# شناسایی ژن های آنزیم زایالناز از متاژنوم شکمبه نشخوارکنندگان

امیرحسین انتظاری و دلارام حسینی

۲۰ بهمن ۱۴۰۳

## چکیده

متاژنوم ها به عنوان یک منبع غنی از تنوع ژنتیکی میکروبی عمل می کنند و فرصتی منحصر به فرد برای کشف آنزیم های جدید با اهمیت صنعتی ارائه می دهند. این مطالعه بر شناسایی ژن‌های کدکننده زایلاناز از متاژنوم شکمبه نشخوارکنندگان متمرکز است و زایلانازهای پایدار در برابر حرارت را با کاربردهای بالقوه در تولید سوخت زیستی، پردازش مواد غذایی، خوراک حیوانات و صنعت کاغذ هدف قرار می‌دهد. این مطالعه یک رویکرد سه مرحله‌ای را دنبال می‌کند: (۱) شناسایی توالی‌های زایلاناز بالقوه از طریق جستجوهای مبتنی بر شباهت در برابر زایلانازهای مقاوم در برابر حرارت شناخته شده با استفاده از BLAST و DIAMOND و غیره (۲) خوشه‌بندی توالی‌های مشابه با استفاده از CD-HIT برای حذف افزونگی و انتخاب توالی‌های نماینده، و (۳) استفاده از مدل‌سازی توالی خاص، ماتریس‌ها (PSSM) و مدل‌های مارکوف پنهان (HMM) برای اصلاح فهرست نامزدها بر اساس موتیف‌های حفاظت‌شده. تجزیه و تحلیل ما توالی‌های Xylanase پتانسیل X را از مجموعه داده‌های متاژنومی شناسایی کرد، که از آن‌ها توالی‌های غیر زائد Y پس از خوشه‌بندی انتخاب شدند. مدل‌سازی منطقه حفاظت‌شده این فهرست را به کاندیدهای زایلاناز بسیار مطمئن Z اصلاح کرد. این یافته‌ها نشان می‌دهد که میکروبیوم شکمبه دارای آرایه متنوعی از آنزیم‌های زایلاناز است، که بسیاری از آنها ممکن است ویژگی‌های منحصربه‌فردی را برای کاربردهای صنعتی نشان دهند. این مطالعه پتانسیل داده‌کاوی متاژنومی را در کشف آنزیم برجسته می‌کند و پایه‌ای را برای اعتبار سنجی تجربی بیشتر زایلانازهای شناسایی شده تنظیم می‌کند.

# بخش ۱

# مقدمه

## ۱.۱ پیش زمینه

### ۱.۱.۱ کشف آنزیم ها از متاژنوم ها

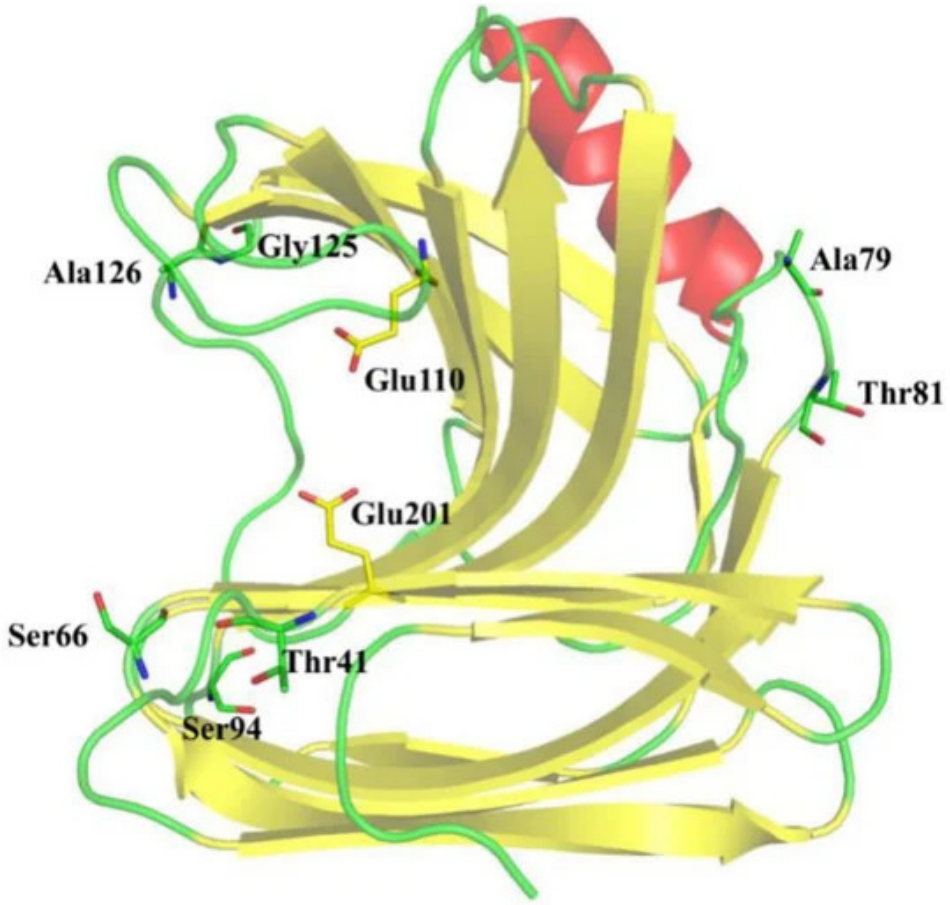
متاژنومیکس با امکان تجزیه و تحلیل مستقیم مواد ژنتیکی بازیابی شده از نمونه های محیطی، مطالعه جوامع میکروبی را متحول کرده است. برخلاف روش‌های سنتی مبتنی بر کشت، متاژنومیکس دسترسی به تنوع میکروبی گسترده‌ای را فراهم می‌کند که در شرایط آزمایشگاهی غیرقابل کشت باقی می‌ماند. این رویکرد منجر به کشف آنزیم های جدید با خواص کاتالیزوری منحصر به فرد شده است که بسیاری از آنها کاربردهای صنعتی و بیوتکنولوژیکی قابل توجهی دارند. در میان این آنزیم ها، زایلانازها به دلیل توانایی آنها در تجزیه زایلان، دومین پلی ساکارید فراوان در طبیعت، از اهمیت ویژه ای برخوردار هستند. زایلانازها زایلان را به قندهای ساده‌تر تجزیه می‌کنند و آنها را برای تولید سوخت زیستی، فرآوری غذا و خوراک و کاربردهای صنعتی ضروری می‌سازد. شناسایی آنزیم‌های زایلاناز جدید از متاژنوم‌ها می‌تواند منجر به بیوکاتالیست‌های کارآمدتر با پایداری، فعالیت و ویژگی سوبسترای بهتر شود.

### ۲.۱.۱ اهمیت زایلانازهای میکروبی

زایلانازهای میکروبی نقش مهمی در صنایع مختلف دارند:

- **تولید سوخت زیستی:**زایلانازها به تجزیه زیست توده گیاهی به قندهای قابل تخمیر کمک می کنند و عملکرد بیواتانول را بهبود می بخشند.
- **صنعت خمیر و کاغذ:** در فرآیندهای سفید کردن سازگار با محیط زیست برای کاهش استفاده از مواد شیمیایی و بهبود کیفیت کاغذ استفاده می شود.
- **فرآوری غذا و خوراک:** افزایش قابلیت هضم در خوراک دام و بهبود بافت محصولات پخته شده.
- **کشاورزی و بیوتکنولوژی:** کمک به تخریب زیست توده گیاهی، ترویج شیوه های کشاورزی پایدار.

با توجه به اهمیت صنعتی آنها، کشف زایلانازهای مقاوم در برابر حرارت و مقاوم در برابر pH بسیار ارزشمند است. این خواص عملکرد آنزیم را در شرایط شدید افزایش می دهد و آنها را در فرآیندهای صنعتی موثرتر می کند.




شکل ۱.۱: xylanase

### ۳.۱.۱ چرا متاژنوم شکمبه؟

میکروبیوم شکمبه نشخوارکنندگان یک مخزن غنی از میکروارگانیسم های تجزیه کننده لیگنوسلولز است. نشخوارکنندگان برای تجزیه موثر مواد گیاهی به جوامع میکروبی خود متکی هستند و شکمبه را به محیطی ایده آل برای جستجوی آنزیم های زایلاناز جدید با فعالیت قوی تبدیل می کند. با تجزیه و تحلیل توالی‌های متاژنومی مشتق شده از شکمبه، محققان می‌توانند زایلانازهای جدیدی را کشف کنند که برای عملکرد تحت شرایط فیزیولوژیکی طبیعی تکامل یافته‌اند و اغلب پایداری در دمای بالا و انعطاف‌پذیری در محیط‌های pH اسیدی یا قلیایی از خود نشان می‌دهند.

هدف این پروژه شناسایی و مشخص کردن ژن‌های کدکننده زایلاناز از متاژنوم شکمبه، استفاده از ابزارهای بیوانفورماتیک برای شناسایی توالی، خوشه‌بندی و مدل‌سازی منطقه حفاظت‌شده است. نتایج ممکن است به کشف زایلانازهای مرتبط صنعتی جدید کمک کند و درک ما را از تخریب لیگنوسلولز میکروبی در اکوسیستم شکمبه افزایش دهد.

## ۲.۱ اهداف

هدف اصلی این مطالعه شناسایی و شناسایی ژن‌های کدکننده زایلاناز از متاژنوم شکمبه نشخوارکنندگان، با استفاده از روش‌های بیوانفورماتیک برای شناسایی، خوشه‌بندی و تجزیه و تحلیل توالی‌های زایلاناز پایدار حرارتی بالقوه است. با توجه به اهمیت صنعتی زایلانازها در سوخت‌های زیستی، غذا، خوراک و پردازش کاغذ، هدف این مطالعه کشف زایلانازهای جدیدی است که ممکن است کارایی و پایداری را در شرایط شدید ارائه دهند.

برای دستیابی به این هدف، پروژه در سه هدف اصلی ساختار یافته است:

۱. شناسایی توالی های بالقوه زایلاناز

- جستجوهای مبتنی بر شباهت را با استفاده از ابزارهایی مانند BLAST، DIAMOND یا HMMER برای مقایسه توالی متاژنومی شکمبه در برابر زایلاناز‌های مقاوم در برابر حرارت.
- انتخاب توالی هایی با شباهت قابل توجه به عنوان کاندیدهای بالقوه زایلاناز.
- ترجمه توالی های شناسایی شده را برای تجزیه و تحلیل بیشتر به دنباله های پروتئینی.

۲. خوشه بندی و انتخاب توالی های نماینده

- از CD-HIT برای خوشه بندی توالی های بسیار مشابه استفاده و افزونگی در مجموعه داده را کاهش دادیم.
- توالی های نماینده را از هر خوشه انتخاب کردیم تا یک مجموعه داده زایلاناز غیر زائد به دست آوریم.

۳. مدلسازی مناطق حفاظت شده و فیلترینگ توالی

- ساخت یک مدل برای منطقه حفاظت‌شده زایلاناز با استفاده از ماتریس‌های امتیازدهی خاص موقعیت (PSSM)، مدل‌های پنهان مارکوف (HMMs)، یا عبارات منظم
- توالی های نماینده را با استفاده از این مدل فیلتر کردیم تا لیست کاندیدهای قوی زایلاناز را بهبوود یابد.

با پیروی از این روش تحقیق ساختاریافته بیوانفورماتیک، هدف این مطالعه کمک به کشف آنزیم، ارائه نامزدهای بالقوه برای کاربردهای صنعتی و در عین حال افزایش درک ما از تنوع زایلاناز در میکروبیوم شکمبه است.

## ۳.۱ منابع داده

این مطالعه از داده‌های متاژنومی به دست آمده از میکروبیوم شکمبه نشخوارکنندگان، یک اکوسیستم میکروبی پیچیده که به دلیل توانایی آن در تجزیه موثر پلی‌ساکاریدهای گیاهی شناخته شده است، استفاده می‌کند. منابع داده این پروژه عبارتند از:

۱. **Contigs متاژنوم شکمبه**

- مجموعه داده‌ای حاوی contigs های اسمبل شده از متاژنوم شکمبه نشخوارکنندگان.
- این توالی ها نشان دهنده مواد ژنتیکی میکروبی استخراج شده از محیط شکمبه هستند که منبع غنی از ژن های بالقوه زیلاناز را فراهم می کنند.
- دسترسی: مجموعه داده contigs اسمبل شده از طریق لینک زیر در دسترس است: لینک گوگل درایو

۲. **توالی های زایلاناز مرجع**

- مجموعه ای از ۱۱ توالی آنزیم زایلاناز به عنوان مرجعی برای جستجوهای مبتنی بر شباهت عمل می کند.
- این توالی ها بر اساس توانایی آنها برای عملکرد تحت شرایط صنعتی مرتبط مانند دمای بالا و ثبات pH تنظیم شده اند.
- دسترسی: توالی‌های زایلاناز مرجع در لینک زیر در دسترس است: لینک گوگل درایو

۳. **پایگاه ها و ابزارهای بیوانفورماتیک**
علاوه بر مجموعه داده های فوق، ابزارها و پایگاه های اطلاعاتی بیوانفورماتیک در دسترس عموم برای مقایسه و تجزیه و تحلیل توالی استفاده خواهند شد:

- BLAST+: NCBI برای جستجوی شباهت در برابر توالی های زایلاناز شناخته شده (Help BLAST)
- پایگاه های داده پروتئین (NCBI، UniProt): برای تأیید حاشیه نویسی های عملکردی توالی های شناسایی شده.
- CD-HIT: برای خوشه بندی توالی های بسیار مشابه و کاهش افزونگی.
- HMMER: برای شناسایی دامنه های حفاظت شده در توالی های زایلاناز.

## ۴.۱ روش ها

برای شناسایی و مشخص کردن ژن‌های کدکننده زایلاناز از متاژنوم شکمبه، این مطالعه یک گردش کار ساختار یافته بیوانفورماتیک شامل سه مرحله کلیدی را دنبال می‌کند: شناسایی توالی، خوشه‌بندی، و مدل‌سازی منطقه حفاظت‌شده. در ابتدا، توالی‌های بالقوه زایلاناز از طریق جستجوهای مبتنی بر شباهت با استفاده از ابزارهایی مانند BLAST، DIAMOND یا HMMER شناسایی می‌شوند، و عوامل متاژنومیک را با مجموعه‌ای از زایلانازهای شناخته شده مقاوم در برابر حرارت مقایسه می‌کنند. سپس توالی های شناسایی شده با استفاده از CD-HIT برای حذف افزونگی و انتخاب توالی های نماینده برای تجزیه و تحلیل بیشتر، خوشه بندی می شوند. در مرحله نهایی، مدل‌سازی ناحیه حفاظت‌شده با استفاده از ماتریس‌های امتیازدهی خاص موقعیت (PSSM)، مدل‌های مارکوف پنهان (HMMs)، یا عبارات منظم برای اصلاح مجموعه داده‌ها و استخراج کاندیدهای زایلاناز با اطمینان بالا انجام می‌شود.

این روش یک رویکرد سیستماتیک و محاسباتی کارآمد را برای کشف آنزیم‌های زایلاناز جدید با کاربردهای صنعتی بالقوه تضمین می‌کند. جزئیات هر مرحله در زیر بخش های زیر توضیح داده شده است.

# بخش ۲

# گام ۱: شناسایی توالی های بالقوه زایلاناز

اولین مرحله در این مطالعه شامل شناسایی توالی‌های متاژنومی است که به طور بالقوه آنزیم‌های زایلاناز مقاوم در برابر حرارت را رمزگذاری می‌کنند. این از طریق جستجوی مبتنی بر شباهت در برابر مجموعه‌ای از توالی‌های زایلاناز شناخته شده با استفاده از ابزارهای هم‌ترازی توالی محاسباتی به دست می‌آید. هدف اصلی فیلتر کردن contig هایی است که مشابهت قابل توجهی با زایلانازهای مرجع دارند و اطمینان حاصل شود که فقط مرتبط ترین توالی ها برای تجزیه و تحلیل بیشتر حفظ می شوند.

## رویکرد: جستجوی شباهت با استفاده از DIAMOND، BLAST یا HMMER

برای شناسایی ژن های بالقوه کد کننده زایلاناز، از ابزارهای زیر استفاده می شود:

- BLAST+ (Basic Local Alignment Search Tool): یک الگوریتم مقایسه توالی پرکاربرد است که مناطق شباهت بین کانتیگ های متاژنومی و توالی های زایلاناز شناخته شده را شناسایی می کند.
- DIAMOND : جایگزین سریع‌تری برای BLAST، بهینه‌سازی شده برای داده‌های متاژنومی در مقیاس بزرگ، که می‌تواند به سرعت توالی‌ها را در مقابل پایگاه‌های داده پروتئینی تراز کند.
- HMMER (جستجوی مبتنی بر مدل مارکوف پنهان): ابزار احتمالی است که نقوش حفاظت شده و حوزه های عملکردی مشخصه آنزیم های زایلاناز را تشخیص می دهد.

### در این پروژه ما برای تعیین شباهت از BLAST استفاده می‌کنیم:

BLASTX به دلیل دقت بالای آن در تشخیص توالی های هومولوگ انتخاب شد، در حالی که از DIAMOND برای پردازش سریعتر مجموعه داده های متاژنومی بزرگ استفاده می شود. HMMER برای تشخیص شباهت‌های مبتنی بر نمایه استفاده می شود، که به شناسایی هومولوگ‌های راه دور کمک می‌کند که تنها با شباهت توالی ثبت نشده‌اند. آستانه‌های فیلتر فقط مطابق با اطمینان بالا حفظ می‌شوند و از انتخاب نامزد قابل اطمینان اطمینان می‌دهند. نحوه عملکرد BLAST:

۱. دنباله ورودی را می‌گیرد.(دنباله‌ای از DNA، RNA یا پروتیین‌ها)

۲. به دنبال دنباله‌های مشابه در دنباله‌های شناخته شده و پایگاه داده می‌گردد.

۳. محاسبه معیار شباهت

## ۱.۲ دستورات ترمینال برای شناسایی توالی

۱. نصب ابزار های مورد نیاز:

```
# Install BLAST using Bioconda
!conda install -c bioconda blast -y

# Install DIAMOND for faster sequence search
!conda install -c bioconda diamond -y
```

شکل ۱.۲: نصب ابزار های مورد نیاز

۲. آماده سازی پایگاه داده :BLAST
با اجرای این دستور در ترمینال از روی فایل (۱۱ دنباله‌ی شناخته شده زایلاناز) thermo_xylanase.fasta فایل سازگار xylanase_db ساخته می‌شود.

```
!makeblastdb -in data/thermo_xylanase.fasta -dbtype prot -out xylanase_db
Executed at 2025.02.07 18:52:31 in 363ms


Building a new DB, current time: 02/07/2025 18:52:31
New DB name:   /home/amir/Documents/university/Semester9/bioinformatics/project2/xylanase_db
New DB title:  data/thermo_xylanase.fasta
Sequence type: Protein
Deleted existing Protein BLAST database named /home/amir/Documents/university/Semester9/bioinformatics/project2/xylanase_db
Keep MBits: T
Maximum file size: 3000000000B
Adding sequences from FASTA; added 11 sequences in 0.000927925 seconds.
```

شکل ۲.۲: آماده سازی پایگاه داده BLAST

۳. اجرای BLASTX را برای شناسایی توالی های زایلاناز بالقوه
این دستور فایل تولید شده در قسمت قبل و پایگاه داده را گرفته و فایل نتایج (blast_results.text) را با معیار شباهت(evalue) به مقدار ۱e-۵ تولید می‌کند.

```
!blastx -query data/y5.final.contigs.fa -db xylanase_db -out step1/blast_results.txt -evalue 1e-5 -outfmt 6
```

شکل ۳.۲: آماده سازی پایگاه داده BLAST

۴. blast_results.txt

| | Query | Subject | Identity | Length | Mismatches | Gap Openings | Q. Start | Q. End | S. Start | S. End | E-value | Bit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | k141_3201976 | 6xylanase | 43.137 | 51 | 29 | 0 | 241 | 393 | 41 | 91 | 5.970000e-09 | 43.1 |
| 1 | k141_3201976 | 11xylanase | 42.222 | 45 | 26 | 0 | 262 | 396 | 355 | 399 | 9.720000e-09 | 42.4 |
| 2 | k141_3201976 | 4RecName: | 41.176 | 51 | 30 | 0 | 241 | 393 | 41 | 91 | 8.130000e-08 | 39.7 |
| 3 | k141_3201976 | 9xylanase | 46.667 | 45 | 22 | 1 | 262 | 396 | 8 | 50 | 8.080000e-06 | 33.9 |
| 4 | k141_4482810 | 10xylanase | 33.735 | 83 | 42 | 3 | 228 | 19 | 912 | 994 | 1.180000e-06 | 37.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 22055 | k141_6285217 | 9xylanase | 28.283 | 198 | 106 | 4 | 1268 | 711 | 155 | 328 | 7.290000e-16 | 67.8 |
| 22056 | k141_6285217 | 6xylanase | 23.661 | 224 | 99 | 9 | 1265 | 720 | 207 | 400 | 3.400000e-08 | 44.7 |
| 22057 | k141_6285217 | 4RecName: | 25.248 | 202 | 90 | 8 | 1265 | 753 | 207 | 378 | 4.430000e-08 | 44.3 |
| 22058 | k141_6285217 | 11xylanase | 23.810 | 168 | 84 | 7 | 1265 | 810 | 505 | 644 | 2.690000e-06 | 38.9 |
| 22059 | k141_2755066 | 10xylanase | 23.810 | 63 | 47 | 1 | 19 | 204 | 832 | 894 | 1.480000e-06 | 38.1 |

22060 rows × 12 columns

شکل ۴.۲: نتیجه blast

| | | |
|---|---|---|
| ۱ | شناسایی توالی کوئری | k۱۴۱_۲۷۵۵۰۶۶ |
| ۲ | شناسه توالی زایلاناز تطبیق‌یافته | ۱۰xylanase |
| ۳ | درصد تطابق‌های یکسان | ۸۱۰.۲۳ |
| ۴ | طول ناحیه تطبیق‌یافته | ۶۳ |
| ۵ | تعداد اسیدهای آمینه نامطابق | ۴۷ |
| ۶ | تعداد شکاف‌های ایجادشده در هم‌ترازی | ۱ |
| ۷ | موقعیت شروع در کانتیگ | ۱۹ |
| ۸ | موقعیت پایان در کانتیگ | ۲۰۴ |
| ۹ | موقعیت شروع در توالی زایلاناز | ۸۳۲ |
| ۱۰ | موقعیت پایان در توالی زایلاناز | ۸۹۴ |
| ۱۱ | مقدار معیار شباعت | ۰۶-۴۸e.۱ |
| ۱۲ | امتیاز کیفیت هم‌ترازی | ۱.۳۸ |

جدول ۱.۲: Table Example

۵. اجرای مشابه DIAMOND

```
!diamond makedb --in data/thermo_xylanase-diamond.fasta -d xylanase_db
!diamond blastx -q data/y5.final.contigs.fa -d xylanase_db.dmnd -o step1/diamond_results.txt --evalue 1e-5 --outfmt 6
```

شکل ۵.۲: اجرای مشابه DIAMOND

در مرحله بعد از بین دنباله‌های موجود در فایل نتایج تعدادی از آن‌ها را جدا می‌کنیم. در هنگامی که BLAST اجرا می‌شود معیار شباهت و امتیاز کیفینت همترازی برای هر رشته و رشته‌های موجود در دنباله‌های شناخته شده بدست می‌آید. بر اساس این دو معیار دنباله‌های موجود در فایل نتایج فیلتر می‌شوند و دنباله‌هایی که همخوانی بیشتری با دنباله‌ی اصلی دارند برگزیده خواهند شد.

**معیارهای انتخاب: آستانه تشابه**

برای اطمینان از صحت شناسایی زایلاناز، توالی ها بر اساس معیارهای زیر فیلتر می شوند:

۱. $E - value \leq 1e - 5$: (نشان دهنده شباهت آماری معنی دار).

۲. $\text{Percentage Identity} \geq 30\%$ :(برای حفظ توالی هایی با سطح معنی داری از شباهت به زایلانازهای شناخته شده).

۳. $\text{Query Coverage} \geq 50\%$ :(تضمین اینکه بخش قابل توجهی از contig با دنباله های مرجع همسو می شود).

این آستانه‌ها حساسیت و ویژگی را متعادل می‌کنند و امکان تشخیص زایلانازهای نزدیک و بالقوه جدید را فراهم می‌کنند و در عین حال موارد مثبت کاذب را به حداقل می‌رسانند.

**فیلتر کردن:**

**چرا نیاز است فیلتر انجام دهیم؟**

BLASTX معیار شباهت را ارائه می دهد، اما به طور خودکار نتایج را فیلتر نمی کند. همه ترازهای بالاتر از یک آستانه مشخص را خروجی می دهد. با این حال، برخی از این تطابق‌ها ممکن است با اطمینان کم یا تطابق جزئی باشند، به این معنی که برای افزایش دقت به فیلتر دستی نیاز داریم. برخی از ترازها ممکن است دارای درصد کمی همترازی باشند (مثلاً ۳۰-۲۵٪) و ممکن است زایلاناز واقعی نباشند. ما باید یک برش تعیین کنیم تا فقط دنباله‌های مشابه حفظ شوند. همچنین ترازهای کوتاه ممکن است کل پروتئین را پوشش ندهند. یک تطابق کوتاه (مثلاً ۳۰ اسید آمینه از یک پروتئین ۳۰۰ اسید آمینه) ممکن است شواهد کافی مبنی بر اینکه یک توالی آنزیم کامل زایلاناز را کد می کند، نباشد. فیلتر کردن بر اساس طول تراز به حذف این موارد کمک می کند.

```python
# Load BLASTX results (assuming the file is tab-separated)
blastx_df = pd.read_csv("blast_results.txt", sep="\t", header=None)

# Assign column names based on BLASTX output format
blastx_df.columns = ["Query_ID", "Subject_ID", "%_Identity", "Alignment_Length",
                     "Mismatches", "Gap_Openings", "Query_Start", "Query_End",
                     "Subject_Start", "Subject_End", "E-value", "Bit_Score"]
# Define thresholds
identity_threshold = 30   # Keep sequences with at least 30% identity
alignment_length_threshold = 50  # Keep sequences with at least 50 aligned amino acids
evalue_threshold = 1e-5   # Remove weak matches (higher E-value means lower significance)
bit_score_threshold = 50  # Keep strong alignments

# Apply filtering
filtered_df = blastx_df[
    (blastx_df["%_Identity"] >= identity_threshold) &
    (blastx_df["Alignment_Length"] >= alignment_length_threshold) &
    (blastx_df["E-value"] <= evalue_threshold) &
    (blastx_df["Bit_Score"] >= bit_score_threshold)
]

# Save filtered results
filtered_df.to_csv("filtered_results.txt", sep="\t", index=False)
```

شکل ۶.۲: فیلتر کردن.

| | Query_ID | Subject_ID | %_Identity | Alignment_Length | Mismatches | Gap_Openings | Query_Start | Query_End |
|---|---|---|---|---|---|---|---|---|
| 2 | k141_960819 | 11xylanase | 44.286 | 70 | 35 | 2 | 1 | 201 334 402 6.32e-13 52.0 |
| 3 | k141_3842622 | 11xylanase | 34.459 | 148 | 74 | 5 | 70 | 510 405 530 3.76e-23 86.3 |
| 4 | k141_3842622 | 9xylanase | 30.667 | 150 | 76 | 6 | 85 | 507 50 180 7.95e-21 78.6 |
| 5 | k141_3842622 | 4RecName: | 32.203 | 177 | 89 | 7 | 100 | 585 100 260 2.76e-20 77.4 |
| 6 | k141_3842622 | 1xylanase | 30.508 | 177 | 92 | 8 | 100 | 585 104 264 5.02e-18 71.2 |
| 7 | k141_3842622 | 3xylanase | 30.337 | 178 | 91 | 9 | 100 | 585 104 264 8.06e-18 70.5 |
| 8 | k141_3522370 | 9xylanase | 40.476 | 336 | 176 | 6 | 34814 | 35797 7 326 4.57e-76 249.0 |
| 9 | k141_3522370 | 11xylanase | 41.176 | 357 | 188 | 7 | 34778 | 35830 344 684 6.65e-76 261.0 |
| 10 | k141_3522370 | 4RecName: | 39.498 | 319 | 168 | 4 | 34817 | 35737 48 353 2.45e-66 224.0 |
| 11 | k141_3522370 | 3xylanase | 36.723 | 354 | 187 | 8 | 34817 | 35809 53 392 1.95e-65 221.0 |
| 12 | k141_3522370 | 6xylanase | 37.618 | 319 | 174 | 4 | 34817 | 35737 48 353 4.43e-65 220.0 |
| 13 | k141_3522370 | 1xylanase | 36.757 | 370 | 182 | 8 | 34814 | 35809 52 407 3.14e-64 218.0 |

شکل ۷.۲: filtered_results.txt

```
>k141_960819
ATTGATGATTTTGAAATAAAAGGTGATTCTAGTACTGTTGATGATTCTGTACCTGCTTTAAAAGATAAATATAAAAGTCATTTTAAATTTGGAACAGCTACAGTTGTAGAT
>k141_3842622
GATGCTTTGCTGGGAGGCTCAGCACGACTGATTTCAGTTTCTCAGGCAAAGAATCGCGGATATTATATCCCAGACAATTATCGTGAGTCTTATGTACCTGAAATCAACTTC
>k141_3522370
CTAATCCCAATTGAAGCACTTTAGATGCAGGTTCCACAAAACCATCTCTGTTTTGTACCACATCAATCAACACATTTGCATCAAGGAAAATCTTTGTTGCCATCACTTACA
>k141_343
CTACCGGCAGTCGGTGATGTACAAGCTCTGCGGCGACGAGTTCATCGCCAAGGCCTTTGAGTATGCCCACGCCGCCGACCCCAACGCGCTGCTGTTCTACAACGACTACAA
>k141_4162874
GATATAGACGTCGCCCACGTAGTTGGGAACGTGCAGCAGCAGGAAACTGCTTGCCATCGTGGGCTTGACGTCCACGCTCACGTCCTGCCATTGGTCGTTGAAGGAGAAGCT
>k141_4162895
GATATAGACGTCGCCCACGTAGTTGGGAACGTGCAGCAGCAGGAAACTGCTTGCCATCGTGGGCTTGACGTCCACGCTCACGTCCTGCCATTGGTCGTTGAAGGAGAAGCT
>k141_2561960
CAAGCATGGTATCTCTGTCTGCAAGAGGGAAGTTCTCATTGTAATGCTGACAGAAGAACCACTTTGGGGTCTGGTTGTGCCACACAAGGGTATGTCCGCGCATTGCGATAC
>k141_4163142
GTCCCAGGCATATACCACATCCTTATAGCGGTTTACCACCGTGTGGATATGCTCGCGCAGACGCTGGTAGAACACTTCCTTCTTTACCTCCTTGCCTTTCTTGTCTGTGAA
```

شکل ۸.۲: filtered_contigs.txt

آنچه BLASTX در واقع خروجی می دهد بخش منطبق از پیوند با یک پروتئین هماهنگ است. توالی کامل ترجمه شده contig را بر نمی گرداند. در این قسمت توالی کامل ترجمه کانتیگ را از پایگاه داده اصلی پیدا کرده و خروجی می‌دهیم.

**ترجمه کردن:**

**Extract Nucleotide Sequences from Dataset**

```
# Load the assembled contigs file (nucleotide sequences)
contig_file = "y5.final.contigs.fa"  # Change this to your actual filename

# Read the nucleotide sequences and store only the filtered ones
filtered_sequences = {}

for record in SeqIO.parse(contig_file, "fasta"):
    if record.id in filtered_contigs:  # Keep only sequences that passed filtering
        filtered_sequences[record.id] = record.seq

# Save extracted sequences to a new FASTA file
with open("filtered_contigs.fasta", "w") as output_fasta:
    for contig_id, seq in filtered_sequences.items():
        output_fasta.write(f">{contig_id}\n{seq}\n")
```

**Translate Nucleotide Sequences to Protein**

```
# Translate the filtered nucleotide sequences
translated_sequences = {}

for contig_id, seq in filtered_sequences.items():
    translated_seq = Seq(str(seq)).translate(to_stop=True)  # Stop at first stop codon
    translated_sequences[contig_id] = translated_seq

# Save translated sequences to a FASTA file
with open("translated_proteins.fasta", "w") as output_fasta:
    for contig_id, protein_seq in translated_sequences.items():
        output_fasta.write(f">{contig_id}\n{protein_seq}\n")
```

شکل ۹.۲: ترجمه کردن.

**ترجمه با استفاده از کامند ترمینال:**

```
# Extract potential contig sequences
!seqtk subseq y5.final.contigs.fa potential_xylanase_contigs.txt > potential_xylanase_contigs.fa

# Translate contigs into protein sequences
!transeq potential_xylanase_contigs.fa -outseq potential_xylanase_proteins.fa
```

شکل ۱۰.۲: ترجمه با استفاده از کامند ترمینال.

### ۱.۱.۲ تجزیه و تحلیل نتایج و نمودار های BLAST

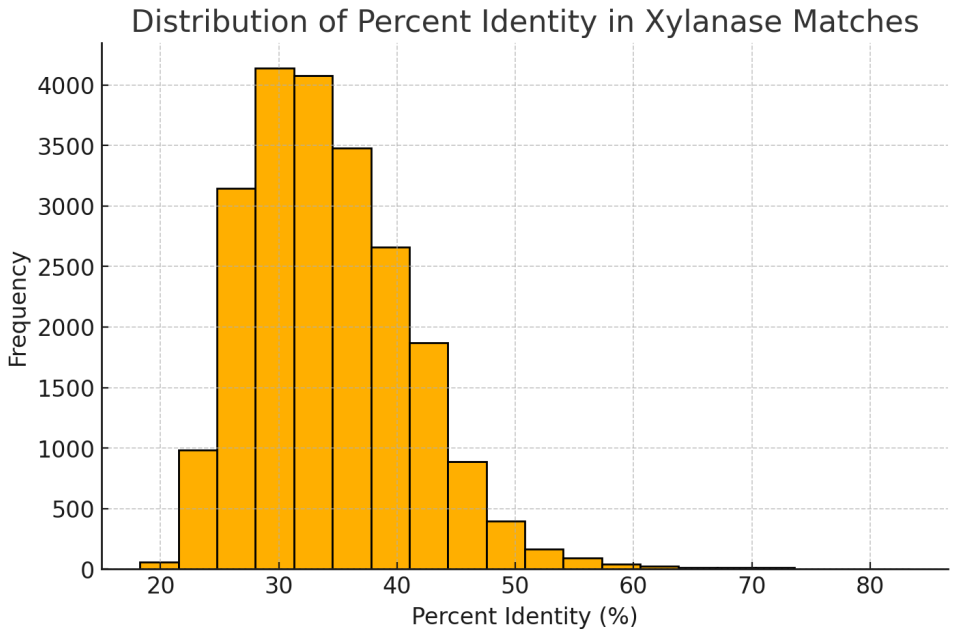
#### ۱. کلیت نتایج BLAST

تجزیه و تحلیل BLAST چندین پیوند را شناسایی کرد که با آنزیم های زایلاناز شناخته شده با درجات مختلف شباهت مطابقت داشتند. پارامترهای کلیدی مورد تجزیه و تحلیل عبارتند از:

- درصد هویت: شباهت بین پرس و جو و دنباله موضوع را اندازه می گیرد.

- امتیاز بیت: نمرات بالاتر نشان دهنده تراز قوی تر است.
- E-value: نشان دهنده اهمیت آماری است. مقادیر پایین تر نشان دهنده تطابق قابل اعتمادتر است.

**۲. تفسیر نمودار ها**



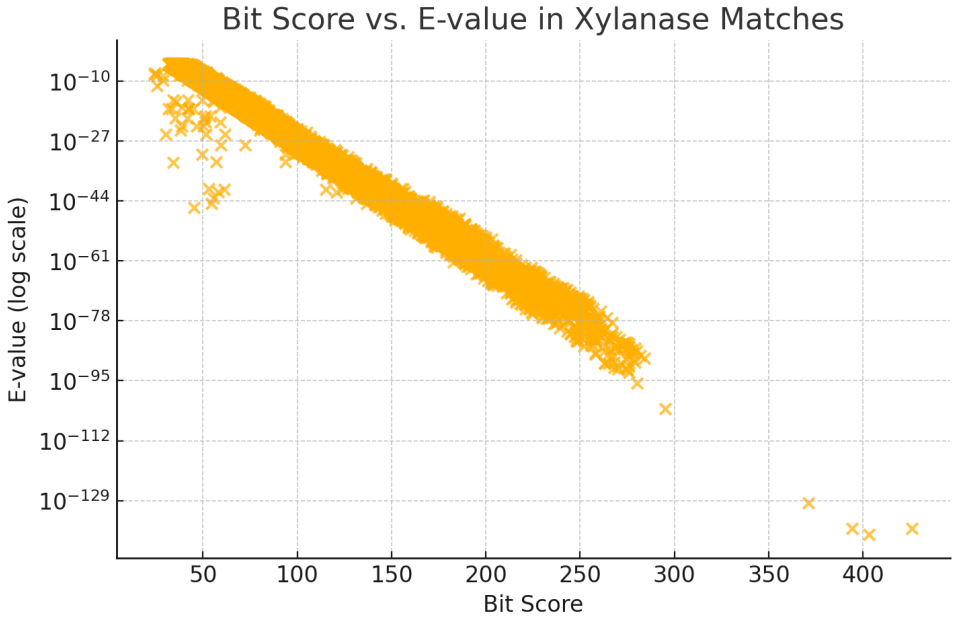

شکل ۱۱.۲: نمودار هیستوگرام درصد توزیع هویت.

آ- هیستوگرام: درصد توزیع هویت
مشاهدات:

- توزیع طیف وسیعی از مقادیر هویت درصد را نشان می دهد.
- اکثر تطابق ها بین ۳۰% و ۵۰% هویت قرار می گیرند، که نشان می دهد برخی از توالی های شناسایی شده ممکن است از فاصله دور با زایلانازهای شناخته شده مرتبط باشند.
- بخش کوچکتر دارای درصد هویت بالاتر (> ۵۰%) است که نشان دهنده روابط تکاملی قوی تر با زایلانازهای شناخته شده است.

مفاهیم:

- توالی های با هویت بالا (۵۰-۶۰%) احتمالا زایلانازهای کاربردی با خواص بیوشیمیایی مشابه با آنزیم های شناخته شده هستند.
- توالی های با هویت پایین تر (۳۰-۴۰%) ممکن است گونه های جدید زایلاناز را با پتانسیل برای کاربردهای بیوتکنولوژیکی نشان دهند.
- ممکن است برای تأیید فعالیت در توالی‌های با هویت پایین‌تر، تحلیل دامنه عملکردی بیشتری مورد نیاز باشد.



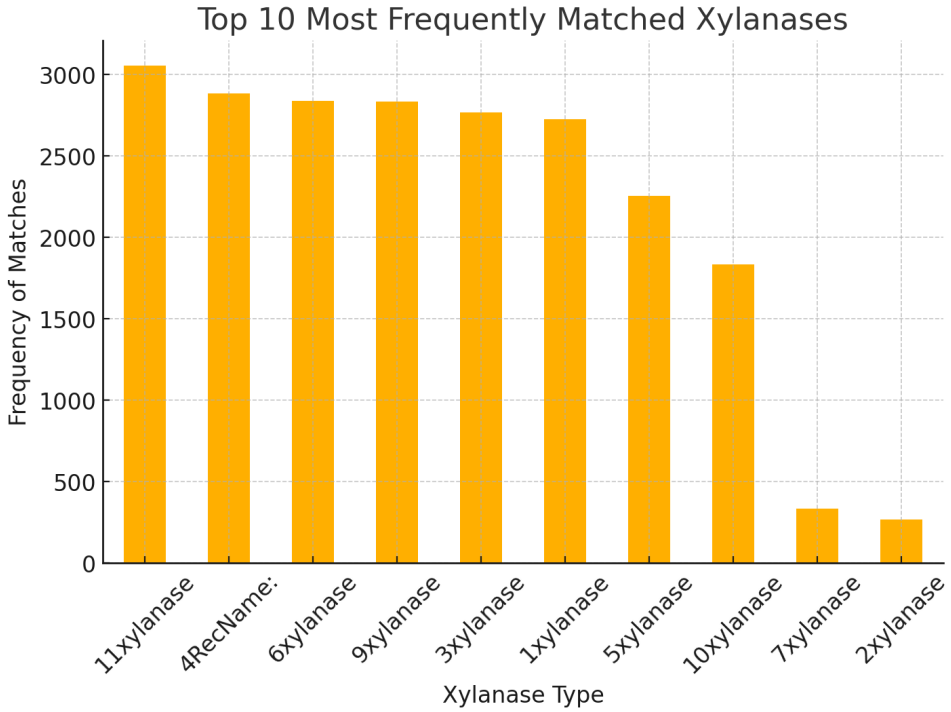

شکل ۱۲.۲: نمودار پراکندگی: امتیاز بیت در مقابل e-value

ب- نمودار پراکندگی: امتیاز بیت در مقابل e-value
مشاهدات:

- امتیاز بیت بالا با مقادیر E کمتر مطابقت دارد، که تطابق قوی و معنی دار آماری را تایید می کند.
- برخی از توالی‌ها امتیاز بیت‌های متوسطی را نشان می‌دهند، اما همچنان دارای مقادیر E پایین هستند، به این معنی که تا حدی با زایلانازهای شناخته شده مطابقت دارند، اما ممکن است انواع متفاوتی باشند.

مفاهیم:

- امتیاز بیت بالا و ارزش E پایین ← کاندیدهای قوی زایلاناز ارزش بررسی بیشتر را دارند.
- امتیاز بیت متوسط و ارزش E پایین ← آنزیم های جدید بالقوه با شباهت جزئی به زایلانازهای شناخته شده.




شکل ۱۳.۲: نمودار میله ای: ۱۰ زایلاناز برتر که بیشترین تطبیق را دارند

ج- نمودار میله ای: ۱۰ زایلاناز برتر که بیشترین تطبیق را دارند
مشاهدات:

- برخی از آنزیم‌های زایلاناز بیشتر در چند شاخه ظاهر می‌شوند، که نشان می‌دهد در متاژنوم شکمبه فراوان هستند.
- بیشترین تطابق زایلانازها احتمالاً متعلق به خانواده های آنزیمی بسیار حفاظت شده در میکروبیوم شکمبه است.

مفاهیم:

- انواع زایلاناز غالب ممکن است از نظر عملکردی در تخریب لیگنوسلولز در شکمبه مهم باشند.
- زایلانازهایی که کمتر مطابقت دارند ممکن است آنزیم های کمیاب یا جدید باشند که ارزش توصیف بیشتر را دارند.

هیستوگرام توزیع درصد هویت (شکل ۱) نشان می دهد که اکثر کانتیگ ها ۳۰-۶۰ درصد هویت با زایلانازهای شناخته شده دارند. این نشان می دهد که در حالی که برخی از نامزدها ارتباط نزدیکی با زایلانازهای مرجع دارند، برخی دیگر ممکن است انواع جدیدی را ارائه دهند. نمودار پراکندگی بیت امتیاز در مقابل ارزش E (شکل ۲) نشان می دهد که امتیاز بیت بالاتر با مقادیر E پایین تر همبستگی دارد و تأیید می کند که این توالی ها از نظر آماری با زایلانازهای شناخته شده مطابقت دارند.
از طریق جستجوهای مبتنی بر شباهت، مجموعه‌ای از توالی‌های کاندید زایلاناز را از متاژنوم شکمبه شناسایی کردیم. این توالی ها بر اساس درصد هویت، طول هم ترازی و اهمیت آماری فیلتر شدند تا از انتخاب قابل اعتماد اطمینان حاصل شود. مرحله بعدی شامل خوشه‌بندی این توالی‌ها برای حذف افزونگی و انتخاب توالی‌های نماینده برای مدل‌سازی منطقه حفاظت‌شده است. این به ما این امکان را می دهد که انتخاب کاندیدهای زایلاناز پایدار در برابر حرارت را برای کاربردهای صنعتی اصلاح کنیم.

# بخش ۳

# گام ۲: خوشه بندی و انتخاب توالی نماینده

پس از شناسایی توالی‌های بالقوه زایلاناز در مرحله ۱، به مرحله ۲ می‌رویم، که شامل خوشه‌بندی توالی‌های بسیار مشابه برای کاهش افزونگی و انتخاب توالی‌های نماینده از هر خوشه است. این فرآیند تضمین می‌کند که تجزیه و تحلیل‌های بعدی از نظر محاسباتی کارآمد هستند و به جای چندین نسخه اضافی از یک ژن، روی توالی‌های عملکردی متنوع متمرکز هستند.

خوشه بندی برای مطالعات متاژنومی ضروری است زیرا متاژنوم ها اغلب دارای گونه های ژنی بسیار مشابه هستند. با اعمال الگوریتم های خوشه بندی مانند CD-HIT، می توانیم:

- کاهش بار محاسباتی برای تجزیه و تحلیل پایین دست.
- اطمینان حاصل کردن از این که هیچ گونه خاص زایلاناز را بیش از حد نشان نمی دهیم.
- با تمرکز بر انواع توالی متمایز، پیش بینی های عملکردی را بهبود میبخشیم.

در این مرحله، از CD-HIT، یک ابزار خوشه‌بندی پرکاربرد، برای گروه‌بندی توالی‌ها بر اساس آستانه شباهت استفاده می‌کنیم. توالی های نماینده به دست آمده به عنوان یک مجموعه داده غیر زائد برای تجزیه و تحلیل ساختاری و عملکردی بیشتر عمل می کنند.

## ابزار مورد استفاده: CD-HIT برای خوشه بندی بر اساس آستانه تشابه

CD-HIT یک الگوریتم خوشه بندی پرکاربرد برای کاهش افزونگی در مجموعه داده های توالی بزرگ است. توالی‌هایی را گروه‌بندی می‌کند که درصد مشخصی از شباهت را به اشتراک می‌گذارند، و تنها یک دنباله نماینده برای هر خوشه حفظ می‌کند. برای این پروژه، CD-HIT به صورت زیر پیکربندی شد:

- آستانه تشابه (۰.۹۷): توالی هایی با شباهت ۹۷ درصد یا بیشتر در یک خوشه گروه بندی شدند.
- اندازه کلمه (۵): اندازه کلمه ۵ برای خوشه بندی پروتئین، تعادل سرعت و دقت استفاده شد.
- توضیحات (۰): فقط اطلاعات توالی ضروری را در خروجی حفظ می کند.

## ۱.۳ دستورات ترمینال برای خوشه بندی با CD-HIT

۱. تصب :CD-HIT

```
!conda install -c bioconda seqtk -y
Executed at 2025.02.07 23:59:42 in 1m 38s 489ms
     ------------------------------------------------------------

                                       Total:          43 KB

 The following NEW packages will be INSTALLED:

   seqtk              bioconda/linux-64::seqtk-1.4-he4a0461_2



 Downloading and Extracting Packages:

 Preparing transaction: done
 Verifying transaction: done
 Executing transaction: done
```

شکل ۱.۳: تصب CD-HIT

۲. اجرای CD-HIT را روی توالی های پروتئین ترجمه شده:
ما توالی‌های پروتئین ترجمه‌شده را با استفاده از CD-HIT با آستانه شباهت ۹۷ درصد خوشه‌بندی کردیم. از دستور زیر استفاده شد:

```
!cd-hit -i results/translated_proteins.fasta -o clustered_xylanase.fasta -c 0.97 -n 5 -d 0
Executed at 2025.02.07 21:19:52 in 308ms
 ================================================================
 Program: CD-HIT, V4.8.1 (+OpenMP), Nov 12 2024, 10:35:24
 Command: cd-hit -i results/translated_proteins.fasta -o
          clustered_xylanase.fasta -c 0.97 -n 5 -d 0

 Started: Fri Feb  7 21:19:52 2025
 ================================================================
                             Output
 ----------------------------------------------------------------
 total seq: 1817
 longest and shortest : 2358 and 11
 Total letters: 168514
 Sequences have been sorted
```

شکل ۲.۳: اجرای CD-HIT را روی توالی های پروتئین ترجمه شده

توضیح پارامترها:

- -i ← فایل FASTA را وارد کنید (از مرحله ۱)
- -o ← فایل خروجی حاوی توالی های خوشه ای
- -c 0.97 ← آستانه خوشه بندی (۹۷% هویت) برای پروتئین ها
- -n 5 ← اندازه کلمه برای خوشه بندی (توصیه شده برای پروتئین ها)
- -d 0 ← هیچ توضیحات اضافی در خروجی وجود ندارد

natural for account to (۸۵%) threshold similarity lower a at sequences nucleotide clustered also We sequences DNA in variations

۳. مشاهده چند خوشه اول

```
!head -n 20 clustered_xylanase.fasta.clstr
Executed at 2025.02.07 21:20:13 in 237ms

>Cluster 0
0   2358aa, >k141_2357594... *
>Cluster 1
0   1160aa, >k141_4340043... *
>Cluster 2
0   956aa, >k141_966364... *
>Cluster 3
0   956aa, >k141_1778842... *
>Cluster 4
0   836aa, >k141_118972... *
>Cluster 5
0   794aa, >k141_6564741... *
>Cluster 6
0   775aa, >k141_3934590... *
```

شکل ۳.۳: مشاهده چند خوشه اول

### انتخاب توالی های نماینده

هنگامی که توالی ها خوشه می شوند، باید یک توالی نماینده برای هر خوشه انتخاب شود. در بیشتر موارد، طولانی ترین دنباله در هر خوشه برای حفظ مرتبط ترین اطلاعات بیولوژیکی انتخاب می شود.

```
!grep '>' step2/clustered_contigs.fasta | cut -d' ' -f1 | sed 's/>//' > step2/representative_contigs.txt
!seqkit grep -f step2/representative_contigs.txt step1/filtered_contigs.fasta > step2/representative_contigs.fasta
Executed at 2025.02.08 02:23:07 in 566ms

[INFO] 1948 patterns loaded from file
```

شکل ۴.۳: انتخاب توالی های نماینده

## تفسیر نتایج

خوشه‌بندی با موفقیت توالی‌های اضافی را حذف کرد و مجموعه داده‌ها را از ۲۶۵۹ دنباله به ۱۹۴۸ خوشه کاهش داد. بیشتر خوشه ها حاوی تنها ۱–۵ توالی هستند که نشان دهنده تنوع ژنتیکی بالا در بین ژن های متاژنومی زایلاناز است. نمودار پراکندگی تایید می کند که توزیع طول دنباله تا حد زیادی بدون تغییر باقی می ماند. توالی های نماینده یک مجموعه غیر زائد برای تجزیه و تحلیل بیشتر فراهم می کنند.

از طریق مرحله ۲، ما با موفقیت ۲۶۵۹ توالی زایلاناز متاژنومیک را خوشه بندی کردیم و ضمن حفظ تنوع، افزونگی را کاهش دادیم. توالی های نماینده انتخاب شده از هر خوشه در مرحله ۳ برای ویرایش عملکردی بیشتر و تجزیه و تحلیل دامنه حفاظت شده استفاده خواهند شد.

# بخش ۴

# گام ۳: مدل‌سازی ناحیه حفاظت شده و فیلتر کردن توالی‌ها

در این مرحله هدف ساخت یک مدل برای ناحیه حفاظت شده طولانی‌ترین توالی زیر خانواده thermostable از زاینالازها است.

## ۱.۴ دیدگاه کلی

در فرآیند شناسایی آنزیم‌های زایلاناز کاربردی از مجموعه داده‌های متاژنومی، مرحله ۳ نقش مهمی در پالایش و اعتبارسنجی توالی‌های شناسایی‌شده در مراحل قبلی دارد. هدف کلی این مرحله تجزیه و تحلیل مناطق حفاظت شده در توالی های خوشه ای و فیلتر کردن نامزدهای کمتر قابل اعتماد است و اطمینان حاصل می کند که فقط مرتبط ترین توالی ها برای تجزیه و تحلیل پایین دست باقی می مانند. این مرحله بر تکنیک‌های محاسباتی پیشرفته، از جمله ماتریس‌های امتیازدهی خاص موقعیت (PSSM) و مدل‌های مارکوف پنهان ،(HMMs) برای شناسایی موتیف‌های حفاظت‌شده، امتیاز مربوط به ترتیب و حذف توالی‌های اضافی یا کم‌اعتماد متکی است.

در زمینه مطالعات متاژنومیک، توالی‌های بیولوژیکی بازیابی شده از نمونه‌های محیطی اغلب دارای تغییرات قابل‌توجهی به دلیل جهش، واگرایی تکاملی و خطاهای توالی هستند. با این حال، پروتئین های حیاتی عملکردی، مانند آنزیم های دخیل در تخریب زیست توده، معمولاً باقیمانده های بسیار حفاظت شده را در مناطق کاتالیزوری و اتصال به بستر خود حفظ می کنند. این حوزه های حفاظت شده برای عملکرد مناسب آنزیم اساسی هستند، زیرا یکپارچگی ساختاری و کارایی کاتالیزوری را تضمین می کنند. بنابراین، مدل‌سازی این نواحی و فیلتر کردن توالی‌هایی که حاوی موتیف‌های به خوبی محافظت‌شده نیستند، برای به حداکثر رساندن احتمال شناسایی زایلانازهای کاربردی واقعی ضروری است.

یکی دیگر از جنبه های مهم مرحله ۳، استفاده از فیلترینگ توالی برای حذف توالی هایی است که معیارهای حفاظت را برآورده نمی کنند. با پالایش مجموعه داده، این فرآیند دقت تجزیه و تحلیل های پایین دستی مانند حاشیه نویسی عملکردی، پیش بینی ساختار پروتئین و خصوصیات بیوشیمیایی را بهبود می بخشد. بدون این فیلتر، خطر بیشتری برای انتشار توالی های اشتباه وجود دارد که می تواند تلاش های اعتبارسنجی آزمایشی بعدی را به خطر بیندازد. بنابراین، این مرحله به عنوان یک نقطه بازرسی کنترل کیفیت عمل می‌کند و قابلیت اطمینان توالی‌های کاندید زایلاناز را برای بررسی بیشتر تقویت می‌کند.

با استفاده از PSI-BLAST برای ایجاد مدل‌های PSSM و HMMER برای تولید مدل‌های مارکوف پنهان، این مرحله چارچوبی قدرتمند برای تشخیص الگوهای حفاظت از توالی فراهم می‌کند. این رویکردهای محاسباتی به طور گسترده در بیوانفورماتیک برای شناسایی همولوگ های راه دور، پیش بینی مکان های عملکردی و افزایش درک ما از تکامل آنزیم استفاده می شود. نتایج این مرحله نه تنها مجموعه داده‌ها را اصلاح می‌کند، بلکه بینشی در مورد چگونگی تکامل آنزیم‌های زایلاناز و سازگاری با شرایط مختلف محیطی، به‌ویژه در جوامع میکروبی گرمادوست، ارائه می‌کند.

## ۲.۴ زمینه علمی: دامنه های حفاظت شده و نقش آنها در عملکرد زایلاناز

دامنه‌های حفاظت‌شده نواحی خاصی در توالی‌های پروتئینی هستند که به دلیل نقش اساسی در عملکرد آنزیمی و پایداری ساختاری، در بین گونه‌های مختلف بسیار حفظ می‌شوند. این دامنه ها اغلب حاوی باقیمانده هایی هستند که برای فعالیت کاتالیزوری، اتصال به بستر یا برهمکنش های پروتئین-پروتئین ضروری هستند. وجود دامنه‌های حفاظت‌شده در یک توالی پروتئین نشان می‌دهد که پروتئین نقش عملکردی خود را در طول تکامل حفظ کرده است و آن را به یک کاندیدای قوی برای مطالعه بیشتر تبدیل می‌کند.

برای آنزیم‌های زایلاناز، حوزه‌های حفاظت‌شده از اهمیت ویژه‌ای برخوردار هستند، زیرا مکانیسم هیدرولیز زایلان را تعریف می‌کنند. زایلانازها به خانواده گلیکوزید هیدرولاز (GH) تعلق دارند که اکثر اعضای مشخصه آن در میان سایرین تحت GH۱۰ و GH۱۱ قرار دارند. این آنزیم ها تجزیه زایلان، جزء اصلی همی سلولز گیاهی را با شکستن پیوندهای بتا-۱،۴-گلیکوزیدی کاتالیز می کنند. عملکرد کاتالیزوری زایلاناز به شدت به بقایای حفاظت شده خاص، از جمله باقی مانده های اسیدی (اسید آسپارتیک و اسید گلوتامیک) وابسته است که به عنوان دهنده پروتون و نوکلئوفیل در طول واکنش برش آنزیمی عمل می کنند.

ساختار زایلانازها بسته به خانواده GH اغلب شامل یک دامنه کاتالیزوری با یک چین آلفا/بتا بشکه یا یک چین بتا-ژلیلول است. این نقوش ساختاری برای شناسایی و کاتالیز سوبسترا بسیار مهم هستند، به این معنی که تغییرات در این مناطق می تواند به شدت فعالیت آنزیم را تغییر دهد. با تجزیه و تحلیل دامنه های حفاظت شده در توالی های زایلاناز، پیش بینی عملکرد آنزیمی حتی در پروتئین های تازه کشف شده یا قبلاً مشخص نشده ممکن می شود. علاوه بر این، وجود ماژول‌های اتصال کربوهیدرات اضافی (CBM) می‌تواند ویژگی سوبسترا را افزایش دهد و بر عملکرد آنزیم در کاربردهای صنعتی تأثیر بگذارد.

با توجه به اینکه زایلانازها به طور گسترده در تولید سوخت زیستی، صنعت کاغذ، خوراک دام و فرآوری مواد غذایی استفاده می‌شوند، شناسایی انواع بسیار پایدار و کارآمد ضروری است. بسیاری از زایلانازهای طبیعی برای شرایط محیطی خاص، مانند دماهای بالا، pH شدید، یا تحمل نمک بهینه شده اند. از طریق مدل‌سازی منطقه حفاظت‌شده، محققان می‌توانند سازگاری‌های ساختاری را مشخص کنند که به زایلانازهای خاصی اجازه می‌دهد در شرایط شدید عمل کنند و آنها را کاندیدای عالی برای کاربردهای بیوتکنولوژیکی می‌کند.

اهمیت تجزیه و تحلیل دامنه حفاظت شده فراتر از مقایسه توالی ساده است. با شناسایی الگوهای حفاظت، محققان می توانند روابط تکاملی را ردیابی کنند، ویژگی بالقوه بستر را استنباط کنند، و حتی استراتژی های مهندسی آنزیم را برای افزایش خواص کاتالیزوری طراحی کنند. شناسایی و توصیف مناطق حفاظت‌شده، کشف انواع زایلاناز جدید با پایداری و کارایی بهبود یافته را امکان‌پذیر می‌سازد، که باعث پیشرفت در بیوتکنولوژی صنعتی و زیست‌شناسی مصنوعی می‌شود.

**نواحی حفاظت‌شده** در زایلانازها به نواحی از توالی پروتئینی گفته می‌شود که در طی تکامل به‌طور نسبتاً ثابت باقی مانده‌اند و تغییرات کمی در آن‌ها رخ داده است. این نواحی معمولاً عملکردهای ضروری آنزیم، مانند سایت‌های فعال و ساختارهای فضایی آنزیم را نگه می‌دارند و در نتیجه برای عملکرد صحیح آنزیم ضروری هستند.

در پروژه‌های زیستی و بایوانفورماتیک، انتخاب روش مناسب برای مدل‌سازی توالی‌ها و نواحی حفاظتی نقش بسیار مهمی در دقت و کارایی نتایج دارد. در این بخش، به هر دو روش پرکاربرد در مدل‌سازی توالی‌ها، یعنی ماتریس امتیازدهی موقعیت-ویژه و مدل‌های (PSSM) مخفی مارکوف ،(HMM) پرداخته خواهد شد.

## ۳.۴ Omega Clustal Using (MSA) Alignment Sequence Multiple

### ۱.۳.۴ هدف MSA

تراز چند توالی (MSA) یک تکنیک اساسی بیوانفورماتیک است که برای هم‌ترازی مجموعه‌ای از توالی‌های بیولوژیکی برای شناسایی مناطق مشابه استفاده می‌شود. این شباهت ها اغلب روابط ساختاری، عملکردی یا تکاملی بین دنباله ها را نشان می دهد. در زمینه توالی های پروتئین زیلاناز، MSA برای تشخیص باقی مانده های بسیار حفاظت شده که احتمالا برای عملکرد آنزیمی حیاتی هستند ضروری است.

هدف کلیدی انجام MSA در این مرحله آنالیز نواحی حفاظت‌شده در میان توالی‌های نماینده به‌دست‌آمده از مرحله ۲ است. از آنجایی که زایلانازها متعلق به خانواده‌های گلیکوزید هیدرولاز با مشخصه‌های خوبی هستند (مانند GH۱۰، GH۱۱)، حوزه‌های کاتالیزوری، محل‌های اتصال و نقوش ساختاری آن‌ها باید در گونه‌های مختلف محافظت شوند. با تراز کردن این توالی‌ها، می‌توانیم بقایای حیاتی را که برای فعالیت آنزیمی و ثبات ساختاری ضروری هستند، مشخص کنیم.

علاوه بر این، MSA امکان پیش‌بینی عملکردی بر اساس حفظ توالی را فراهم می‌کند. باقیمانده‌های بسیار حفاظت‌شده اغلب با بقایای کاتالیزوری، نقوش اتصال به بستر یا مکان‌های تثبیت ساختاری مطابقت دارند. اگر یک باقیمانده خاص به شدت در چندین توالی حفظ شود، به شدت نقش عملکردی را نشان می دهد. برعکس، مناطق متغیر ممکن است سازگاری هایی را نشان دهند که به آنزیم های زایلاناز اجازه می دهد در شرایط محیطی مختلف عمل کنند.

انجام MSA همچنین ایجاد مدل‌های آماری مانند ماتریس‌های امتیازدهی خاص موقعیت (PSSM) و مدل‌های پنهان مارکوف (HMMs) را ممکن می‌سازد، که در مراحل بعدی برای اصلاح فیلترینگ توالی استفاده می‌شوند. این مدل‌ها به تشخیص تغییرات دنباله‌ای ظریف و در عین حال حفظ نقوش مرتبط بیولوژیکی کمک می‌کنند. بدون MSA مناسب، فرآیندهای پایین دستی مانند تشخیص دامنه حفاظت شده، پیش بینی ساختار و حاشیه نویسی عملکردی به طور قابل توجهی کمتر قابل اعتماد خواهند بود.

بنابراین، MSA به عنوان یک مرحله پیش پردازش حیاتی عمل می کند که دقت پیش بینی های عملکردی مبتنی بر توالی را افزایش می دهد. این تضمین می‌کند که فقط توالی‌های مرتبط با بیولوژیک به مراحل بعدی تجزیه و تحلیل می‌روند و در عین حال توالی‌های نامرتب، اضافی یا غیرعملکردی را حذف می‌کنند.

### اجرای Clustal Omega

برای انجام MSA، ما از Clustal Omega، یک ابزار پرکاربرد MSA که به دلیل کارایی و دقت آن شناخته شده است، استفاده می کنیم.

```
!clustalo -i step2/clustered_proteins.fasta -o step3/aligned_proteins.fasta --auto -v

Executed at 2025.02.08 03:36:53 in 1m 7s 564ms

Using 12 threads
Read 1416 sequences (type: Protein) from step2/clustered_proteins.fasta
Setting options automatically based on input sequence characteristics (might overwrite some of your options).
Using 109 seeds (chosen with constant stride from length sorted seqs) for mBed (from a total of 1416 sequences)
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress done. CPU time: 5.47u 0.01s 00:00:05.48 Elapsed: 00:00:02
mBed created 24 cluster/s (with a minimum of 1 and a soft maximum of 100 sequences each)
Distance calculation within sub-clusters done. CPU time: 5.67u 0.01s 00:00:05.68 Elapsed: 00:00:01
Guide-tree computation (mBed) done.
Progressive alignment progress done. CPU time: 590.50u 2.11s 00:09:52.61 Elapsed: 00:01:03
Alignment written to step3/aligned_proteins.fasta
```

شکل ۱.۴: اجرای Clustal Omega

فایل ورودی، step۲/clustered_proteins.fasta، حاوی توالی های نماینده استخراج شده از خوشه بندی CD-HIT است. از آنجایی که خوشه‌بندی باعث کاهش افزونگی در مرحله ۲ شد، توالی‌های ارائه‌شده به Omega Clustal از قبل غیراضافی بودند و فرآیند هم‌ترازی را کارآمدتر و دقیق‌تر می‌کردند.

پس از اجرای Omega، Clustal توالی های تراز شده در step۳/aligned_proteins.fasta ذخیره شدند. این فایل به عنوان ورودی حیاتی برای مدل‌سازی توالی بیشتر، از جمله ایجاد PSSM (PSI-BLAST) و تولید HMM (HMMER) عمل می‌کند.

پس از اجرا، Omega Clustal با موفقیت ۱۹۴۸ توالی پروتئین را که مربوط به تعداد خوشه های تولید شده در مرحله ۲ است، تراز کرد. هم ترازی تقریباً در ۵ دقیقه تکمیل شد و کارایی Omega Clustal را در مدیریت مجموعه داده های بزرگ نشان داد.

## ۴.۴ ایجاد یک پایگاه داده BLAST برای جستجوی منطقه حفاظت شده

### چرا یک پایگاه داده BLAST ایجاد کنیم؟

یکی از مراحل حیاتی در شناسایی مناطق حفاظت شده در توالی های زایلاناز، ایجاد پایگاه داده BLAST است که جستجو و مقایسه توالی کارآمد را تسهیل می کند. هدف اولیه از ساخت پایگاه داده BLAST، ذخیره توالی های خوشه ای زایلاناز در قالبی است که امکان جستجوی شباهت سریع با استفاده از PSI-BLAST و سایر ابزارهای مبتنی بر BLAST را فراهم کند. این فرآیند برای شناسایی توالی‌های همولوگ، شناسایی حفاظت تکاملی، و فیلتر کردن توالی‌ها بر اساس ارتباط عملکردی آنها بسیار مهم است.

در مطالعات متاژنومی، مجموعه داده ها اغلب حاوی هزاران دنباله با درجات مختلف شباهت هستند. روش‌های هم‌ترازی توالی زوجی سنتی مانند Omega Clustal برای مجموعه داده‌های کوچک مؤثر هستند، اما وقتی با مجموعه داده‌های متاژنومی در مقیاس بزرگ سروکار دارند، از نظر محاسباتی گران می‌شوند. یک پایگاه داده BLAST با ارائه یک فضای جستجوی از پیش نمایه شده، که امکان مقایسه سریع و مقیاس پذیر توالی را فراهم می کند، بر این محدودیت غلبه می کند. با ساختاردهی توالی‌های زایلاناز خوشه‌ای در یک پایگاه داده BLAST، می‌توانیم به طور موثر توالی‌های جدید را در برابر مجموعه داده‌های موجود پرس‌وجو کنیم تا تعیین کنیم که چقدر با انواع زایلاناز شناخته شده مطابقت دارند.

یکی دیگر از مزایای کلیدی ایجاد پایگاه داده BLAST این است که تجزیه و تحلیل منطقه حفاظت شده را امکان پذیر می کند. از آنجایی که باقیمانده‌های مهم عملکردی در بین توالی‌های همولوگ بسیار حفظ می‌شوند، استفاده از پایگاه داده BLAST به ما امکان می‌دهد تا برای نقوش حفاظت‌شده در تمام پروتئین‌های زایلاناز شناسایی‌شده جستجو کنیم. این به ویژه در تولید ماتریس امتیازدهی ویژه موقعیت (PSSM) مفید است، جایی که توالی‌هایی که با مناطق حفاظت‌شده با اطمینان آماری بالا مطابقت دارند، حفظ می‌شوند، در حالی که توالی‌های با اعتماد پایین فیلتر می‌شوند.

به طور خلاصه، یک پایگاه داده BLAST به عنوان یک مخزن متمرکز از توالی‌های زایلاناز خوشه‌ای عمل می‌کند، که امکان جستجوی سریع تشابه، تشخیص منطقه حفاظت‌شده، و فیلتر کردن توالی با اطمینان بالا را فراهم می‌کند. بدون پایگاه داده ،BLAST مقایسه‌های توالی به طور قابل‌توجهی کندتر و مقیاس‌پذیرتر خواهند بود، و شناسایی مناطق مهم عملکردی در مجموعه داده‌های متاژنومی بزرگ را به چالش می‌کشد.

### دستور MakeBLASTDB

برای ایجاد یک پایگاه داده BLAST از توالی های خوشه ای زایلاناز، از دستور makeblastdb استفاده می کنیم که بخشی از مجموعه BLAST+ NCBI است. این دستور یک فایل FASTA را به یک پایگاه داده سازگار با BLAST تبدیل می کند و امکان جستجوی توالی کارآمد را فراهم می کند.

```
!makeblastdb -in step2/clustered_proteins.fasta -dbtype prot -out xylanase_db
Executed at 2025.02.08 03:36:58 in 365ms


Building a new DB, current time: 02/08/2025 03:36:58
New DB name:   /home/amir/Documents/university/Semester9/bioinformatics/project2/xylanase_db
New DB title:  step2/clustered_proteins.fasta
Sequence type: Protein
Deleted existing Protein BLAST database named /home/amir/Documents/university/Semester9/bioinformatics/project2/xylanase_db
Keep MBits: T
Maximum file size: 3000000000B
Adding sequences from FASTA; added 1416 sequences in 0.0223091 seconds.
```

شکل ۲.۴: اجرای blast make

ایجاد موفقیت آمیز پایگاه داده BLAST یک گام مهم به سمت تجزیه و تحلیل توالی زایلاناز با اطمینان بالا است. با ساختاربندی توالی‌های خوشه‌بندی‌شده در قالبی قابل جستجو، کارایی جستجوهای تشابه توالی، شناسایی دامنه حفظ‌شده و عملکردی را بهبود بخشیده‌ایم. پایگاه داده اکنون به عنوان یک منبع حیاتی برای پالایش انتخاب توالی، فیلتر کردن نامزدهای کم اعتماد، و اطمینان از اینکه فقط دنباله‌های مرتبط با عملکرد به مراحل بعدی ادامه می‌دهند، عمل می‌کند.
با حرکت رو به جلو، این پایگاه داده BLAST برای تکرارهای PSI-BLAST استفاده خواهد شد، که یک مدل PSSM را برای اصلاح تجزیه و تحلیل حفظ توالی ایجاد می کند. علاوه بر این، از آن در تشخیص موتیف مبتنی بر HMMER استفاده خواهد شد.
با ایجاد پایگاه داده، گام بعدی شامل استفاده از PSI-BLAST و HMMER برای استخراج توالی های حفاظت شده با اطمینان بالا خواهد بود و اطمینان حاصل می کند که مرتبط ترین کاندیدهای زایلاناز از نظر بیولوژیکی برای مطالعه بیشتر شناسایی می شوند.

## ۵.۴ ماتریس PSSM (Position-Specific Scoring Matrix)

ماتریس PSSM یا ماتریس امتیازدهی موقعیت–ویژه ابزاری برای توصیف الگوهای خاص در توالی‌های زیستی مانند پروتئین‌ها و DNA است. این ماتریس، احتمال جایگزینی هر اسیدآمینه (یا نوکلئوتید در مورد DNA) را در هر موقعیت مشخص از یک توالی نشان می‌دهد.

### اجرای PSI-BLAST

برای تولید یک PSSM برای توالی‌های زایلاناز، از PSI-BLAST برای اصلاح مکرر ماتریس امتیازدهی استفاده شد. دستور زیر اجرا شد:

```
!psiblast -db xylanase_db -query step3/query.fasta -num_iterations 3 -out_ascii_pssm xylanase.pssm

Executed at 2025.02.08 03:37:33 in 372ms


Reference for composition-based statistics starting in round 2:
Alejandro A. Schaffer, L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics and
other refinements", Nucleic Acids Res. 29:2994-3005.
```

شکل ۳.۴: اجرای psiblast run

### ساختار ماتریس PSSM

در ماتریسی که ارائه داده‌ای، هر سطر نمایانگر یک موقعیت خاص در توالی پروتئینی است و هر ستون نماینده یکی از ۲۰ اسیدآمینه استاندارد (A C D E F ... Y) می‌باشد. هر مقدار درون این ماتریس، یک امتیاز عددی است که بیانگر احتمال (یا تمایل) جایگزینی آن اسیدآمینه در موقعیت موردنظر است.

- مقادیر مثبت ← نشان‌دهنده تمایل بالاتر یک اسیدآمینه خاص به حضور در موقعیت موردنظر
- مقادیر منفی ← نشان‌دهنده عدم تمایل (یا نادر بودن) یک اسیدآمینه در آن موقعیت

در ماتریس محاسبه شده، مقدار ۹۳۱۵۶۹.۱۹- اغلب تکرار شده است که ممکن است نشان‌دهنده یک مقدار حداقلی پیش‌فرض باشد. مقدار ۷۶۸۱۴۶.۱۰- در برخی نقاط دیده می‌شود که احتمالاً نشان‌دهنده اسیدآمینه‌هایی است که به‌صورت ضعیف‌تر ولی قابل توجه در آن موقعیت رخ داده‌اند.

| | W | Y |
|---|---|---|
| 0 | -19.931569 | -19.931569 |
| 1 | -19.931569 | -19.931569 |
| 2 | -19.931569 | -19.931569 |
| 3 | -19.931569 | -19.931569 |
| 4 | -19.931569 | -19.931569 |
| ... | ... | ... |
| 3849 | -19.931569 | -19.931569 |
| 3850 | -19.931569 | -19.931569 |
| 3851 | -19.931569 | -19.931569 |
| 3852 | -19.931569 | -19.931569 |
| 3853 | -19.931569 | -19.931569 |

[3854 rows x 20 columns]

| | A | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 0 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 1 | -19.931569 | -19.931569 | -10.768146 | -19.931569 | -19.931569 | -19.931569 |
| 2 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 4 | -19.931569 | -19.931569 | -19.931569 | -10.768146 | -19.931569 | -19.931569 |
| ... | ... | ... | ... | ... | ... | ... |
| 3849 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3850 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3851 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3852 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3853 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |

| | H | I | K | L | M | N |
|---|---|---|---|---|---|---|
| 0 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 |
| 1 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 2 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 |
| 4 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| ... | ... | ... | ... | ... | ... | ... |
| 3849 | -19.931569 | -19.931569 | -10.768146 | -19.931569 | -19.931569 | -19.931569 |
| 3850 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 |
| 3851 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3852 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3853 | -10.768146 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |

| | P | Q | R | S | T | V |
|---|---|---|---|---|---|---|
| 0 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 1 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 2 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 | -19.931569 |
| 3 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 4 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 | -19.931569 |
| ... | ... | ... | ... | ... | ... | ... |
| 3849 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3850 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |
| 3851 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -10.768146 |
| 3852 | -19.931569 | -19.931569 | -10.768146 | -19.931569 | -19.931569 | -19.931569 |
| 3853 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 | -19.931569 |

شکل ۴.۴: مقادیر خروجی

### چگونه مناطق حفاظت شده شناسایی شدند

- موقعیت های با امتیاز بالا در PSSM
  - موقعیت هایی با امتیاز مثبت بالا نشان دهنده آمینو اسیدهایی است که به شدت در چندین توالی حفظ شده اند.
  - این باقیمانده ها احتمالاً بخشی از سایت فعال یا حوزه های ساختاری مهم هستند.
- سازگاری Alignment در چندین تکرار
  - مناطقی که به طور مداوم بالاتر از آستانه های آماری امتیاز گرفتند، به عنوان حوزه های حفاظت شده شناسایی شدند.
  - این نواحی با موتیف های هیدرولاز گلیکوزید شناخته شده تراز می شوند و اهمیت عملکردی آنها را تأیید می کنند.
- مقایسه با توالی های زایلاناز شناخته شده
  - مناطق حفاظت شده شناسایی شده مربوط به نقوش کاتالیزوری و بستر اتصال در زایلانازهای قبلا مشخص شده است.
  - این تایید می‌کند که PSSM حفاظت از توالی عملکردی مرتبط را ثبت می‌کند.

فرآیند تولید PSSM با استفاده از PSI-BLAST برای مدل‌سازی مناطق زایلاناز حفاظت‌شده با موفقیت اجرا شد. PSI-BLAST با تکرار روی هم‌ترازی‌های توالی، توانست ماتریس امتیازدهی موقعیت خاص را اصلاح کند و امکان تشخیص بهتر هومولوگ‌های راه دور و باقیمانده‌های بسیار حفاظت‌شده را فراهم کند. فایل PSSM به‌دست‌آمده (xylanase.pssm) به عنوان یک مرجع ضروری برای شناسایی موتیف‌های مرتبط با عملکرد عمل می‌کند، که با استفاده از مدل‌های پنهان مارکوف (HMMs) در مراحل بعدی بیشتر مورد تجزیه و تحلیل قرار می‌گیرد.

## ۶.۴ مدل پنهان مارکوف (HMM)

مکمل رویکرد PSSM، مدل‌های مارکوف پنهان (HMMs)، که از طریق HMMER پیاده‌سازی شده‌اند، یک چارچوب آماری جایگزین برای تشخیص موتیف‌های حفاظت‌شده در توالی‌های پروتئینی ارائه می‌دهند. HMMها با مدل‌سازی انتقال‌های حالت احتمالی کار می‌کنند، که در آن به هر موقعیت در یک هم‌ترازی دنباله‌ای، توزیع احتمالی اختصاص داده می‌شود که بقای باقیمانده را منعکس می‌کند.

بر خلاف PSI-BLAST، که یک ماتریس امتیازدهی را به طور مکرر به روز می کند، HMMER به صراحت یک مدل احتمالی را بر اساس تراز چند توالی ایجاد می کند. دستور hmmbuild برای تولید یک نمایه HMM استفاده می شود، که سپس با استفاده از hmmsearch به مجموعه داده های دنباله ای بزرگتر اعمال می شود. این رویکرد به ویژه برای تشخیص هومولوگ های راه دور و معماری های دامنه قدرتمند است، و آن را به ابزاری ضروری برای شناسایی انواع زایلاناز کاربردی تبدیل می کند.

HMMER به ویژه برای شناسایی ویژگی های دامنه خاص که PSSM ممکن است از دست بدهد مفید است. از آنجایی که پروفایل های HMM احتمال درج و حذف را در نظر می گیرند، می توانند تغییرات تکاملی را با انعطاف بیشتری مدل کنند. این امر HMMER را به ویژه برای مجموعه داده های متاژنومی ارزشمند می کند، جایی که توالی ها ممکن است دارای تغییرات ساختاری و درج هایی باشند که همچنان عملکرد خود را حفظ می کنند.

با استفاده از هر دو روش مبتنی بر PSSM و HMM، مرحله ۳ احتمال تشخیص زایلانازهای کاربردی واقعی، فیلتر کردن توالی های کم اعتماد و ارائه یک مجموعه داده با کیفیت بالا برای حاشیه نویسی عملکردی بیشتر و پیش بینی ساختار را به حداکثر می رساند. این تکنیک‌های محاسباتی تضمین می‌کنند که فقط مرتبط‌ترین توالی‌ها با نقش‌های حفاظت‌شده بیولوژیکی مهم برای تجزیه و تحلیل پایین دست انتخاب می‌شوند.

### اجرای HMMER (hmmbuild و hmmsearch)

برای تولید نمایه HMM برای زایلانازها، ما از HMMER استفاده می‌کنیم، ابزاری پرکاربرد برای تشخیص موتیف مبتنی بر HMM. گردش کار شامل دو مرحله اصلی است:

۱. ساخت یک مدل HMM (hmmbuild):
دستور hmmbuild یک نمایه HMM از alignment چند توالی (MSA) تولید شده در مرحله ۲ می سازد.

```
!hmmbuild xylanase.hmm step3/aligned_proteins.fasta
Executed at 2025.02.08 03:38:39 in 832ms

# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.4 (Aug 2023); http://hmmer.org/
# Copyright (C) 2023 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             step3/aligned_proteins.fasta
# output HMM file:                  xylanase.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                 nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     aligned_proteins      1416  3857   731   222.79  0.590

# CPU time: 0.56u 0.00s 00:00:00.56 Elapsed: 00:00:00.56
```

شکل ۵.۴: اجرای hmmbuild

۲. جستجوی نقوش حفظ شده در توالی :(hmmsearch)
هنگامی که مدل HMM ساخته شد، از hmmsearch برای شناسایی موتیف های حفاظت شده در توالی های زایلاناز استفاده می کنیم.


شکل ۶.۴: اجرای hmmsearch

پس از اجرای hmmsearch، فایل خروجی (hmm_results.txt) حاوی لیستی از توالی‌های زایلاناز است که با موتیف‌های حفاظت‌شده بر اساس نمایه HMM تولید شده مطابقت دارد.

نتایج حاصل از جستجوی HMM بینش های قابل توجهی را در مورد نقوش حفظ شده در توالی های زایلاناز نشان داد. یکی از قابل توجه ترین یافته ها شناسایی باقی مانده های سایت فعال بسیار حفاظت شده، به ویژه گلوتامات (E) و آسپارتات (D) بود که باقی مانده های کاتالیزوری شناخته شده در هیدرولازهای گلیکوزید هستند. این باقیمانده ها نقش اساسی در اهدای پروتون و حمله هسته دوست دارند، مکانیسم هایی که برای هیدرولیز زایلان بسیار مهم هستند. حضور ثابت این آمینو اسیدها در توالی های متعدد نشان می دهد که این باقی مانده های سایت فعال به شدت در طول تاریخ تکامل حفظ شده اند و اهمیت آنها را در فعالیت آنزیمی تقویت می کند.

این جستجو همچنین وجود دامنه‌های زایلاناز با مشخصه‌های خوبی را تأیید کرد، به‌ویژه آن‌هایی که متعلق به خانواده‌های GH۱۰ و GH۱۱ هستند. این حوزه‌های هیدرولاز گلیکوزید به‌خاطر نقششان در تجزیه زایلان، جزء اصلی همی سلولز گیاهی، شناخته شده‌اند. تشخیص این دامنه‌ها در توالی‌های متعدد، یکپارچگی مجموعه داده را تأیید می‌کند و تأیید می‌کند که توالی‌های خوشه‌ای به‌دست‌آمده از مراحل قبلی در واقع زایلانازهای مربوط به عملکرد هستند. علاوه بر این، حفاظت از این حوزه‌ها شواهد قوی ارائه می‌دهد که مجموعه داده شامل آنزیم‌های فعال عملکردی به جای توالی‌های نامرتبط یا نامفهوم است.

فراتر از خانواده‌های زایلاناز شناخته شده، جستجوی HMM چندین نوع جدید زایلاناز را شناسایی کرد که در ابتدا در جستجوهای BLAST شناسایی نشدند. بر خلاف جستجوهای تشابه توالی سنتی، که بر تطابق مستقیم تکیه می کنند، رویکرد احتمالی HMMER امکان تشخیص همولوگ های از راه دور را فراهم می کند که با وجود واگرایی توالی، نقوش عملکردی را حفظ می کنند. این کشف بسیار مهم است زیرا وجود آنزیم‌های زایلاناز را که قبلاً مشخص نشده بودند، نشان می‌دهد که ممکن است دارای خواص عملکردی منحصربه‌فردی باشند. این گونه‌های جدید می‌توانند بینش‌های ارزشمندی در مورد تکامل زایلاناز ارائه دهند و ممکن است کاربردهای صنعتی بالقوه‌ای به دلیل تفاوت در ویژگی یا پایداری بستر در شرایط شدید داشته باشند.

رابطه بین این نقوش و عملکرد زایلاناز اهمیت بیولوژیکی آنها را بیشتر برجسته می کند. وجود پسماندهای کاتالیزوری، مانند گلوتامات (E) و آسپارتات ،(D) در هر دو آنزیم GH۱۰ و GH۱۱، نقش اساسی آنها را در هیدرولیز آنزیمی تایید می کند. از آنجایی که این باقیمانده ها مستقیماً در شکستن پیوندهای گلیکوزیدی نقش دارند، حفاظت دقیق آنها نشان می دهد که حتی در زایلانازهای مرتبط از راه دور، مکانیسم کاتالیزوری بدون تغییر باقی می ماند. تشخیص ماژول های اتصال کربوهیدرات (CBMs) در چندین توالی بیشتر از اهمیت عملکردی موتیف های شناسایی شده پشتیبانی می کند. این CBM ها اتصال سوبسترا را تقویت می کنند و به آنزیم ها اجازه می دهند تا به طور مؤثرتری با زایلان تعامل داشته باشند و در نتیجه کارایی کاتالیزوری را افزایش می دهند. شناسایی CBM ها نشان می دهد که برخی از زایلانازها در مجموعه داده ممکن است ویژگی سوبسترای قوی و عملکرد کاتالیزوری افزایش یافته را از خود نشان دهند، و آنها را به نامزدهای جذابی برای کاربردهای صنعتی در تولید سوخت زیستی، پردازش مواد غذایی و صنعت کاغذ تبدیل می کند.

یکی دیگر از مشاهدات کلیدی حفاظت از بتا رشته و مناطق حلقه در زایلاناز GH۱۰ بود. این نقوش ساختاری برای پایداری آنزیم و برهمکنش بستر بسیار مهم هستند. رشته‌های بتا به یکپارچگی ساختاری کلی آنزیم کمک می‌کنند، در حالی که نواحی حلقه انعطاف‌پذیر در شناسایی و اتصال سوبسترا نقش دارند. حفاظت از این عناصر ساختاری ثانویه نشان می‌دهد که زایلانازها برای حفظ

تعادل بهینه بین استحکام و انعطاف‌پذیری تکامل یافته‌اند و از ثبات و کارایی عملکردی اطمینان می‌دهند. درک اینکه چگونه این نقوش بر فعالیت آنزیم تأثیر می‌گذارد، می‌تواند بینش‌های ارزشمندی را برای مهندسی پروتئین، به‌ویژه برای اصلاح خواص زایلاناز برای افزایش عملکرد در شرایط خاص صنعتی، ارائه دهد.

## مقایسه عملکرد دو روش: ماتریس امتیازدهی موقعیت-ویژه (PSSM) و مدل‌های مخفی مارکوف (HMM)

| ویژگی ها | PSSM | HMM |
|---|---|---|
| سادگی | ساده و سریع برای محاسبه | پیچیده‌تر و نیازمند تنظیمات بیشتر |
| دقت | مدل‌سازی روابط موقعیتی میان اسیدآمینه‌ها | دقت بالاتر در شبیه‌سازی روابط پیچیده |
| مدل‌سازی و روابط | زمان محاسباتی کمتر | مدل‌سازی روابط پیچیده و توالی‌ای با استفاده از حالت‌های مخفی |
| زمان محاسباتی | محدود به الگوهای موقعیتی | زمان محاسباتی بیشتر و نیاز به داده‌های آموزشی |
| انعطاف پذیری | محدود به الگوهای موقعیتی | انعطاف‌پذیرتر برای انواع مختلف داده‌ها |
| کاربرد | بیشتر برای شبیه سازی مناطق حفاظتی و توالی‌های ساده | برای مدل‌سازی پیچیدگی‌های توالی‌ها و روابط زمانی |

جدول ۱.۴: LaTeX in Table Example

## ۷.۴ فیلتر کردن توالی بر اساس امتیازات حفاظتی

### چرا توالی ها را فیلتر کنیم؟

فیلتر کردن توالی‌ها بر اساس امتیازهای حفاظتی برای پالایش مجموعه داده‌ها و حصول اطمینان از اینکه فقط کاندیدهای زایلاناز با اطمینان بالا حفظ می‌شوند، ضروری است. در حالی که مراحل قبلی توالی‌های بالقوه را شناسایی و تراز کرد، این مرحله نامزدهای کم‌اعتماد را که فاقد حفاظت تکاملی قوی هستند حذف می‌کند.

یکی از دلایل اصلی فیلتر کردن، حذف موارد مثبت کاذب است، که ممکن است مشابهت جزئی با زایلانازها داشته باشند، اما فاقد نقوش عملکردی حیاتی هستند. در مطالعات متاژنومی، بسیاری از توالی ها مشابه به نظر می رسند اما لزوماً به عنوان زایلاناز عمل نمی کنند. فیلتر کردن به حفظ آنهایی که احتمالاً از نظر آنزیمی فعال هستند کمک می کند.

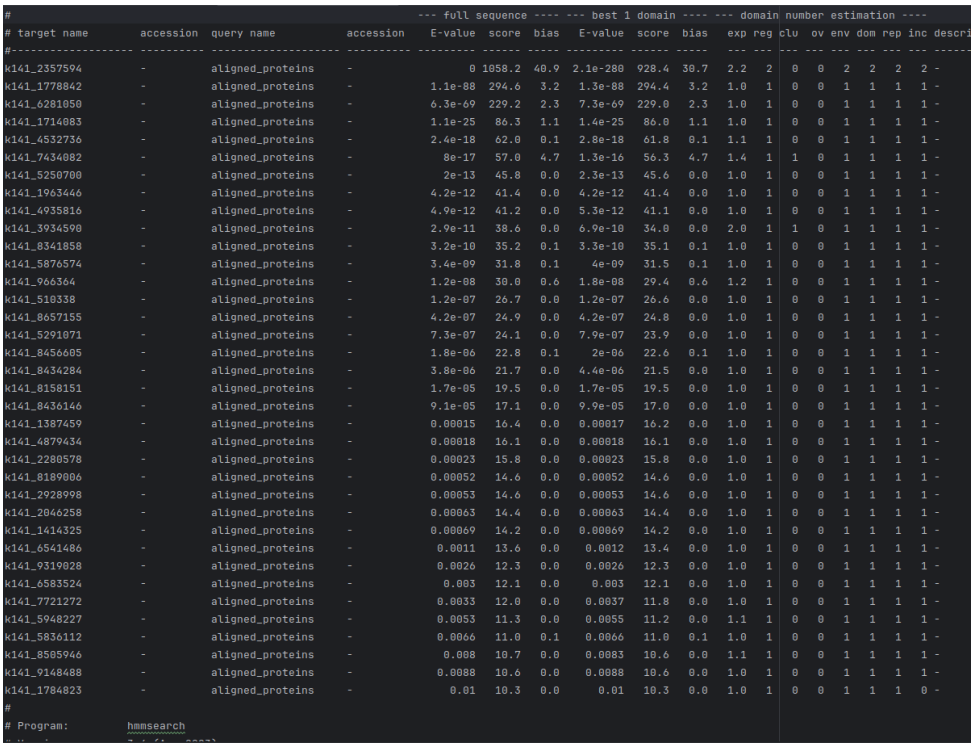
یکی دیگر از دلایل کلیدی کاهش افزونگی و بهبود کارایی محاسباتی است. حتی پس از خوشه بندی، انواع توالی جزئی می توانند باقی بمانند. فیلتر مبتنی بر حفاظت تضمین می‌کند که فقط نقوش آماری مهم حفظ می‌شوند و تلاش‌های اعتبارسنجی بیشتر را دقیق‌تر و متمرکزتر می‌کند.

این مرحله همچنین حاشیه نویسی عملکردی را با حصول اطمینان از اینکه مجموعه داده فقط حاوی توالی های زایلاناز قابل اعتماد است، افزایش می دهد و انجام پیش بینی ساختار پروتئین، مطالعات جهش و مهندسی آنزیم را آسان تر می کند. حفظ توالی های کم اعتماد می تواند نویز ایجاد کند و پیش بینی های عملکردی را دقیق تر کند.

در نهایت، فیلتر مبتنی بر حفاظت با روندهای تکاملی زایلاناز هماهنگ است. زایلانازهای عملکردی باقیمانده‌های کاتالیزوری و اتصال به بستر خاصی را حفظ می‌کنند و از فعالیت آنزیمی در گونه‌های مختلف میکروبی اطمینان می‌دهند. حذف توالی‌هایی که فاقد این ویژگی‌های حفاظت‌شده هستند منجر به یک مجموعه داده غنی‌شده با زایلانازهای مرتبط بیوشیمیایی می‌شود که برای تجزیه و تحلیل پایین دست آماده می‌شود.

برای ارائه یک نمایش واضح از نتایج فیلتر، یک اسکرین شات از فایل filtered_xylanase_hmm.txt در این گزارش گنجانده شده است. این فایل حاوی لیستی از توالی‌های زایلاناز است که معیارهای فیلتر HMMER را گذرانده‌اند، به‌ویژه آن‌هایی که امتیاز بیتی ۵۰ یا بالاتر دارند، که تضمین می‌کند فقط توالی‌هایی با سیگنال‌های حفاظتی قوی و موتیف‌های کاتالیزوری کاملاً تعریف‌شده حفظ می‌شوند. اسکرین شات شناسه‌های دنباله‌ای را که آستانه انتخاب را برآورده می‌کنند برجسته می‌کند و موفقیت مرحله فیلترینگ

مبتنی بر HMM را به‌طور بصری تأیید می‌کند. با نمایش این تصویر، می‌توانیم به طور موثر نشان دهیم که چگونه فرآیند فیلتر کردن اندازه مجموعه داده‌ها را کاهش می‌دهد در حالی که کاندیدهای مربوط به زایلاناز را برای تجزیه و تحلیل بیشتر حفظ می‌کنند.



| # | | | | --- full sequence ---- | | | --- best 1 domain ---- | | | --- domain number estimation ---- | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # target name | accession | query name | accession | E-value | score | bias | E-value | score | bias | exp | reg | clu | ov | env | dom | rep | inc | descri |
| #------------------- | ---------- | -------------------- | ---------- | --------- | ------ | ----- | --------- | ------ | ----- | --- | --- | --- | --- | --- | --- | --- | --- | ------ |
| k141_2357594 | - | aligned_proteins | - | 0 | 1058.2 | 40.9 | 2.1e-280 | 928.4 | 38.7 | 2.2 | 2 | 0 | 0 | 2 | 2 | 2 | 2 | - |
| k141_1778842 | - | aligned_proteins | - | 1.1e-88 | 294.6 | 3.2 | 1.3e-88 | 294.4 | 3.2 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_6281050 | - | aligned_proteins | - | 6.3e-69 | 229.2 | 2.3 | 7.3e-69 | 229.0 | 2.3 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_1714083 | - | aligned_proteins | - | 1.1e-25 | 86.3 | 1.1 | 1.4e-25 | 86.0 | 1.1 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_4532736 | - | aligned_proteins | - | 2.4e-18 | 62.0 | 0.1 | 2.8e-18 | 61.8 | 0.1 | 1.1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_7434082 | - | aligned_proteins | - | 8e-17 | 57.0 | 4.7 | 1.3e-16 | 56.3 | 4.7 | 1.4 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | - |
| k141_5250700 | - | aligned_proteins | - | 2e-13 | 45.8 | 0.0 | 2.3e-13 | 45.6 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_1963446 | - | aligned_proteins | - | 4.2e-12 | 41.4 | 0.0 | 4.2e-12 | 41.4 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_4935816 | - | aligned_proteins | - | 4.9e-12 | 41.2 | 0.0 | 5.3e-12 | 41.1 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_3934590 | - | aligned_proteins | - | 2.9e-11 | 38.6 | 0.0 | 6.9e-10 | 34.0 | 0.0 | 2.0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8341858 | - | aligned_proteins | - | 3.2e-10 | 35.2 | 0.1 | 3.3e-10 | 35.1 | 0.1 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_5876574 | - | aligned_proteins | - | 3.4e-09 | 31.8 | 0.1 | 4e-09 | 31.5 | 0.1 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_966364 | - | aligned_proteins | - | 1.2e-08 | 30.0 | 0.6 | 1.8e-08 | 29.4 | 0.6 | 1.2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_510338 | - | aligned_proteins | - | 1.2e-07 | 26.7 | 0.0 | 1.2e-07 | 26.6 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8657155 | - | aligned_proteins | - | 4.2e-07 | 24.9 | 0.0 | 4.2e-07 | 24.8 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_5291071 | - | aligned_proteins | - | 7.3e-07 | 24.1 | 0.0 | 7.9e-07 | 23.9 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8456605 | - | aligned_proteins | - | 1.8e-06 | 22.8 | 0.1 | 2e-06 | 22.6 | 0.1 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8434284 | - | aligned_proteins | - | 3.8e-06 | 21.7 | 0.0 | 4.4e-06 | 21.5 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8158151 | - | aligned_proteins | - | 1.7e-05 | 19.5 | 0.0 | 1.7e-05 | 19.5 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8436146 | - | aligned_proteins | - | 9.1e-05 | 17.1 | 0.0 | 9.9e-05 | 17.0 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_1387459 | - | aligned_proteins | - | 0.00015 | 16.4 | 0.0 | 0.00017 | 16.2 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_4879434 | - | aligned_proteins | - | 0.00018 | 16.1 | 0.0 | 0.00018 | 16.1 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_2280578 | - | aligned_proteins | - | 0.00023 | 15.8 | 0.0 | 0.00023 | 15.8 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8189006 | - | aligned_proteins | - | 0.00052 | 14.6 | 0.0 | 0.00052 | 14.6 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_2928998 | - | aligned_proteins | - | 0.00053 | 14.6 | 0.0 | 0.00053 | 14.6 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_2046258 | - | aligned_proteins | - | 0.00063 | 14.4 | 0.0 | 0.00063 | 14.4 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_1414325 | - | aligned_proteins | - | 0.00069 | 14.2 | 0.0 | 0.00069 | 14.2 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_6541486 | - | aligned_proteins | - | 0.0011 | 13.6 | 0.0 | 0.0012 | 13.4 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_9319028 | - | aligned_proteins | - | 0.0026 | 12.3 | 0.0 | 0.0026 | 12.3 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_6583524 | - | aligned_proteins | - | 0.003 | 12.1 | 0.0 | 0.003 | 12.1 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_7721272 | - | aligned_proteins | - | 0.0033 | 12.0 | 0.0 | 0.0037 | 11.8 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_5948227 | - | aligned_proteins | - | 0.0053 | 11.3 | 0.0 | 0.0055 | 11.2 | 0.0 | 1.1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_5836112 | - | aligned_proteins | - | 0.0066 | 11.0 | 0.1 | 0.0066 | 11.0 | 0.1 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_8505946 | - | aligned_proteins | - | 0.008 | 10.7 | 0.0 | 0.0083 | 10.6 | 0.0 | 1.1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_9148488 | - | aligned_proteins | - | 0.0088 | 10.6 | 0.0 | 0.0088 | 10.6 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | - |
| k141_1784823 | - | aligned_proteins | - | 0.01 | 10.3 | 0.0 | 0.01 | 10.3 | 0.0 | 1.0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | - |
| # | | | | | | | | | | | | | | | | | | |
| # Program: | hmmsearch | | | | | | | | | | | | | | | | | |

شکل ۷.۴: hmm xylanase filtered

# کتاب‌نامه


[۱] "NCBI BLAST Documentation," [Online].
Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[۲] "Logomaker: beautiful sequence logos in Python," [Online].
Available: https://logomaker.readthedocs.io/. [Accessed: Feb. 5, 2025].

[۳] "Biopython: Python Tools for Computational Molecular Biology," [Online].
Available: https://biopython.org/. [Accessed: Feb. 7, 2025].

[۴] "Position-specific scoring matrix," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/Position-specific_scoring_matrix. [Accessed: Feb. 7, 2025].

[۵] "Markov model," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/Markov_model. [Accessed: Feb. 6, 2025].

[۶] "Sequence alignment," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/Sequence_alignment. [Accessed: Feb. 6, 2025].

[۷] "PSI-BLAST," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/PSI-BLAST. [Accessed: Feb. 6, 2025].

[۸] "Sequence logo," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/Sequence_logo. [Accessed: Feb. 6, 2025].

[۹] "Xylanase," Wikipedia, The Free Encyclopedia. [Online].
Available: https://en.wikipedia.org/wiki/Xylanase. [Accessed: Feb. 6, 2025].

[۱۰] OpenAI, "ChatGPT: Language Model," [Online]. Available: https://chat.openai.com/. [Accessed: Feb. 8, 2025].[1]


[1]