مجله علمی ـ پژوهشی زیست‌فناوری گیاهان زراعی
سال دوازدهم، شماره یکم، پیاپی سی‌ونهم، پائیز ۱۴۰۱ (۶۷-۵۵)

Crop Biotech.
Autumn (2022) 12(1): 55-67 DOI: 10.30473/cb.2023.66084.1892

«مقاله پژوهشی»

شناسایی خصوصیات احتمالی یک پروتئین ناشناخته از آرابیدوپسیس با استفاده از روش‌های بیوانفورماتیکی

مهین پوراسمعیل[۱]، مقصود پژوهنده[۲]*

۱. دکتری، گروه بیوتکنولوژی کشاورزی، دانشکده کشاورزی، دانشگاه شهید مدنی آذربایجان، کیلومتر ۳۵ جاده تبریز- آذرشهر، ایران.
۲. دانشیار، گروه بیوتکنولوژی کشاورزی، دانشکده کشاورزی، دانشگاه شهید مدنی آذربایجان، کیلومتر ۳۵ جاده تبریز- آذرشهر، ایران.

(تاریخ دریافت: ۱۴۰۱/۰۶/۲۱ - تاریخ پذیرش: ۱۴۰۱/۰۹/۲۹)

Identification of possible properties of an unknown protein from Arabidopsis using bioinformatics methods

Mahin Pouresmaei[1], Maghsoud Pazhouhandeh[2*]
1. Ph.D., Biotechnology Department, Agriculture Faculty, Azarbaijan Shahid Madani University, Km 35 Tabriz-Azarshahr Road, Tabriz, Iran.
2. Associate Professor, Biotechnology Department, Agriculture Faculty, Azarbaijan Shahid Madani University, Km 35 Tabriz-Azarshahr Road, Tabriz, Iran.
(Received: Sep. 12, 2022 - Accepted: Dec. 20, 2022)

چکیده

امروزه توالی ژنوم اکثر موجودات شناسایی شده است و این اطلاعات در شناخت عملکرد و ویژگی‌های موجود مفید هستند. در این میان اطلاعات پردازش نشده‌ای وجود دارد که با پیشرفت تکنولوژی و استفاده از ابزارهای بیوانفورماتیک می‌توان از آن‌ها برای مطالعه پروتئین‌ها و ژن‌های ناشناخته استفاده کرد. در این تحقیق توالی یک ژن با عملکرد ناشناخته از گیاه *Arabidopsis thaliana* با شماره دسترسی X91953.1 در پایگاه NCBI برای بررسی و مطالعه ساختار و عملکرد احتمالی مورد استفاده قرار گرفت. این ژن مربوط به کروموزوم شماره یک در آرابیدوپسیس تالیانا بوده و با ٦٧٦ جفت باز، پروتئینی با ۱۵۰ اسیدآمینه به وزن مولکولی تقریبا ۱۵ کیلودالتون تولید می‌کند. با استفاده از سرورهای بیوانفورماتیک خصوصیات توالی ژن و پروتئین مربوطه بررسی و مشخص شد که دارای ۱۸ نوع موتیف تنظیمی است که وظایف برخی از آن‌ها شناخته شده است که می‌توان به پاسخ به نور و فعالیت عناصر Cis برای بیان در مریستم اشاره کرد. آنالیزها نشان داد این پروتئین دارای ۳۸ موتیف است که سه مورد در آن حفاظت شده با تکرار بالاست. این پروتئین دارای سیگنال پپتید در انتهای آمینی خود بوده و به فضای خارج سلولی تراوش می‌شود. بنابراین احتمال حضور آن در فضای بین سلولی بیشتر از هسته و اندامک‌های درون سلولی است. محل تنظیم یک microRNA نیز روی رونوشت این ژن وجود دارد و این microRNA در پاسخ به شوری و همچنین در جنین فعالیت دارد. این پروتئین ناشناخته با پروتئین دیگری در آرابیدوپسیس با شماره دسترسی UPF0540 At1g62000 دارای حدود ۹۰ درصد همولوژی است که برای بررسی‌های بیشتر برای شناسایی نقش پروتئین مورد نظر می‌تواند مورد استفاده قرار گیرد. این پروتئین در ۱۰ بافت مختلف عمدتا در جنین و آندوسپرم دانه بیان می‌شود. براساس همه آنالیزهای صورت گرفته دو وظیفه تمایز پوشش بذر و فرایند بیوسنتز مواد مترشحه در اثر نور را برای این پروتئین می‌توان پیش بینی کرد که در ادامه کار روش‌های آزمایشگاهی برای تست عملکرد منسوب شده به آن توصیه می‌شود.

واژه‌های کلیدی: آرابیدوپسیس تالیانا، بیوانفورماتیک، پایگاه اطلاعات پروتئین، پیش‌بینی عملکرد پروتئین.

Abstract

Today, the genome sequence of most organisms has been identified, and this information is useful in understanding the function and characteristics of organisms. In the meantime, there is unprocessed information that can be used to study unknown proteins and genes with the advancement of technology and the use of bioinformatics tools. In this research, the sequence of a gene with unknown function from *Arabidopsis thaliana* with accession number of X91953.1 in NCBI database was used to investigate and study its structure and possible function. This gene is related to chromosome number one in *Arabidopsis thaliana* and with 676 base pairs, it produces a protein with 150 amino acids and a molecular weight of approximately 15 kD. By using bioinformatics servers, the characteristics of both gene and protein sequences were investigated and it was found that it has 18 types of regulatory motifs, the functions of some of which are known, which can be related to the response to light and the activity of *Cis* elements for expression in the meristem. The analyzes showed that this protein has 38 motifs, three of them are conserved with high frequency. This protein has a signal peptide at its Nt and is leaked into the extracellular space. Therefore, its presence in the intercellular space is more likely than the nucleus and intracellular organelles. There is also a regulation site of a microRNA on its transcript and this microRNA is active in response to salinity and also in the embryo. This unknown protein has about 90% homology with another protein in Arabidopsis with accession number of UPF0540 (At1g62000), which can be used for further studies to identify the role of the desired protein. This protein is expressed in 10 different tissues, mainly in embryo and seed endosperm. Based on all the analyzes carried out, two functions of seed coat differentiation and the biosynthesis of secreted substances due to light can be predicted for this protein. In the continuation of this work, laboratory methods are recommended for testing the functions attributed to this gene.

Keywords: *Arabidopsis thaliana*, Bioinformatics, Predict protein function, Protein database.

* نویسنده مسئول: مقصود پژوهنده
E-mail: pazhouhandeh@gmail.com

مقدمه

آرابیدوپسیس تالیانا (*Arabidopsis thaliana*) یک گیاه مدل متعلق به خانواده براسیکاسه بوده و نخستین گیاهی است که توالی ژنوم آن کاملا شناسایی شده است و به عنوان یک گیاه مدل در تحقیقات استفاده می‌شود. دلیل استفاده از این گیاه به عنوان مدل تحقیقاتی بیشتر به خاطر خصوصیات رشدی آن از جمله ارتفاع کوتاه، تولید بذر زیاد و کوتاه بودن چرخه زندگی است. محتوای ژنی آرابیدوپسیس تالیانا ۱۲۵ میلیون جفت باز و دارای ۲۵۴۹۸ ژن بوده و ۱۱۰۰۰ پروتئین در خانواده‌های مختلف را تشکیل می‌دهد (Swarbreck *et al*., 2007). تعداد زیادی از پروتئین‌ها در موجودات مختلف وجود دارد که همه یا قسمتی از خصوصیات و عملکرد آن‌ها هنوز ناشناخته باقی مانده است. این داده‌های پروتئینی شامل دو دسته پروتئین‌های جدا شده از موجودات مختلف و پروتئین‌هایی است که به صورت فرضی و از روی توالی DNA موجودات شناسایی شده‌اند (Hawkins, 2007). برای بررسی و تعیین خصوصیات یک پروتئین روش‌های مختلفی وجود دارد: ۱- تعیین ساختار پروتئین با استفاده از X-ray و Nuclear Magnetic Resonance (NMR)، ۲- بیان پروتئین و خالص‌سازی و تعیین توالی، ۳- مطالعه از طریق برهمکنش پروتئین-پروتئین و پروتئین-DNA، ۴- استفاده از روش‌های بیوانفورماتیک. در این میان استفاده از روش‌های بیوانفورماتیک در مطالعه پروتئین‌ها یک روش مؤثر و کم هزینه است. ابزارهای بیوانفورماتیکی موجود به صورت پایگاه داده‌ها و ابزارهای آنلاین و همچنین ابزارهای آفلاین برای مطالعه ساختار و عملکرد پروتئین‌ها مورد استفاده قرار می‌گیرد (Bollag *et al*., 1996). همچنین برای پیش‌بینی نقش یک پروتئین شش روش توصیه شده: ۱- براساس توالی، ۲- براساس ساختار، ۳- برهمکنش، ۴- براساس آزمایشات پروتئومیکس، ۵- براساس کارکرد ۶- ترکیب روش‌های ذکر شده،که بیشترین روش استفاده مربوط به استفاده از مقایسه توالی‌ها است. سرورهای مختلفی برای پیش‌بینی عملکرد و ساختار یک پروتئین براساس توالی استفاده می‌شود (جدول ۱) (Edwards & Cottage, 2003; Hawkins, 2007). پروتئین‌های با عملکرد ناشناخته بخش قابل توجهی از ژنوم‌های توالی‌یابی شده را تشکیل می‌دهند. تعریف نقش این پروتئین‌ها برای درک فرآیندهای سلولی ضروری است (Shumilin *et al*., 2012). در سال ۲۰۰۶ در توالی ژنوم گیاه Arabidopsis از ابزارهای بیوانفورماتیک برای شناسایی ساختارهای ثانویه، شناسایی دمین و شناسایی اطلاعات بیانی استفاده شد و مشخص گردید که آنالیزهای بیوانفورماتیک منبع ارزشمندی برای ژنومیک عملکردی هستند. انتظار می‌رود که داده‌ها، قوانین و پیش‌بینی‌های انجام شده با استفاده از ابزارهای بیوانفورماتیکی در توالی ژنوم Arabidopsis الهام بخش پیشرفت در زمینه مطالعات آتی باشد (Clare *et al*., 2006). مطالعات زیادی به بررسی نقش احتمالی یک ژن یا پروتئین آن با روش بیوانفورماتیکی پرداخته‌اند. به طور مثال، تحقیقی روی یک پروتئین ناشناخته در *E. coli* با کمک بررسی موتیف‌ها، ساختار 3D و تغییرات پس از ترجمه، وظایف احتمالی آن را مشخص و احتمال تغییرات پس از ترجمه آن را مشخص نمود (Pino *et al*., 2019). در یک مطالعه دیگر برای بررسی عملکرد یک پروتئین ناشناخته در گیاه آرابیدوپسیس از آنالیزهای بیوانفورماتیکی استفاده شد (Azizi-Dargahlou & Fazeli-Nasab, 2022). در مطالعه‌ای دیگر برای بررسی وظایف یک پروتئین در گیاه *Brachypodium distachyon* از ابزارهای بیوانفورماتیکی برای مطالعه موتیف‌ها و دمین‌ها استفاده گردید (Hajibarat *et al*., 2018). مطالعه دیگری نشان داد که یافتن یک پروتئین با وظیفه مشخص که شریک مولکولی یک پروتئین ناشناس در شبکه تعامل پروتئین‌ها تشخیص داده می‌شود برای کشف وظیفه آن اهمیت زیادی دارد (Kuroda *et al*., 2006). می‌توان گفت که با توسعه علم بیوانفورماتیک، استفاده از این علم برای شناخت

خصوصیات پروتئین نسبت به سایر روش‌های شناسایی پروتئین مقرون‌به‌صرفه‌تر است. پروتئین‌های گیاهی دارای نقش آنزیمی، ساختاری و عملکردی مختلفی (فتوسنتز، بیوسنتز، انتقال، ایمنی و ...) هستند. این پروتئین‌ها همچنین به عنوان محیط ذخیره‌سازی برای برآوردن نیازهای رشد و تغذیه جنین عمل می‌کنند. در سلول‌های گیاهی، دستگاه گلژی اندامک کلیدی برای سنتز پلی ساکارید و گلیکولیپید، گلیکوزیلاسیون پروتئین و مرتب سازی پروتئین به سمت بخش‌های مختلف سلولی است (Rasheed *et al*., 2020).

در پروتئین‌ها، موتیف یک توالی حفاظت شده کوتاه است که عملکرد مشخصی دارد و در پروتئین‌های مختلف تکرار شده است. یک موتیف معمولی، مانند موتیف انگشت روی (Zn)، ده تا بیست اسیدآمینه طول دارد. دمین همچنین یک الگوی توالی حفظ شده است که به‌عنوان یک واحد عملکردی و ساختاری مستقل تعریف می‌شود (Wang *et al*., 2021). دمین‌ها معمولا طولانی تر از موتیف‌ها هستند. موتیف‌ها و دمین‌ها از نظر تکاملی بیشتر از سایر نواحی پروتئین حفاظت شده هستند. بنابراین شناسایی موتیف‌ها و دمین‌ها در پروتئین‌ها از نظر طبقه‌بندی توالی‌های پروتئینی و تعیین عملکرد آن بسیار مهم است (Xiong, 2006). پیش‌بینی ساختار پروتئین با استفاده از بیوانفورماتیک می‌تواند شامل جستجوهای تشابه توالی، زیرهم‌چینی‌های چندگانه توالی‌ها، شناسایی و مشخص‌سازی دمین‌ها و موتیف‌ها، پیش‌بینی ساختار ثانویه و ساخت مدل‌های سه‌بعدی برای پروتئین باشد (Edwards & Cottage, 2003).

در این مطالعه برای بهره‌گیری از ابزارهای بیوانفورماتیکی در جهت کشف نکات مهم بیولوژیکی، از گیاه مدل آرابیدوپسیس تالیانا که اطلاعات نسبتا زیاد و کاملی نسبت به سایر گیاهان در دسترس هست یک توالی ژن کاندید ناشناخته (mRNA) از پایگاه داده NCBI به شماره دسترسی X91953.1 انتخاب شد که توالی آن موجود بود اما هیچ وظیفه و نقشی برای آن مورد تحقیق و بررسی قرار نگرفته بود. توالی ژن و پروتئین آن با استفاده از نرم‌افزارها و برنامه‌های بیوانفورماتیکی مورد بررسی و مطالعه قرار گرفت تا حداقل بصورت تئوری اطلاعاتی در زمینه نقش و وظیفه آن بدست آید. بدیهی است که پس از یافتن نقش احتمالی برای این ژن، بسیار هدف‌دارتر و منسجم‌تر می‌توان با روشهای بیولوژیکی و ژنتیکی به اثبات آن پرداخت.

جدول ۱. اسامی سرورهای مورد استفاده برای پیش بینی عملکرد پروتئین براساس توالی

Table 1. Web resources for prediction of protein function by sequence

Website	Description	URL
GoFigure	GO annotation by homology search	http://udgenome.ags.udel.edu/frm go.html
FASTA	Sequence homology search	http://www.ebi.ac.uk/fasta33/
BLAST suite	Sequence homology search	http://www.ncbi.nih.gov/BLAST/
Pfam	Protein family alignment database	http://www.sanger.ac.uk/Software/Pfam/
PROSITE	Functional motif database	http://us.expasy.org/prosite/
Blocks	Database of conserved regions of proteins	http://blocks.fhcrc.org/
SMART	Domain-based annotation resource	http://smart.embl-heidelberg.de/
PRINTS	Protein fingerprint database	http://bioinf.man.ac.uk/dbbrowser/PRINTS/
ProtFun	GO functional category prediction	http://www.cbs.dtu.dk/services/ProtFun/
PFP	GO function prediction server	http://dragon.bio.purdue.edu/pfp
Gotcha	GO function prediction software	http://www.compbio.dundee.ac.uk/Software/Gotcha/gotcha.html
KEGG Genes	Genome sequence collection	http://www.genome.jp/kegg/genes.html
COG	Clusters of orthologous groups of proteins	http://www.ncbi.nlm.nih.gov/COG/

مواد و روش‌ها

برای کشف نقش احتمالی و آنالیز خصوصیات ژن و پروتئین ناشناخته از گیاه آرابیدوپسیس (Gasteiger *et al*., 2005) توالی ژن (mRNA) مربوط به آن با شماره دسترسی X91953.1 با اندازه ۶۷۶ جفت باز از پایگاه NCBI استخراج شد. برای بررسی ویژگی‌های آن ابتدا توالی موردنظر در NCBI بلاست چندگانه (نوکلئوتید-نوکلئوتید، نوکلئوتید-پروتئین و پروتئین- پروتئین) شد و از نتایج بلاست برای رسم درخت فیلوژنتیک از نرم‌افزار MEGA6 استفاده گردید که رسم درخت فیلوژنیک با روش N.J (Neighbor-Joining) و بوت استراپ ۱۰۰۰ صورت گرفت.

برای بررسی ویژگی و عملکرد این پروتئین از سرورهای مختلفی جهت پیدا کردن وزن مولکولی، نقطه‌ی ایزوالکترونیک، ساختار سه بعدی، موتیف، دمین، عناصر تنظیمی، محل بیان استفاده گردید (Carthew & Sontheimer, 2009) که در این سرورها توالی ژن یا پروتئین و یا شماره دسترسی ژن مورد نظر وارد و به بررسی ویژگی و عملکرد پروتئین پرداخته شد (جدول ۲). برای شروع آنالیز توالی پروتئین مربوط به X91953.1 از سایت UniProt به فرمت fasta ذخیره گردید. برای بررسی و تجزیه تحلیل شباهت این پروتئین با سایر پروتئین‌ها و همچنین بررسی وظیفه و محل بیان آن از سایت UniProt استفاده شد.

شناسایی عناصر تنظیمی می‌تواند به پیش بینی عملکرد ژن‌ها کمک کند. بنابراین شناسایی ناحیه پروموتری توالی و عناصر تنظیمی ژن X91953.1 با استفاده از برنامه plantCARE انجام شد. در توالی یک پروتئین برای شناسایی موتیف‌های حفاظت شده از ارزیابی سایت MEME استفاده شد. این سرور موتیف‌های محافظت‌شده و نیز تکرار موتیف‌ها، ماکزیمم تعداد موتیف و محدوده موتیف‌ها را پیش‌بینی می‌کند (Hu *et al*., 2015). دمینِ پروتئین، بخش حفاظت شده یک توالی است که مستقل از بقیه زنجیره پروتئین تکامل یافته و عمل می‌کند. برای شناسایی و آنالیز دمین پروتئین از سایت SMART استفاده شد (Hajibarat *et al*., 2018). در ادامه کار و برای شناسایی موتیف‌های اختصاصی در توالی پروتئین مورد نظر از سایت MOTIF Search براساس کتابخانه‌های موتیف PROSITE، NCBI-CDD، Pfam بررسی صورت گرفت (Finn *et al*., 2013; Marchler-Bauer *et al*., 2012; Sigrist *et al*., 2012).

برای بررسی ویژگی‌های فیزیکوشیمیایی پروتئین مورد نظر و محاسبه وزن مولکولی و نقطه ایزوالکتریک تئوریتیکال توالی مورد نظر در سایت ExPASY (Gasteiger *et al*., 2005) ابزار ProtParam مورد استفاده قرار گرفت. همچنین از ابزار SWISS-MODEL سایت ExPASY نیز برای بررسی ساختار سه‌بعدی (Bienert *et al*., 2017) و از ابزار TargetP برای تعیین عملکرد پروتئین استفاده شد (Emanuelsson *et al*., 2007). همچنین برای شناخت بهتر وظیفه این برهمکنش پروتئین مورد نظر با سایر پروتئین‌ها در گیاه آرابیدوپسیس توسط وب سایت STRING رسم گردید.

miRNA از RNAهای کوچک یا از ژنهای Mir نشات می‌گیرند و در تنظیم بیان سایر ژن‌ها نقش دارند. این گروه از RNAهای غیر کد کننده در حدود ۱۸-۲۵ نوکلئوتید را شامل شده و بیان ژن را پس از رونویسی با تجزیه mRNA یا مهار ترجمه آن‌ها کنترل می‌کند (Carthew & Sontheimer, 2009). مسیر miRNA یک مکانیسم ضروری در تنظیم فرایندهای زیستی مختلف بوده و در گیاهان در فرایندهای اساسی مثل تمایز، پاسخ به تنش‌های زیستی و غیر زیستی و پاسخ به هورمون‌ها نقش دارد (Borges *et al*., 2011). برای یافتن احتمال وجود یک توالی هدف miRNA در روی ترانسکریپت‌های این ژن از سایت mirbase استفاده شد.

جدول ۲. اسامی و آدرس پایگاه داده‌های استفاده شده در این پژوهش

Table 2. Name and URL address of the databases exploited in this research

Database	URL
Expasy	http://web.expasy.org/compute_pi/
GSDS	http://gsds.cbi.pku.edu.cn/
MEME	http://meme-suite.org/
SMART	http://smart.embl-heidelberg.de/
PlantCare	http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
UniProt	https://www.uniprot.org/
MOTIF Search	https://www.genome.jp/tools/motif/
CELLO	http://cello.life.nctu.edu.tw/
AntiBP Server	https://webs.iiitd.edu.in/raghava/antibp/submit.html
STRING	https://string-db.org
AodPred	http://www.mirbase.org
Mirbase	http://www.mirbase.org

نتایج و بحث

همولوژی و تاکسونومی

جهت شناسایی وظیفه و نقش یک پروتئین ناشناخته آرابیدوپسیس، توالی ژن مربوط به آن (mRNA) با شماره دسترسی X91953.1 برداشت شده و توالی نوکلئوتیدی و همچنین پروتئینی آن در سایت NCBI بصورت چهارگانه بلاست (BLAST) شد و نتایج همه بلاستها مطابق هم بوده و مشخص گردید توالی مورد نظر با توالی‌های از ژنوم آرابیدوپسیس با شماره دسترسی CP002684.1 و NM-104889.2 و AY058148.1 و AC000375.1 دارای شباهت ۹۸/۹۶ درصد و توالی دیگری مربوط به قسمتی از ژنوم آرابیدوپسیس با شماره دسترسی BX81769.1 دارای شباهت ۹۸/۹۳ درصد است.

برای پی بردن به موقعیت تاکسونومیک این پروتئین از آنالیز سایت UniProt استفاده شد و مشخص گردید که موقعیت تاکسونومیک پروتئین مورد نظر به قرار زیر است و در موجود دیگری این توالی گزارش نشده است:

Eukaryota› Viridiplantae› Streptophyta› Embryophyta› Tracheophyta› Spermatophyta› Magnoliophyta› eudicotyledons› Gunneridae› Pentapetalae› rosids› malvids› Brassicales› Brassicaceae› Camelineae› Arabidopsis.

نتایج آنالیز توالی مورد نظر در سایت ExPASY (Gasteiger *et al*., 2005) ابزار ProtParam مورد نشان داد که این پروتئین دارای ۱۵۰ اسید آمینه بوده و وزن مولکولی آن ۱۵۱۲۷/۱۹ دالتون و نقطه ایزوالکترونیک آن ۱۰/۶۵ است. توالی نوکلئوتیدی این ژن و موقعیت کدون آغازین و پایان آن و همچنین توالی اسیدآمینه آن عبارتند از:

ATGgaatgccacaaagtttcttgtgcttctcgtgatt
ggcgttttgtgtgccattgtcaccgcaaggcaggtca
aagatttgtccacagagaccaaattaggggcctctct
tccaaaaacgactaccaaaggcattggagctcagctt
tctgcaactggctcaactttcagcagcagcagtgtcg
ttagctatcttaatggttttaacaatcccaaaggccc
aggggccaatgcattcgaaagtggctccacatttacc
agcggacaagtcactgccaagggtcgcaacgcaagag
tttcttctgcaagtgcttctaccgctacaggtgaggc
tgcagctgcagtgactcgcaaagctgctgctgcacgt
gcaaagggtaaggtagcttccgcatcaagggtgaagg
gttcctctgagaagaagaagaaggaccgcaaaggaaa
aaatgatTGAgcgtgaggtcatctcatattctcacac
atgcgtattcctcaaacctacatattaattatatcct
aaaaactacaaactcacaatctctttcttgttcataa
ataaaaccagagattgtcacctctaaatattattatt
atgaaatgctactctttctcatgtaatggctttttta
aaataaaataaatgcaataaatatgatcttatagt

MNATKFLVLLVIGVLCAIVTARQVKDLSTETKLG
ASLPKTTTKGIGAQLSATGSTFSSSSVVSYLNGFNNP
KGPGANAFESGSTFTSGQVTAKGRNARVSSASASTAT
GEAAAAVTRKAAAARAKGKVASASRVKGSSEKKKKDR
KGKND*

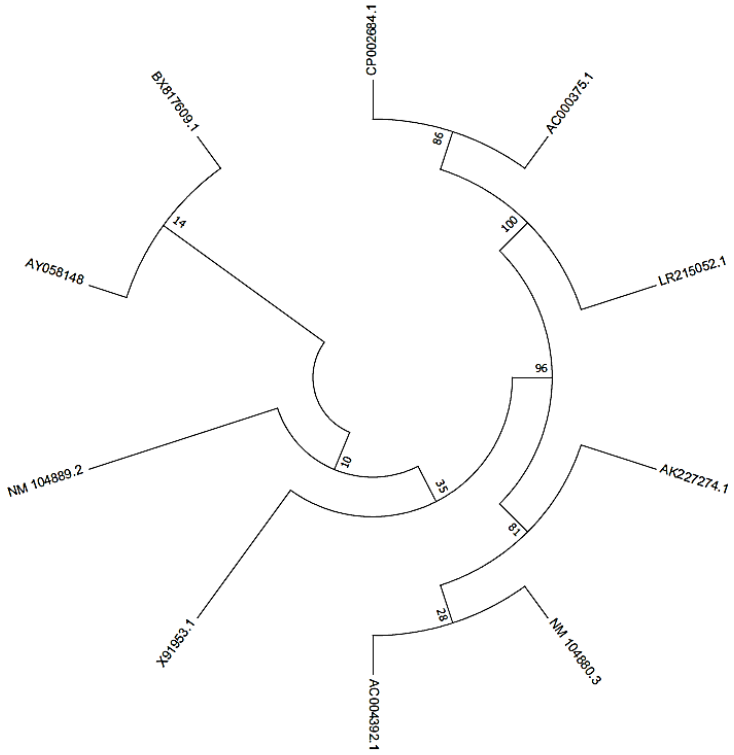
رسم درخت فیلوژنتیک

درخت فیلوژنتیک (فیلوژنی یا درخت تکاملی) یک نمودار انشعاب یا درختی است که روابط تکاملی

بین گونه‌های مختلف بیولوژیکی یا موجودات دیگر را بر اساس شباهت‌ها و تفاوت‌های ویژگی‌های فیزیکی یا ژنتیکی آنها نشان می‌دهد (Baum, 2008). رابطه فیلوژنتیک پروتئین مورد نظر با سایر پروتئین‌های موجود در موجودات، براساس نتایج بلاست در سایت NCBI و استفاده از توالی‌هایی که دارای بیشترین میزان همولوژی (بالای ۹۰٪) و بیشترین میزان هم‌پوشانی (بالای ۹۸ ٪) با پروتئین موردنظر بودند بررسی شد. شماره‌های دسترسی توالی‌های که در آن‌ها میزان همولوژی و هم‌پوشانی به طور همزمان بالا بود. آنالیز توالی‌ها با CLUSTAL W انجام گرفت و در نهایت رسم درخت فیلوژنیک با روش Neighbor-Joining (N.J) و بوت استراپ ۱۰۰۰ در نرم‌افزار MEGA6 انجام گرفت (Allahi *et al*., 2017) (شکل ۱).

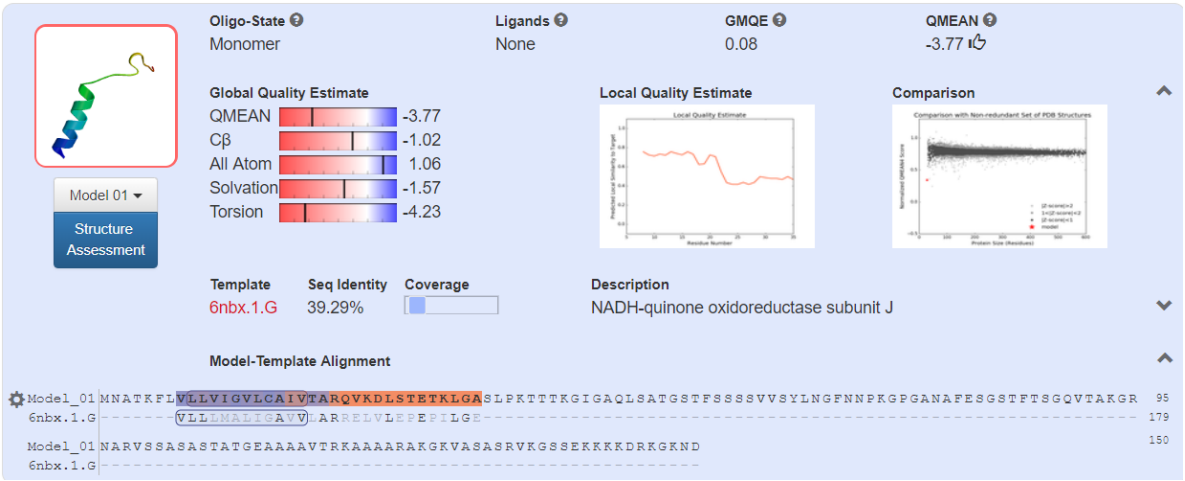
بررسی ساختار سه بعدی و محل جایگیری پروتئین

برای درک بهتر عملکرد پروتئین، تعامل آن با اجزای دیگر (لیگاندها، پروتئین، DNA و ...) و فهم اثرات فنوتیپیکی جهش‌ها، ساختار سه بعدی پروتئین می‌تواند مفید باشد. بنابراین ساختار سه بعدی این پروتئین با استفاده از ابزار SWISS-MODEL سایت ExPASY به‌دست آمد (شکل ۲). پروتئین‌ها مولکول‌های بزرگ و پیچیده‌ای هستند که از واحدهایی به نام اسیدهای آمینه تشکیل شده‌اند که با اتصال به یکدیگر، زنجیره‌های بلندی را تشکیل می‌دهند.



شکل ۱- رابطه فیلوژنتیک X91953.1 با سایر توالی‌های گیاه آرابیدوپسیس. با استفاده از نتایج سایت NCBI و توالی‌های دارای بیشترین همولوژی هم ردیف سازی ۱۰ نمونه اول برای رسم درخت فیلوژنتیک با استفاده از نرم‌افزار MEGA.6 براساس روش (N.J) neighbor joining و آزمون بوت استراپ استفاده شد.

Figure 1. The Phylogenetic Relationship of X91953.1 Protein with Other Proteins. Using the results of NCBI site and sequences with the most homology, the alignment of the first 10 samples was used to plot the phylogenetic tree using the MEGA.6 software based on the neighbor joining method (N.J) and bootstrap test.



شکل ۲. پیش‌بینی ساختار سه بعدی پروتئین: ساختار سه بعدی این پروتئین با استفاده از ابزار SWISS-MODEL سایت ExPASY به‌دست آمد.

Figure 2. Prediction of 3D structure of the protein: The three-dimensional structure of this protein was obtained using the EXISSY SWISS-MODEL tool.

در طبیعت ۲۰ نوع مختلف اسیدآمینه وجود دارد که به صورت‌های مختلفی با هم ترکیب می‌شوند و انواع بسیار گوناگونی از پروتئین‌ها را ایجاد می‌کنند. این توالی اسیدهای آمینه و بنیان‌های مختلف موجود در آنها و اتصال و ارتباط بین این بنیان‌ها در اسید آمینه‌های مختلف موجب می‌شود ساختار سه‌بعدی منحصر به فرد هر پروتئین و عملکرد خاص آن تعیین شود. پروتئین دارای چهار ساختار می‌باشد و ساختار سوم ساختاری است که در آن زنجیره‌های پلی پپتیدی فعال می‌شوند. در این سطح، هر پروتئین یک شکل سه بعدی خاصی دارد و گروه‌های عاملی را در سطح بیرونی خود ارائه می‌کند به آن اجازه می‌دهد با مولکول‌های دیگر برهم‌کنش داشته باشد و عملکرد منحصربه‌فرد خود را داشته باشد (Baum, 2008).

تجزیه و تحلیل توالی اینترون و اگزون و شناسایی موتیف‌ها و دمین‌ها

بررسی ساختار اگزون–اینترونی می‌تواند شواهد بیشتری در تائید گروه‌بندی درخت فیلوژنتیکی و روابط تکاملی ارائه نماید، زیرا این نوع تنوع ساختاری اغلب نقش مهمی در تکامل خانواده‌های ژنی دارد (Zhang *et al.*, 2012). برای بررسی الگوی توزیع ساختار اینترون و اگزون از طریق مقایسه توالی CDS نسبت به DNA ژنومی ژن مورد نظر، با استفاده از سایت GSDS با وارد کردن شماره دسترسی ژن مورد نظر (X91953.1) صورت گرفت (Hu *et al.*, 2015). مشخص گردید روی توالی مورد نظر یک اینترون در انتهای '5 می‌تواند وجود داشته باشد که بیشتر یک 5'UTR است و موقعیت اگزون از نوکلئوتید ۱۴–۴۶۶ روی ژن می‌باشد. این ژن فقط یک اگزون دارد و اینترونی داخل آن دیده نمی‌شود بنابراین برای بیان هترولوگ آن در باکتری و برای همسانه‌سازی به روش ساده‌تر می‌توان آن را از روی DNA تکثیر داد و نیازی به استخراج RNA و سنتز cDNA برای همسانه‌سازی نیست.

براساس نتایج سرور MOTIF Search مشخص گردید در توالی پروتئین مورد نظر ۳۸ موتیف وجود دارد که هر کدام وظیفه خاصی به پروتئین می‌دهند (شکل ۳). برخی وظایفی که به خاطر وجود این موتیف‌ها می‌توان در این پروتئین پیش‌بینی کرد عبارتند از: زیر واحد گاما و تئو DNA پلیمرازIII،

60S ،alpha-ketoglutarate decarboxylase ،ribosomal protein L19-like protein ،conjugal transfer protein TrbL NADH ،dehydrogenase subunit C ،cytoskeletal ،protein RodZ ،nitrous-oxide reductase ،flagellin .Predicted solute binding protein در آنالیز دیگری که برای شناسایی موتیف‌های حفاظت شده در توالی این پروتئین انجام گرفت مشخص گردید در توالی این پروتئین سه موتیف حفاظت شده وجود دارد که از این میان می‌توان به electron transport complex protein RnfC (PRK05035) و dihydrolipoyllysine-succinyltransferase residue (PRK05704) اشاره کرد.

بررسی وجود دمین نیز در این پروتئین با SMART نشان داد که پروتئین مورد نظر دارای چهار دمین به نام‌های transmembrane region (از کدون ۲۴-۷) و سه دمین دیگر به نام‌های low complexity (از کدون ۶۳-۵۰ و ۱۳۲-۹۷ و ۱۴۸-۱۳۵) است.

شکل ۳. موتیف‌های موجود در توالی پروتئین. الف) شناسایی موتیف‌ها توسط سایت MOTIF SEARCH صورت گرفت که وجود این موتیف‌ها وظایف خاصی به پروتئین می‌دهد. ب) موتیف‌های حفاظت شده پروتئین توسط سایت MEME سه موتیف حفاظت شده برای این پروتئین از طریق سایت MEME گزارش شد که در این شکل موقعیت هر یک از این موتیف‌ها در توالی نشان داده شده است.

Figure 3. The motifs in the protein sequence. A: Identification of motifs by the MOTIF SEARCH site, where these motifs have specific functions for the protein. B: Protein-protected motifs were reported by the site MEME 3 Protected Motif for this protein via the MEME site, in which the position of each of these motifs is shown in sequence.

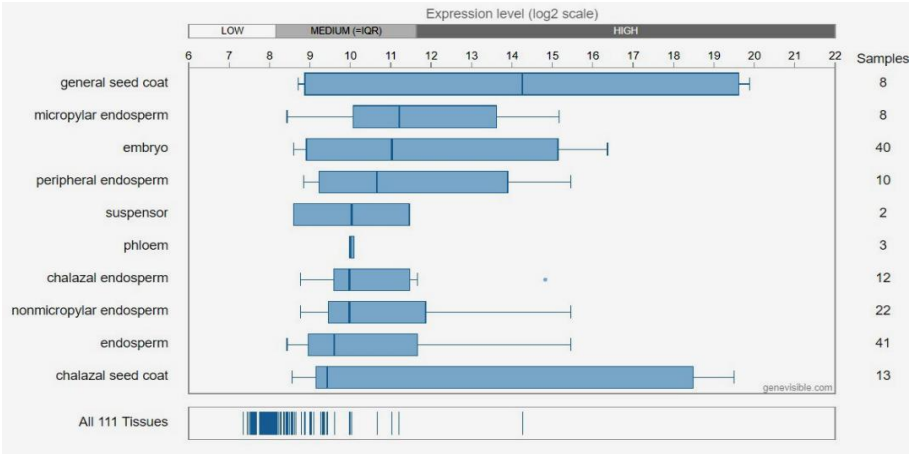
Crop Biotech.
Autumn (2022) 12(1): 55-67 DOI: 10.30473/cb.2023.66084.1892

شناسایی توالی‌های پروموتور و بررسی بیان ژن

هر ژن در ناحیه تنظیمی خود دارای توالی‌های منحصربه‌فردی از عناصر تنظیمی است که بیان ژن را در زمان و مکان مختلف برای انجام وظایف خود تنظیم می‌کند. بنابراین، درک جامع از این عناصر تنظیمی و نحوه کنترل بیان ژن می‌تواند در تعیین نقش ژن مفید باشد (Guo *et al*., 2021). شناسایی توالی تنظیمی مربوط به ژن مورد نظر از طریق سایت PlantCARE صورت گرفت و نتایج تجزیه و تحلیل پایگاه داده ۱۸ نوع مختلف توالی یا عناصر تنظیمی برای این پروتئین ناشناخته شناسایی کرد که برخی از این توالی تنظیمی در گیاهان دیگر مثل ذرت، یولاف، توتون، نخود، سیب زمینی کلزا، کلم نیز شناخته شده و هر کدام وظیفه خاصی را بر عهده دارند که در خصوص این پروتئین نیز وجود این توالی تنظیمی باعث شده است به واسطه برخی از آن‌ها پروتئین وظیفه خاصی را داشته باشد (جدول ۳).

یکی از روش‌های بسیار کارآمد در تعیین عملکرد پروتئین‌ها پیش بینی بیوانفورماتیکی محل جایگیری و فعالیت پروتئین‌ها در سلول‌ها می‌باشد. جایگیری پروتئین‌ها در میتوکندری، کلروپلاست، سیتوپلاسم و یا هسته تا میزان زیادی تعیین کننده نقش آن پروتئین می‌باشد، بنابراین برای تعیین عملکرد پروتئین‌ها از ابزار TargetP سایت EXPASY استفاده شد (Emanuelsson *et al*., 2007) و نتایج وجود سیگنال پپتید در توالی پروتئینی را نشان داد که به احتمال فراوان پروتئین در مسیرهای ترشحی وارد می‌شود. نتایج بررسی با وب سایتCELLO v.2.5 برای شناسایی جایگاه سلولی پروتئین (Bienert *et al*., 2017) نتایج نشان داد که این پروتئین درهسته، میتوکندری، و فضای خارج سلولی قرار دارد که میزان حضور آن در فضای خارج سلولی نسبت به داخل میتوکندری و هسته بیشتر است که این امر احتمالاً به دلیل وجود سیگنال پپتید و وارد شدن پروتئین مذکور در مسیر ترشحی باشد.

بر اساس اطلاعات موجود در سایت UniProt برای این پروتئین می‌توان این ژن در روی کروموزوم شماره یک گیاه آرابیدوپسیس تالیانا قرار داشته و دو وظیفه تمایز پوشش بذر و فرایند بیوسنتز مواد مترشحه در اثر نور را پیش بینی کرد (Tsai *et al*., 2017). در مورد بیان پروتئین نیز از میان ۱۱۱ بافت آزمایش شده توسط وب سایت GENEVESTIGATOR محل بیان این ژن در ۱۰ بافت مختلف گیاهی است و اکثرا در آندوسپرم (Pignatta *et al*., 2014)، بذر (Mergner *et al*., 2020) و جنین گیاهی (Pignatta *et al*., 2014) می‌باشد (شکل ۴).

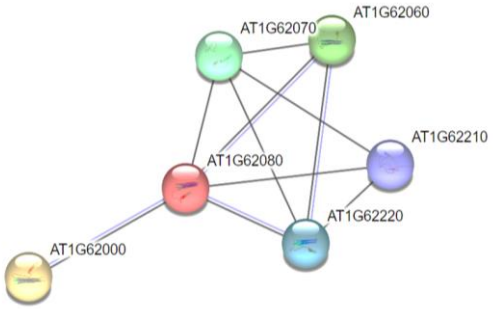


شکل ۴. پیش‌بینی محل بیان این پروتئین به کمک داده‌های سایت GENEVESTIGATOR

* نویسنده مسئول: مقصود پژوهنده
E-mail: pazhouhandeh@gmail.com

Figure 4. The prediction of the protein expression by GENEVESTIGATO.

عملکرد مولکولی یک پروتئین با وظایف ناشناخته معمولاً بر اساس توالی یا شباهت ساختاری آن به پروتئین‌هایی با عملکرد مشخص تعیین می‌شود (Galperin & Koonin, 2004). براساس اطلاعات موجود در سایت UniProt مشخص گردید این پروتئین با پروتئین AT1G62000 در گیاه آرابیدوپسیس تالیانا دارای ۹۰ درصد هومولوژی است. برهمنکش پروتئین مورد مطالعه با سایر پروتئین‌های دیگر گیاه آرابیدوپسیس تالیانا در قالب نقشه ژنی آورده شده است (شکل ۵). براساس آنالیز سایت STRING این پروتئین با پروتئین‌های AT1G62000، AT1G62070، AT1G62220 AT1G6206 و AT1G62210 دارای برهمکنش است. از بین پروتئین‌هایی که با پروتئین مورد مطالعه میانکنش نشان دادند هیچ یک جزو پروتئین‌های شناخته شده در آرابیدوپسیس نبودند. انتولوژی ژن نشان داد این پروتئین در فرآیند بیوسنتزی موسیلاژ و پوشش بذر نقش داشته و دارای سیگنال پپتید در انتهای N ترمینال خود و در موقعیت ۲۱-۱ است.



شکل ۵. شبکه برهکنش پروتئین مورد نظر با سایر پروتئین‌های گیاه آرابیدوپسیس

Figure 5. Interaction network of the target protein with other proteins in Arabidopsis thaliana.

جدول ۳. اسامی عناصر تنظیمی و عملکرد آن‌ها

Table 3. The names of the regulatory elements and their functions

Site Name	Sequence	Function
ABRE	CACGTG ACGTG	cis-acting element involved in the abscisic acid responsiveness
AE-box	AGAAACTT	part of a module for light response
ARE	AAACCA	cis-acting regulatory element essential for the anaerobic induction
Box 4	ATTAAT	part of a conserved DNA module involved in light responsiveness
CAAT-box	CCAAT	common cis-acting element in promoter and enhancer regions
CAT-box	GCCACT	cis-acting regulatory element related to meristem expression
CTAG-motif	ACTAGCAGAA	
ERE	ATTTCATA ATTTTAAA	
G-Box	CACGTG	cis-acting regulatory element involved in light responsiveness
GATA-motif	AAGGATAAGG	part of a light responsive element
MBS	CAACTG	MYB binding site involved in drought-inducibility
MYC	CATGTG	
Myb	CAACTG	
STRE	AGGGG	
TATA-box	TATA ATTATA	core promoter element around -30 of transcription start
TGA-element	AACGAC	auxin-responsive element
Unnamed 4	CTCC	
W box	TTGACC	

شناسایی miRNA و ژن هدف مرتبط با آن

بررسی‌های مربوط به شناسایی miRNA با استفاده از سرور mirbase صورت گرفت و مشخص گردید miRNA به نام ath-miR5021 در توالی X91953.1 وجود دارد که ۱۸ نوکلئوتیدی (ugagaagaagaagaagaga) بوده و توالی ۴۴۴-۴۲۷ از این ژن را مورد هدف قرار می‌دهد. در یک مطالعه انجام شده در گیاه آرابیدوپسیس مشخص گردید این miRNA ژن 2Fe-2S ferredoxin را هدف قرار می‌دهد که وظیفه بیولوژیکی آن مرتبط با iron sulfur cluster binding بوده و در نتیجه موجب پاسخ به شوری در گیاهان آرابیدوپسیس تالیانا می‌شود (Verma *et al*., 2014). در مطالعه دیگر فعالیت microRNA در آرابیدوپسیس در germline نر بررسی گردید و مشخص شد miR5021 می‌تواند در دانه گرده حضور داشته و موجب خاموشی برخی ژن‌ها در مسیر تمایز دانه گرده شود (Borges *et al*., 2011).

نتیجه‌گیری کلی

برای شناسایی عملکرد پروتئین و ژن در گذشته از روش‌های *in-vivo* یا *in-vitro* استفاده می‌شد اما امروزه پیشرفت ابزارهای بیوانفورماتیک و همچنین افزایش پایگاه‌های اطلاعاتی بیولوژیکی، روش *in-silico* می‌تواند گام بزرگی در شناسایی عملکرد ژن‌ها و پروتئین‌های ناشناخته باشد. آنالیزهای بیوانفورماتیکی روی توالی ژن گیاه آرابیدوپسیس تالیانا با شماره دسترسی X91953.1 به اندازه‌ی bp ۶۷۶ نشان داد که محصول این ژن پروتئینی با ۱۵۰ اسیدآمینه و وزن مولکولی ۱۵۱۲۷/۱۹ دالتون است. به احتمال فراوان این پروتئین دارای سیگنال پپتید بوده و در مسیرهای ترشحی وارد می‌شود و دارای ۳۸ موتیف، سه موتیف حفاظت شده و چهار دمین بوده که به واسطه وجود هریک احتمال می‌رود پروتئین دارای وظایف مختلفی باشد. این ژن دارای ۱۸ نوع توالی تنظیمی است که وظایف برخی از آن‌ها شناخته شده است که از جمله این وظایف می‌توان به پاسخ به نور، فعالیت عناصر Cis برای بیان در مریستم اشاره کرد. محل تنظیم یک microRNA نیز روی رونوشت این ژن وجود دارد که این microRNA در پاسخ به شوری و همچنین در جنین فعالیت دارد. در نهایت براساس همه آنالیزهای صورت گرفته دو وظیفه تمایز در پوشش بذر و فرایند بیوسنتز مواد مترشحه در اثر نور را برای این پروتئین می‌توان پیش بینی کرد. یافته‌های بیوانفورماتیکی تحقیق حاضر برای این ژن ناشناخته که هیچ اطلاعاتی در مورد آن موجود نبود، بسیار ارزشمند بوده و راهنمایی خواهد بود تا هدفمندتر با روش‌های آزمایشگاهی بیولوژی و ژنتیک مولکولی تایید و اثبات شوند.

REFERENCES

Allahi, S., Sohani, M. M., & Hasani Kumleh, H. (2017). In silico identification of the PLD gene family and analysis of their expression pattern in response to salt stress in Medicago truncatula. *Genetic Engineering and Biosafety Journal, 6*(1), 143-156.

Azizi-Dargahlou, S., & Fazeli-Nasab, B. (2022). Identification of the Properties and Function of the Unknown Protein with Accession Number AT2G15110. 1 on the TAIR Website. *Gene, Cell and Tissue*, e122297. doi: 10.5812/gct-122297.

Baum, D. (2008). Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education, 1*(1), 190.

Bienert, S., Waterhouse, A., Beer, T. A. P. d., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res, 45*, 313-319.

Bollag, D. M., Rozycki, M. D., & Edelstein, S. J. (1996). *Protein Methods, 2nd Edition* (2nd Edition ed.): Published by Wiley Publishers.

Borges, F., Pereira, P. A., Slotkin, R. K., Martienssen, R. A., & Becker, J. D. (2011). MicroRNA activity in the Arabidopsis male germline. *Journal of experimental botany, 62*(5), 1611-1620.

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell, 136*(4), 642-655. doi:10.1016/j.cell.2009.01.035

Clare, A., Karwath, A., Ougham, H., & King, R. D. (2006). Functional bioinformatics for Arabidopsis thaliana. *Bioinformatics, 22*(9), 1130-1136.

Edwards, Y. J., & Cottage, A. (2003). Bioinformatics methods to predict protein structure and function. *Molecular biotechnology, 23*(2), 139-166.

Edwards, Y. J. K., & Cottage, A. (2003). Bioinformatics Methods to Predict Protein Structure and Function. *Molecular Biotechnology, 23*, 139-166.

Emanuelsson, O., Brunak, S., Von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols, 2*(4), 953-971.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., ... Mistry, J. (2013). Pfam: the protein families database. *Nucleic acids research, 42*(D1), D222-D230.

Galperin, M. Y., & Koonin, E. V. (2004). 'Conserved hypothetical'proteins: prioritization of targets for experimental study. *Nucleic acids research, 32*(18), 5452-5463.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*: The Proteomics Protocols Handbook.

Guo, X., Ohler, U., & Yildirim, F. (2021). How to find genomic regions relevant for gene regulation. *Medizinische Genetik, 33*(2), 157-165.

Hajibarat, Z., Saidi, A., & Hajibarat, Z. (2018). Bioinformatics analysis of MADS-box in Brachypodium distachyon. *Crop Biotechnology, 8*(23), 1-15.

Hawkins, T. (2007). function prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology, 5*, 1-30.

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., & Gao, G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics $V 31*(8), 1296-1297.

Kuroda, K., Kato, M., Mima, J., Ueda, M. (2006). Systems for the detection and analysis of protein-protein interactions. *Appl Microbiol Biotechnol. 71(2),* 127-36. doi: 10.1007/s00253-006-0395-5.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., ... Bryant, S. H. (2012). CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research, 41*(D1), D348-D352. doi:10.1093/nar/gks1243.

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., ... Messerer, M. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature, 579*(7799), 409-414.

Pino, L., Lin, A., Bittremieux, W. (2019). 2018 YPIC Challenge: A Case Study in Characterizing an Unknown Protein Sample. *Journal of Proteome Research. 18(11),* 3936-3943. doi: 10.1021/acs.jproteome.9b00384.

Pignatta, D., Erdmann, R. M., Scheer, E., Picard, C. L., Bell, G. W., & Gehring, M. (2014). Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife, 3*, e03198.

Shumilin, I. A., Cymborowski, M., Chertihin, O., Jha, K. N., Herr, J. C., Lesley, S. A., ... Minor, W. (2012). Identification of unknown protein function using metabolite cocktail screening. *Structure, 20*(10),1715-1725.

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., ... Xenarios, I. (2012). New and continuing developments at PROSITE. *Nucleic acids research, 41*(D1), D344-D347.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., ... Ploetz, L. (2007). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research, 36*(suppl_1), D1009-D1014.

Tsai, A. Y., Kunieda, T., Rogalski, J., Foster, L. J., Ellis, B. E., & Haughn, G. W. (2017). Identification and Characterization of Arabidopsis Seed Coat Mucilage Proteins. *Plant Physiol, 173*, 1059-1074.

Verma, S. S., Sinha, R., Rahman, M., Megha, S., Deyholos, M. K., & Kav, N.N. (2014). miRNA-mediated posttranscriptional regulation of gene expression in ABR17-transgenic Arabidopsis thaliana under salt stress. *Plant molecular biology reporter, 32*(6), 1203-1218.

Wang, Y., Zhang, H., Zhong, H., & Xue, Z. (2021). Protein domain identification methods and online resources. *Computational and structural biotechnology journal, 19*, 1145-1153.

Xiong, J. (2006). *Essential bioinformatics*: Cambridge University Press.

Zhang, Y., Gao, M., Singer, S., Fei, Z., Wang, H., & Wang, X. (2012). Genome-wide identification and analysis of the TIFY gene family in grape. *PLoS One, 7*, e44465.