.98	2.36	2.36	0	131.18	113.16	206.32
12	Lys	K	1	P	9.74	2.18	2.18	10.53	146.19	128.18	213.74
13	Met	M	0	N	5.47	2.28	2.28	0	14921	131.2	216.63
14	Phe	F	0	N	5.48	1.83	1.83	0	165.19	147.18	227.46
15	Pro	P	0	N	6.3	1.99	1.99	0	115.13	97.12	155.07
16	Ser	S	0	P	5.68	2.21	2.21	0	105.09	87.08	128.27
17	Thr	T	0	P	5.6	2.09	2.09	0	119.12	101.11	138.58
18	Trp	W	0	P	5.89	2.83	2.83	0	204.23	186.22	269.35
19	Tyr	Y	0	P	5.66	2.2	2.2	10.07	181.19	183.18	241.54
20	Val	V	0	N	5.96	2.32	2.32	0	117.15	99.13	169.82

در نهایت کاربر می تواند یکی یا چند تا از فیلد های موجود را پر کند و نرم افزار تمام قطعات پروتئینی که ویژگی های مورد نظر کاربر را دارند را پیدا و به او نمایش دهد. این قطعات می توانند در ادامه برای پیدا شدن قطعه ای که بتوانند پروتئین نو ترکیب مد نظر را بسازد مورد استفاده قرار بگیرد.

در مرحله نخست توسعه نرم افزار ابتدا نسخه ماه جولای بانک اطلاعات پروتئین ها (PDB) که دارای بیش از ۱۹۶۹۷۰ ساختار سوم پروتئین بود دریافت شد. ساختار های سوم دریافتی را فیلتر و با میزان دقت مشخص جداسازی کردیم و تعداد ۷۸۰۰۰ پروتئین را استخراج کردیم.

بسیاری از پروتئین ها به یک دیگر شبیه هستند ولی برای این نرم افزار ما ناچار به انتخاب نمونه های غیر تکراری هستیم چرا که با انتخاب پروتئین های مشابه نتایج نهایی متعلق به گروه های خاصی خواهند بود و وزن دسته ها توزیع داده ها را بر هم خواهد زد. به همین علت پروتئین هایی با تشابه ۹۵ درصد از بانک کنار گذاشته شده و تنها با ۷۸۰۰۰ پروتئین باقی مانده بانک داده پر شده است.

در مرحله ی بعد برای تبدیل فایل های اطلاعات ساختار سوم پروتئین ها (فایل های pdb) به فایل های اطلاعات ساختار دوم (فایل های dssp) یک اسکریپت پایتونی نوشته شد. (در این مرحله از اسکریپت پایتونی الگوریتم dssp استفاده شد .) اطلاعات مورد نیاز شامل نام پروتئین ها، نام زنجیره های پلی پپتیدی و محتوای اسید امینه ها، ساختار دوم، میزان سطح در دسترس در هر پروتئین و موقعیت X,Y,Z مرکز تک تک اسید امینه ها استخراج شد و به صورت ستون هایی در یک فایل متنی قرار گرفت .

در این مرحله فایل خروجی اسکریپت پایتونی قبلی را وارد یک پایگاه داده MYSQL کردیم و به صورت یک جدول بزرگ ذخیره کردیم .سپس قطعات پپتیدی را به طول های ۳ تا ۲۰ اسید آمینه از روی جدول ساختیم. و محتوا را به صورت جدول های جدیدی که شامل ستون های توالی آمینو اسید های موجود، توالی ساختار دوم هر قطعه، قطبیت هر توالی و فاصله بین ابتدا تا انتهای قطعه پپتیدی بود قرار دادیم.

برای اتصال پایگاه داده و وب سرور اسکریپتی نوشته شد تا ارتباط بین درخواست کاربر و سیستم مدیریت پایگاه داده را ایجاد کند.

یک اسکریپت پایتونی هم برای مدیریت ورود محتوای جدید به پایگاه داده با هدف بروز نگه داشتن پایگاه داده با پایگاه اطلاعات پروتئین ها (PDB) ایجاد شد.

به این ترتیب نرم افزار پرو دآ (ProDA)* و رابط کاربری اش با هدف یافتن قطعات پپتیدی مناسب از میان بیش از 40 میلیون قطعه پپتیدی موجود در پایگاه داده ایجاد و توسعه پیدا کرد .

۳. نتیجه‌گیری

امروزه پروتئین ها در صنایع مختلفی از جمله بیوتکنولوژی، دارویی، بهداشتی، آرایشی، غذایی، شیمیایی و کشاورزی کاربرد دارند و تولید پروتئین های نو ترکیبی که عملکرد های جدیدی و مشخصی داشته باشند یکی از نیاز های اصلی پروسه ساخت پروتئین در این صنایع است. با در نظر داشتن این موضوع نرم افزار پرو دآ و ابزار های دیگر بیوانفورماتیکی با هدف تسهیل ساخت قطعات پروتئینی طراحی شدند تا بتواند علاوه بر کاهش هزینه های آزمایشگاهی به یافتن مجموعه ی کاملی از راه حل کمک کند.

این نرم افزار پروتئین هایی را از بانک داده پروتئین ها استخراج می کند و با پردازش آنها را به قطعات پپتیدی به طول سه تا بیست آمینواسید تبدیل می کنند و به کاربران اجازه می دهد تا با تنظیم پارامتر های ورودی رابط کاربری بین چهل میلیون قطعه پپتیدی به وجود آمده قطعه پروتئینی که شرایط مد نظرشان را دارد را پیدا کنند.

۴. مراجع

* ProDA (Protein Design Assistant)

[1] J. E. Murray, N. Laurieri, and R. Delgoda, *Proteins*. Elsevier Inc., 2017.
[2] P. J. Carter, "Introduction to current and future protein therapeutics: A protein engineering perspective," *Exp. Cell Res.*, vol. 317, no. 9, pp. 1261–1269, 2011, doi: 10.1016/j.yexcr.2011.02.013.
[3] R. A. Chica, "Protein Engineering, Design and Selection," *Protein Eng. Des. Sel.*, vol. 33, p. 2020, 2020, doi: 10.1093/protein/gzaa024.
[4] R. Zaman *et al.*, "Current strategies in extending half-lives of therapeutic proteins," *J. Control. Release*, vol. 301, no. March, pp. 176–189, 2019, doi: 10.1016/j.jconrel.2019.02.016.
[5] C. Nelson, *Lehninger's principles of biochemistry*, vol. 53, no. 9. 2013.
[6] Y. Zhang and C. Sagui, "Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI," *J. Mol. Graph. Model.*, vol. 55, pp. 72–84, 2015, doi: 10.1016/j.jmgm.2014.10.005.
[7] I. M. L. Saur, R. Panstruga, and P. Schulze-Lefert, "NOD-like receptor-mediated plant immunity: from structure to cell death," *Nat. Rev. Immunol.*, vol. 21, no. 5, pp. 305–318, 2021, doi: 10.1038/s41577-020-00473-z.