Protein Domains and Distant Relatives

PSI-BLAST, Clustering, Multiple Sequence Alignments and HMMER

Brandi Cantarel, Ph.D
UTSW, Department of Bioinformatics
Programming for Biology 2018

Protein Domain Take Home

- Protein divergence is not uniform over a protein - some parts are more conserved than others
- Position specific scoring matrices can capture the specific patterns of conservation at different sites in a protein
- PSI-BLAST combines searching, multiple alignment, and PSSMs
- Statistical estimates are difficult with PSSMs, use PSI- SEARCH and PSI-PRSS
- HMMER3 creates HMM models of a protein family from a multiple sequence alignment
- Iterative PSSM/HMM searches may be contaminated by Homologous Overextension
- Single models cannot capture diverse families (PFAM Clans)
- Protein domains can be identified using RPS-BLAST or CDD searching

Inferring Homology from Statistical Significance

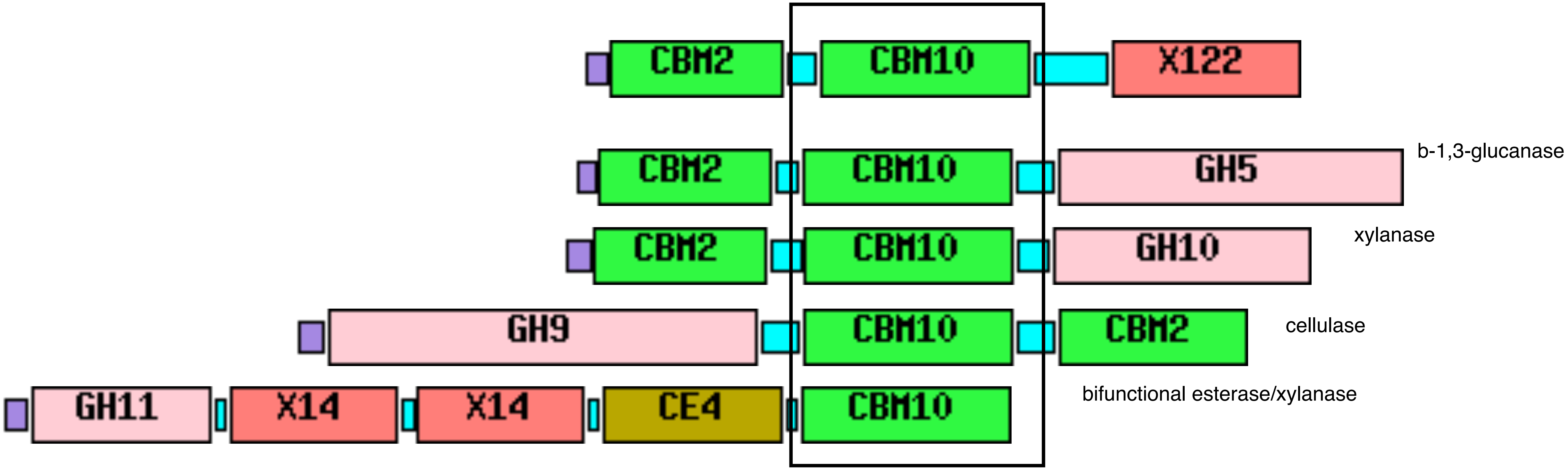
Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences

If a similarity is NOT *RANDOM,* then it must be NOT *UNRELATED*

Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

Protein Domains are structural units that can pair with different partners.

Homology in Domains



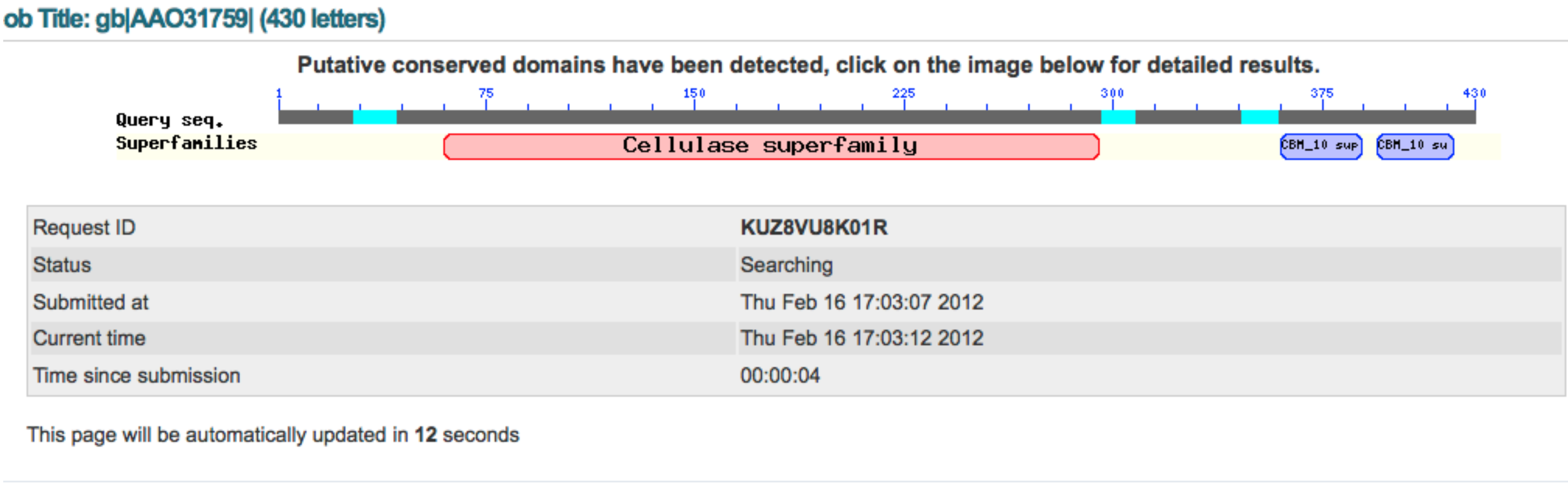
b-1,3-glucanase

xylanase

cellulase

bifunctional esterase/xylanase

Imagine you are searching with a protein with multiple domains



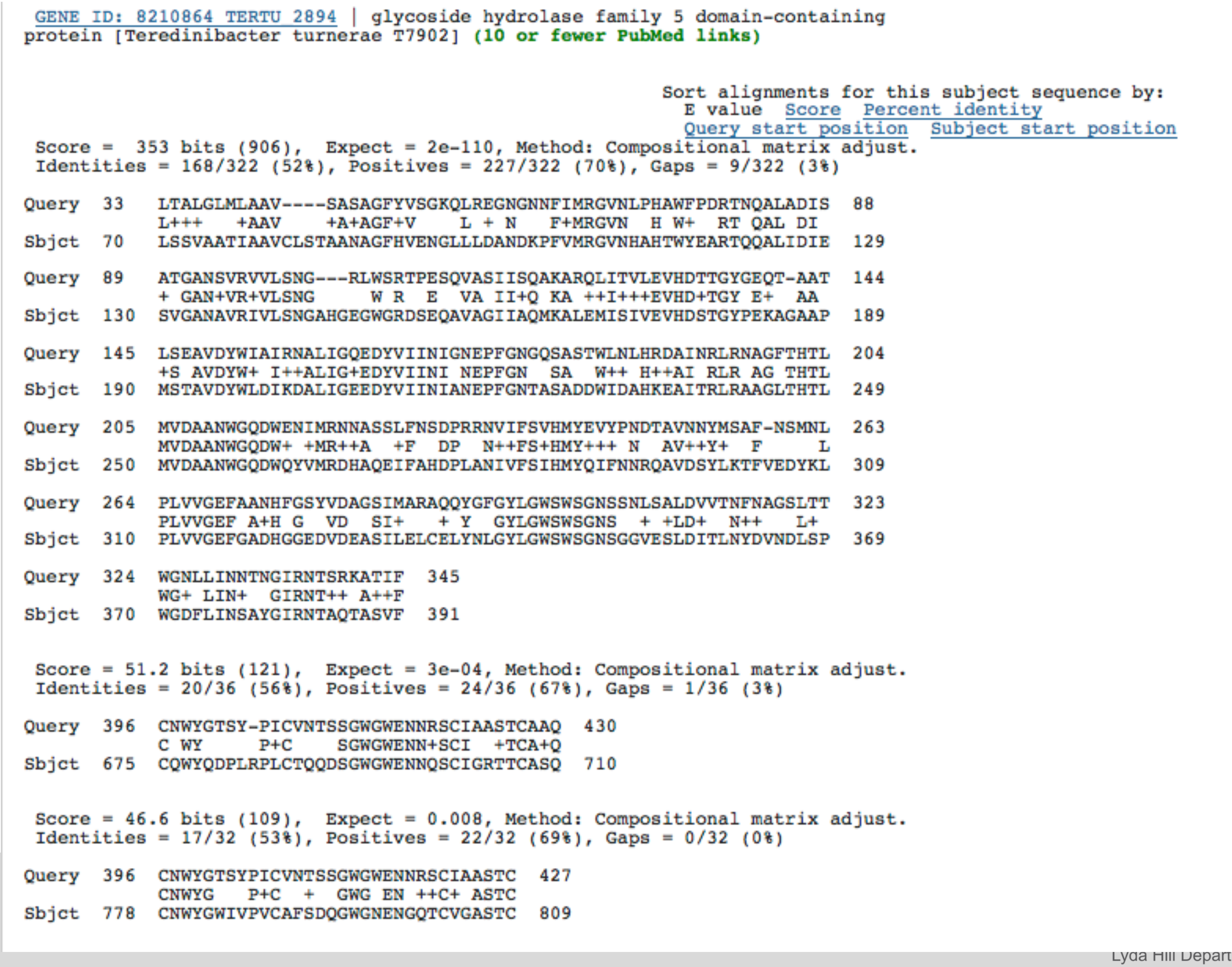
ob Title: gb|AAO31759| (430 letters)

Putative conserved domains have been detected, click on the image below for detailed results.

Request ID	**KUZ8VU8K01R**
Status	Searching
Submitted at	Thu Feb 16 17:03:07 2012
Current time	Thu Feb 16 17:03:12 2012
Time since submission	00:00:04

This page will be automatically updated in **12** seconds

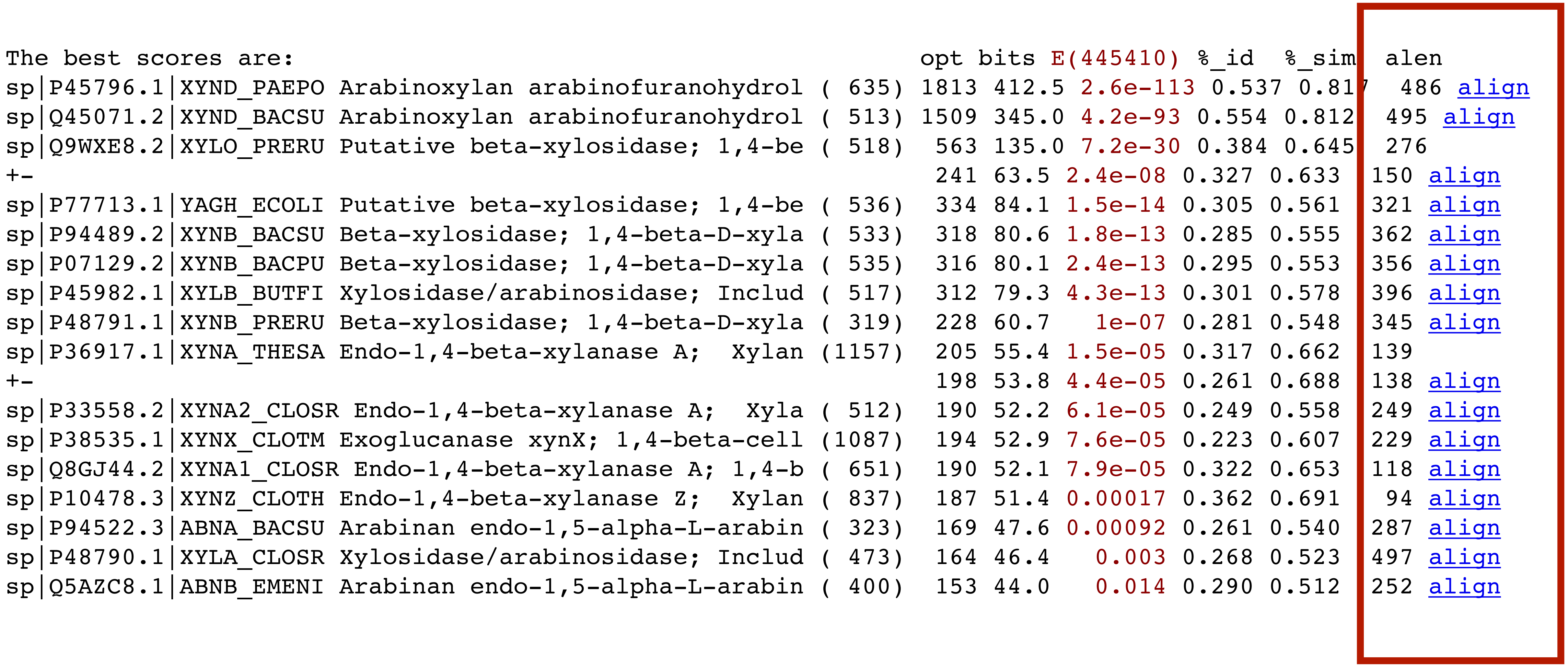
BLAST Reports Multiple Highest Scoring Pairs



```
 GENE ID: 8210864 TERTU_2894 | glycoside hydrolase family 5 domain-containing
protein [Teredinibacter turnerae T7902] (10 or fewer PubMed links)


                                                     Sort alignments for this subject sequence by:
                                                       E value  Score  Percent identity
                                                       Query start position  Subject start position
 Score =  353 bits (906),  Expect = 2e-110, Method: Compositional matrix adjust.
 Identities = 168/322 (52%), Positives = 227/322 (70%), Gaps = 9/322 (3%)

Query  33   LTALGLMLAAV----SASAGFYVSGKQLREGNGNNFIMRGVNLPHAWFPDRTNQALADIS  88
            L+++   +AAV    +A+AGF+V    L + N   F+MRGVN  H W+  RT QAL DI
Sbjct  70   LSSVAATIAAVCLSTAANAGFHVENGLLLDANDKPFVMRGVNHAHTWYEARTQQALIDIE  129

Query  89   ATGANSVRVVLSNG---RLWSRTPESQVASIISQAKARQLITVLEVHDTTGYGEQT-AAT  144
            + GAN+VR+VLSNG     W R  E  VA II+Q KA ++I+++EVHD+TGY E+  AA
Sbjct  130  SVGANAVRIVLSNGAHGEGWGRDSEQAVAGIIAQMKALEMISIVEVHDSTGYPEKAGAAP  189

Query  145  LSEAVDYWIAIRNALIGQEDYVIINIGNEPFGNGQSASTWLNLHRDAINRLRNAGFTHTL  204
            +S AVDYW+ I++ALIG+EDYVIINI NEPFGN  SA  W++ H++AI RLR AG THTL
Sbjct  190  MSTAVDYWLDIKDALIGEEDYVIINIANEPFGNTASADDWIDAHKEAITRLRAAGLTHTL  249

Query  205  MVDAANWGQDWENIMRNNASSLFNSDPRRNVIFSVHMYEVYPNDTAVNNYMSAF-NSMNL  263
            MVDAANWGQDW+ +MR++A  +F  DP  N++FS+HMY+++ N  AV++Y+  F     L
Sbjct  250  MVDAANWGQDWQYVMRDHAQEIFAHDPLANIVFSIHMYQIFNNRQAVDSYLKTFVEDYKL  309

Query  264  PLVVGEFAANHFGSYVDAGSIMARAQQYGFGYLGWSWSGNSSNLSALDVVTNFNAGSLTT  323
            PLVVGEF A+H G  VD  SI+   + Y  GYLGWSWSGNS  + +LD+  N++   L+
Sbjct  310  PLVVGEFGADHGGEDVDEASILELCELYNLGYLGWSWSGNSGGVESLDITLNYDVNDLSP  369

Query  324  WGNLLINNTNGIRNTSRKATIF  345
            WG+ LIN+  GIRNT++ A++F
Sbjct  370  WGDFLINSAYGIRNTAQTASVF  391


 Score = 51.2 bits (121),  Expect = 3e-04, Method: Compositional matrix adjust.
 Identities = 20/36 (56%), Positives = 24/36 (67%), Gaps = 1/36 (3%)

Query  396  CNWYGTSY-PICVNTSSGWGWENNRSCIAASTCAAQ  430
            C WY     P+C     SGWGWENN+SCI  +TCA+Q
Sbjct  675  CQWYQDPLRPLCTQQDSGWGWENNQSCIGRTTCASQ  710


 Score = 46.6 bits (109),  Expect = 0.008, Method: Compositional matrix adjust.
 Identities = 17/32 (53%), Positives = 22/32 (69%), Gaps = 0/32 (0%)

Query  396  CNWYGTSYPICVNTSSGWGWENNRSCIAASTC  427
            CNWYG   P+C  +  GWG EN ++C+ ASTC
Sbjct  778  CNWYGWIVPVCAFSDQGWGNENGQTCVGASTC  809
```

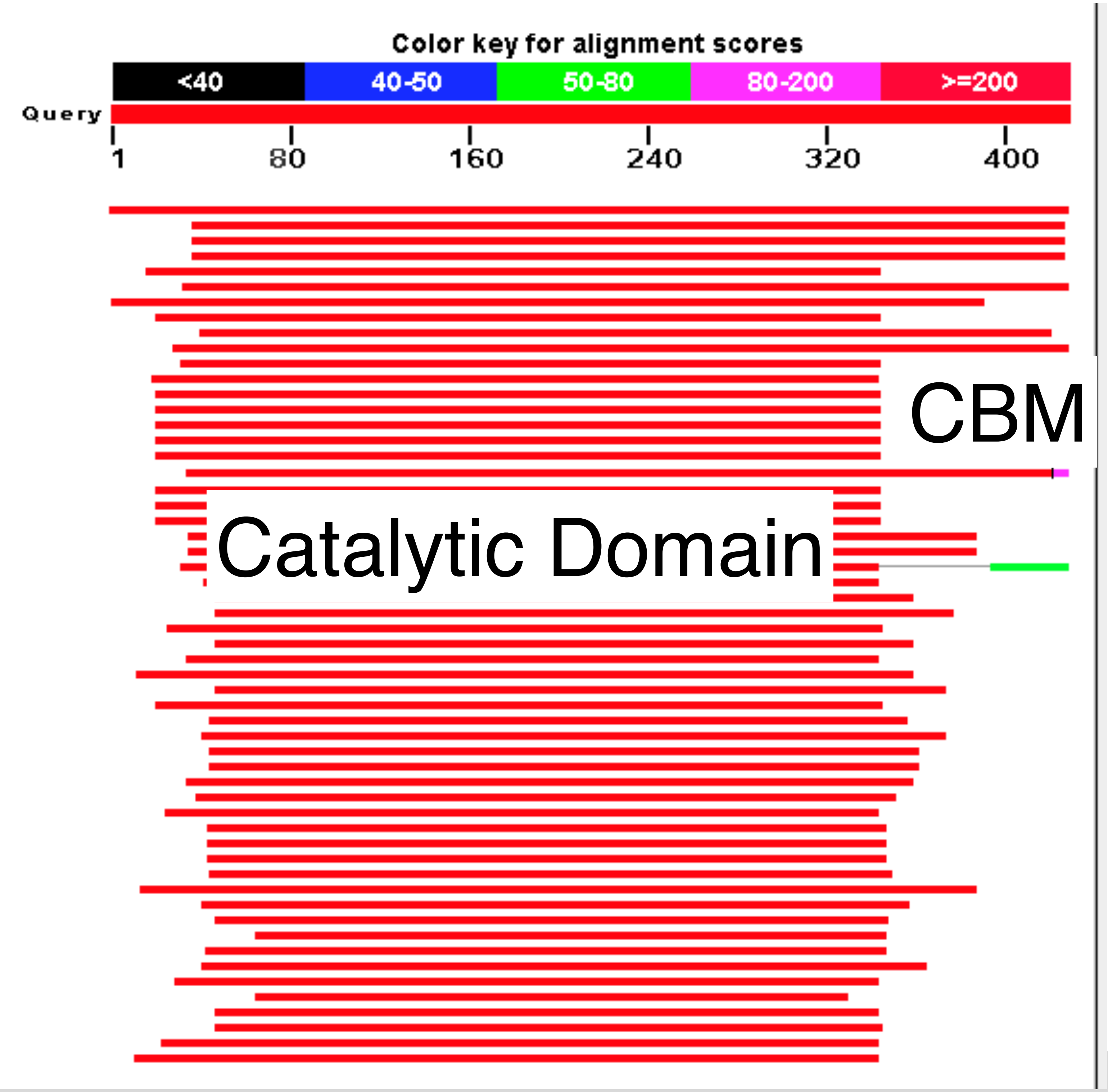
Homology in Domains

Xylanase



```
The best scores are:                                           opt bits E(445410) %_id  %_sim  alen
sp|P45796.1|XYND_PAEPO Arabinoxylan arabinofuranohydrol ( 635) 1813 412.5 2.6e-113 0.537 0.817  486 align
sp|Q45071.2|XYND_BACSU Arabinoxylan arabinofuranohydrol ( 513) 1509 345.0 4.2e-93 0.554 0.812  495 align
sp|Q9WXE8.2|XYLO_PRERU Putative beta-xylosidase; 1,4-be ( 518)  563 135.0 7.2e-30 0.384 0.645  276
+-                                                              241 63.5 2.4e-08 0.327 0.633 150 align
sp|P77713.1|YAGH_ECOLI Putative beta-xylosidase; 1,4-be ( 536)  334 84.1 1.5e-14 0.305 0.561 321 align
sp|P94489.2|XYNB_BACSU Beta-xylosidase; 1,4-beta-D-xyla ( 533)  318 80.6 1.8e-13 0.285 0.555 362 align
sp|P07129.2|XYNB_BACPU Beta-xylosidase; 1,4-beta-D-xyla ( 535)  316 80.1 2.4e-13 0.295 0.553 356 align
sp|P45982.1|XYLB_BUTFI Xylosidase/arabinosidase; Includ ( 517)  312 79.3 4.3e-13 0.301 0.578 396 align
sp|P48791.1|XYNB_PRERU Beta-xylosidase; 1,4-beta-D-xyla ( 319)  228 60.7   1e-07 0.281 0.548 345 align
sp|P36917.1|XYNA_THESA Endo-1,4-beta-xylanase A;  Xylan (1157)  205 55.4 1.5e-05 0.317 0.662 139
+-                                                              198 53.8 4.4e-05 0.261 0.688 138 align
sp|P33558.2|XYNA2_CLOSR Endo-1,4-beta-xylanase A;  Xyla ( 512)  190 52.2 6.1e-05 0.249 0.558 249 align
sp|P38535.1|XYNX_CLOTM Exoglucanase xynX; 1,4-beta-cell (1087)  194 52.9 7.6e-05 0.223 0.607 229 align
sp|Q8GJ44.2|XYNA1_CLOSR Endo-1,4-beta-xylanase A; 1,4-b ( 651)  190 52.1 7.9e-05 0.322 0.653 118 align
sp|P10478.3|XYNZ_CLOTH Endo-1,4-beta-xylanase Z;  Xylan ( 837)  187 51.4 0.00017 0.362 0.691  94 align
sp|P94522.3|ABNA_BACSU Arabinan endo-1,5-alpha-L-arabin ( 323)  169 47.6 0.00092 0.261 0.540 287 align
sp|P48790.1|XYLA_CLOSR Xylosidase/arabinosidase; Includ ( 473)  164 46.4   0.003 0.268 0.523 497 align
sp|Q5AZC8.1|ABNB_EMENI Arabinan endo-1,5-alpha-L-arabin ( 400)  153 44.0   0.014 0.290 0.512 252 align
```

Not all hits are to the full protein



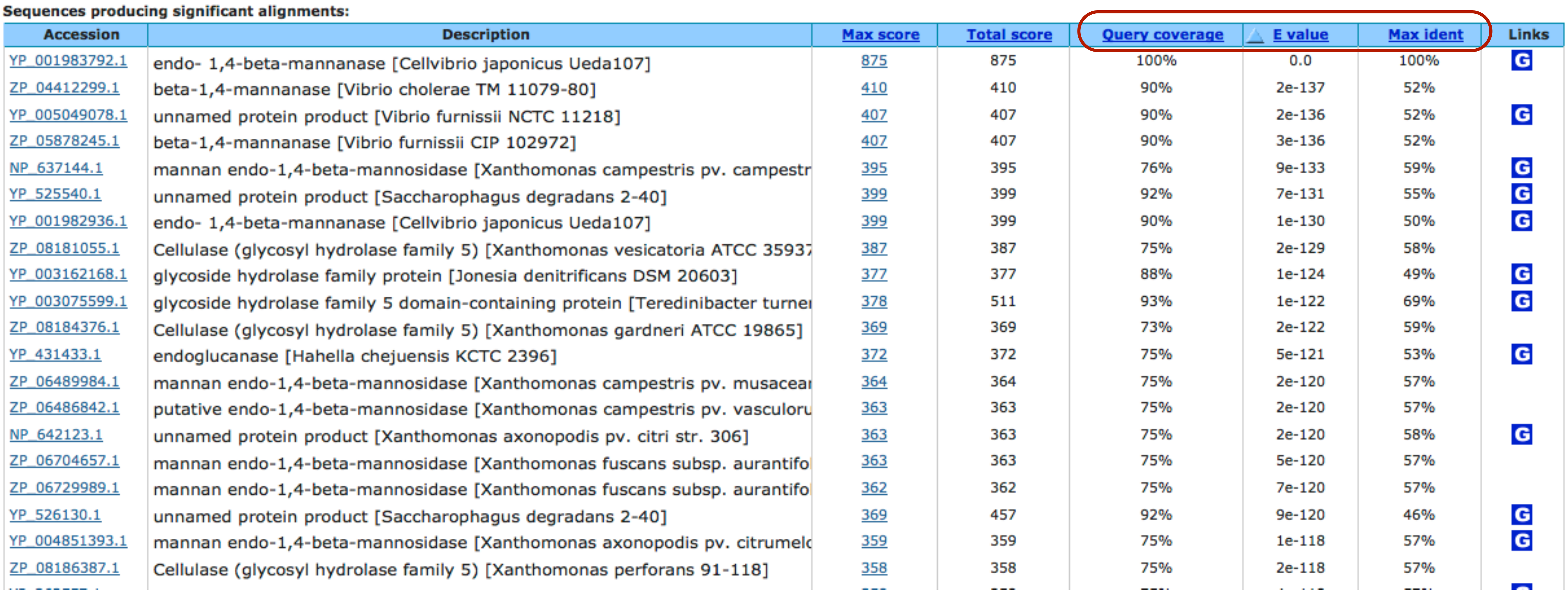
CBM

Catalytic Domain

Look at the Alignment Coverage

Score

E-value



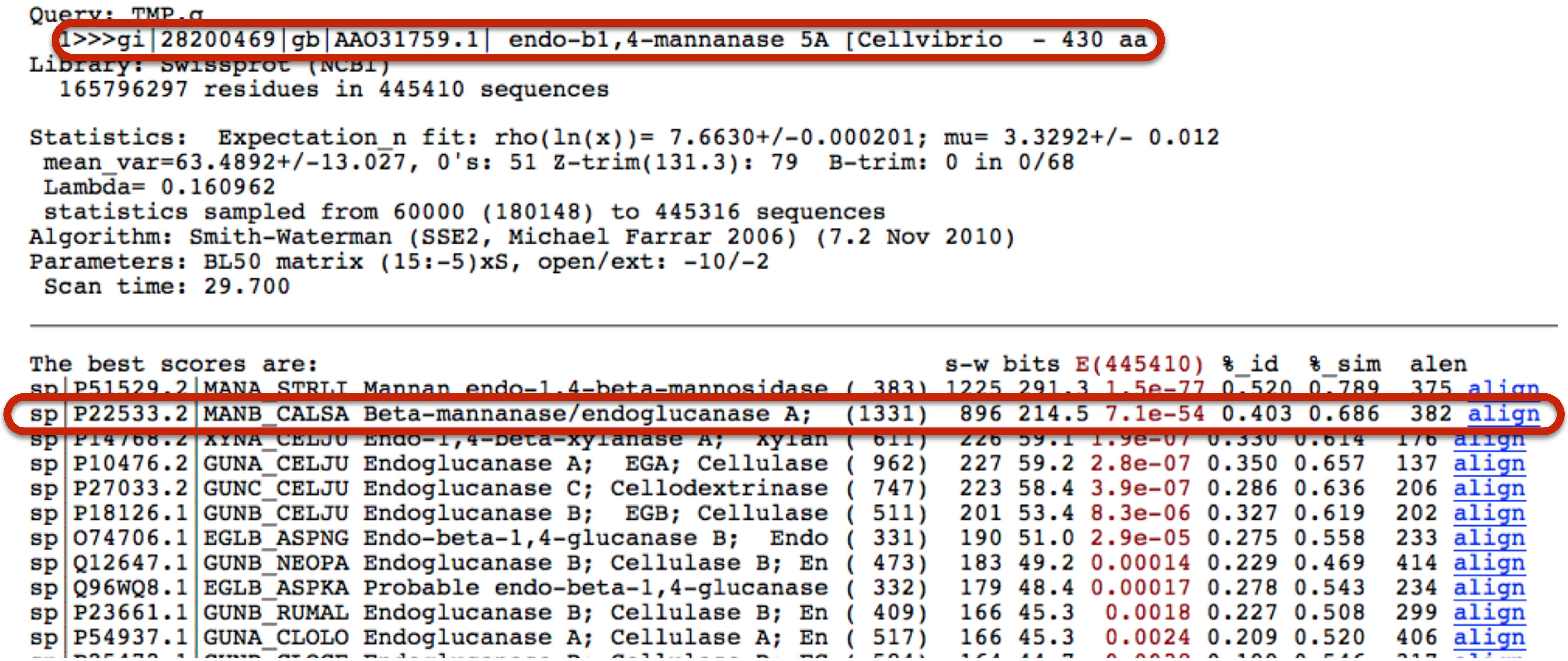
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
YP_001983792.1	endo- 1,4-beta-mannanase [Cellvibrio japonicus Ueda107]	875	875	100%	0.0	100%	G
ZP_04412299.1	beta-1,4-mannanase [Vibrio cholerae TM 11079-80]	410	410	90%	2e-137	52%	
YP_005049078.1	unnamed protein product [Vibrio furnissii NCTC 11218]	407	407	90%	2e-136	52%	G
ZP_05878245.1	beta-1,4-mannanase [Vibrio furnissii CIP 102972]	407	407	90%	3e-136	52%	
NP_637144.1	mannan endo-1,4-beta-mannosidase [Xanthomonas campestris pv. campestr	395	395	76%	9e-133	59%	G
YP_525540.1	unnamed protein product [Saccharophagus degradans 2-40]	399	399	92%	7e-131	55%	G
YP_001982936.1	endo- 1,4-beta-mannanase [Cellvibrio japonicus Ueda107]	399	399	90%	1e-130	50%	G
ZP_08181055.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas vesicatoria ATCC 35937	387	387	75%	2e-129	58%	
YP_003162168.1	glycoside hydrolase family protein [Jonesia denitrificans DSM 20603]	377	377	88%	1e-124	49%	G
YP_003075599.1	glycoside hydrolase family 5 domain-containing protein [Teredinibacter turnei	378	511	93%	1e-122	69%	G
ZP_08184376.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas gardneri ATCC 19865]	369	369	73%	2e-122	59%	
YP_431433.1	endoglucanase [Hahella chejuensis KCTC 2396]	372	372	75%	5e-121	53%	G
ZP_06489984.1	mannan endo-1,4-beta-mannosidase [Xanthomonas campestris pv. musacear	364	364	75%	2e-120	57%	
ZP_06486842.1	putative endo-1,4-beta-mannosidase [Xanthomonas campestris pv. vasculoru	363	363	75%	2e-120	57%	
NP_642123.1	unnamed protein product [Xanthomonas axonopodis pv. citri str. 306]	363	363	75%	2e-120	58%	G
ZP_06704657.1	mannan endo-1,4-beta-mannosidase [Xanthomonas fuscans subsp. aurantifo	363	363	75%	5e-120	57%	
ZP_06729989.1	mannan endo-1,4-beta-mannosidase [Xanthomonas fuscans subsp. aurantifo	362	362	75%	7e-120	57%	
YP_526130.1	unnamed protein product [Saccharophagus degradans 2-40]	369	457	92%	9e-120	46%	G
YP_004851393.1	mannan endo-1,4-beta-mannosidase [Xanthomonas axonopodis pv. citrumelc	359	359	75%	1e-118	57%	G
ZP_08186387.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas perforans 91-118]	358	358	75%	2e-118	57%	

Coverage

MaxID

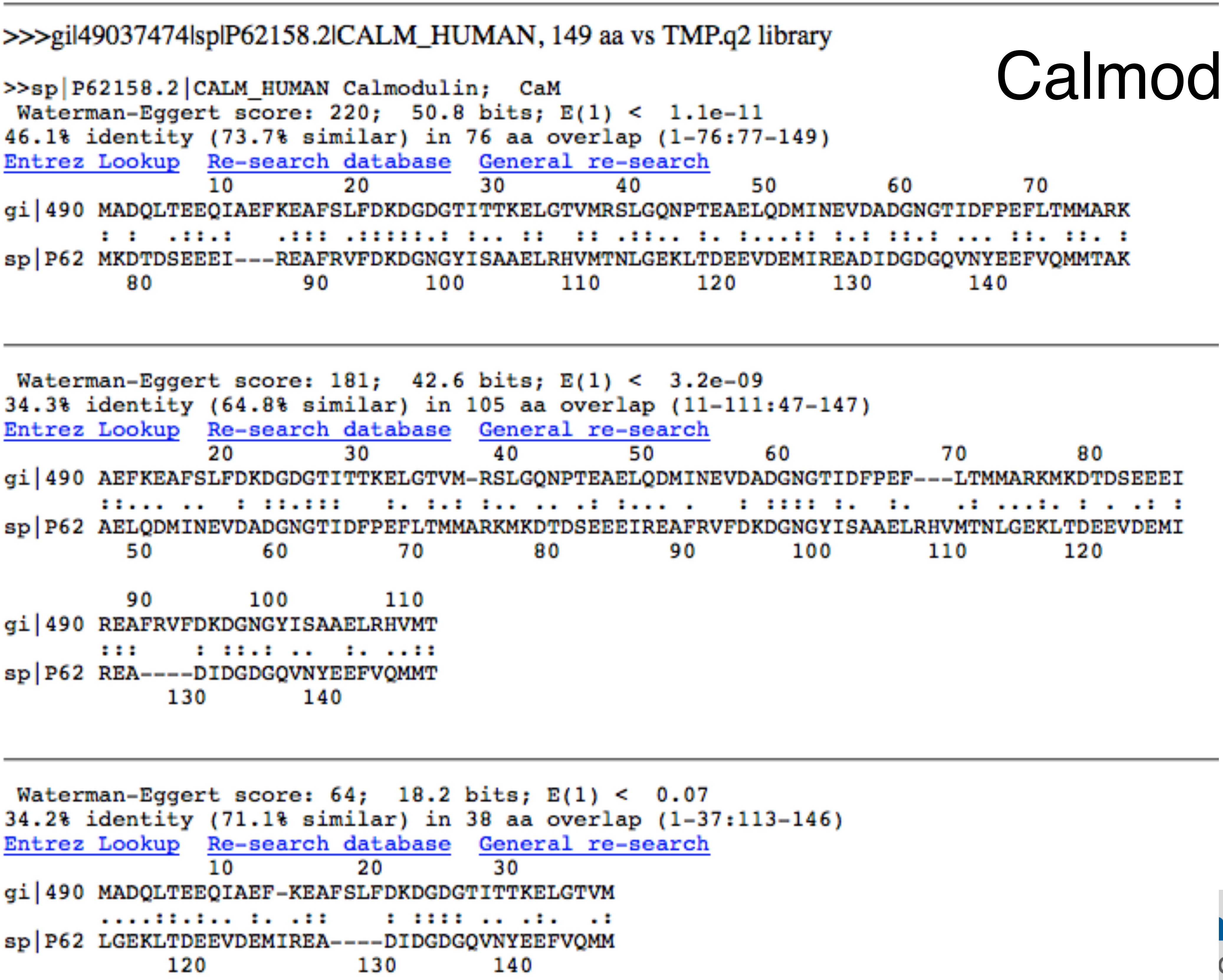
Examine The Alignment Length



```
Query: TMP.q
  1>>>gi|28200469|gb|AAO31759.1| endo-b1,4-mannanase 5A [Cellvibrio  - 430 aa
Library: Swissprot (NCBI)
  165796297 residues in 445410 sequences

Statistics:  Expectation_n fit: rho(ln(x))= 7.6630+/-0.000201; mu= 3.3292+/- 0.012
 mean_var=63.4892+/-13.027, 0's: 51 Z-trim(131.3): 79  B-trim: 0 in 0/68
 Lambda= 0.160962
 statistics sampled from 60000 (180148) to 445316 sequences
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2010)
Parameters: BL50 matrix (15:-5)xS, open/ext: -10/-2
 Scan time: 29.700

The best scores are:                                      s-w bits E(445410) %_id  %_sim  alen
sp|P51529.2|MANA_STRLI Mannan endo-1,4-beta-mannosidase ( 383) 1225 291.3 1.5e-77 0.520 0.789  375 align
sp|P22533.2|MANB_CALSA Beta-mannanase/endoglucanase A;  (1331)  896 214.5 7.1e-54 0.403 0.686  382 align
sp|P14768.2|XYNA_CELJU Endo-1,4-beta-xylanase A;  Xylan ( 611)  226 59.1 1.9e-07 0.330 0.614  176 align
sp|P10476.2|GUNA_CELJU Endoglucanase A;  EGA; Cellulase ( 962)  227 59.2 2.8e-07 0.350 0.657  137 align
sp|P27033.2|GUNC_CELJU Endoglucanase C; Cellodextrinase ( 747)  223 58.4 3.9e-07 0.286 0.636  206 align
sp|P18126.1|GUNB_CELJU Endoglucanase B;  EGB; Cellulase ( 511)  201 53.4 8.3e-06 0.327 0.619  202 align
sp|O74706.1|EGLB_ASPNG Endo-beta-1,4-glucanase B;  Endo ( 331)  190 51.0 2.9e-05 0.275 0.558  233 align
sp|Q12647.1|GUNB_NEOPA Endoglucanase B; Cellulase B; En ( 473)  183 49.2 0.00014 0.229 0.469  414 align
sp|Q96WQ8.1|EGLB_ASPKA Probable endo-beta-1,4-glucanase ( 332)  179 48.4 0.00017 0.278 0.543  234 align
sp|P23661.1|GUNB_RUMAL Endoglucanase B; Cellulase B; En ( 409)  166 45.3  0.0018 0.227 0.508  299 align
sp|P54937.1|GUNA_CLOLO Endoglucanase A; Cellulase A; En ( 517)  166 45.3  0.0024 0.209 0.520  406 align
```

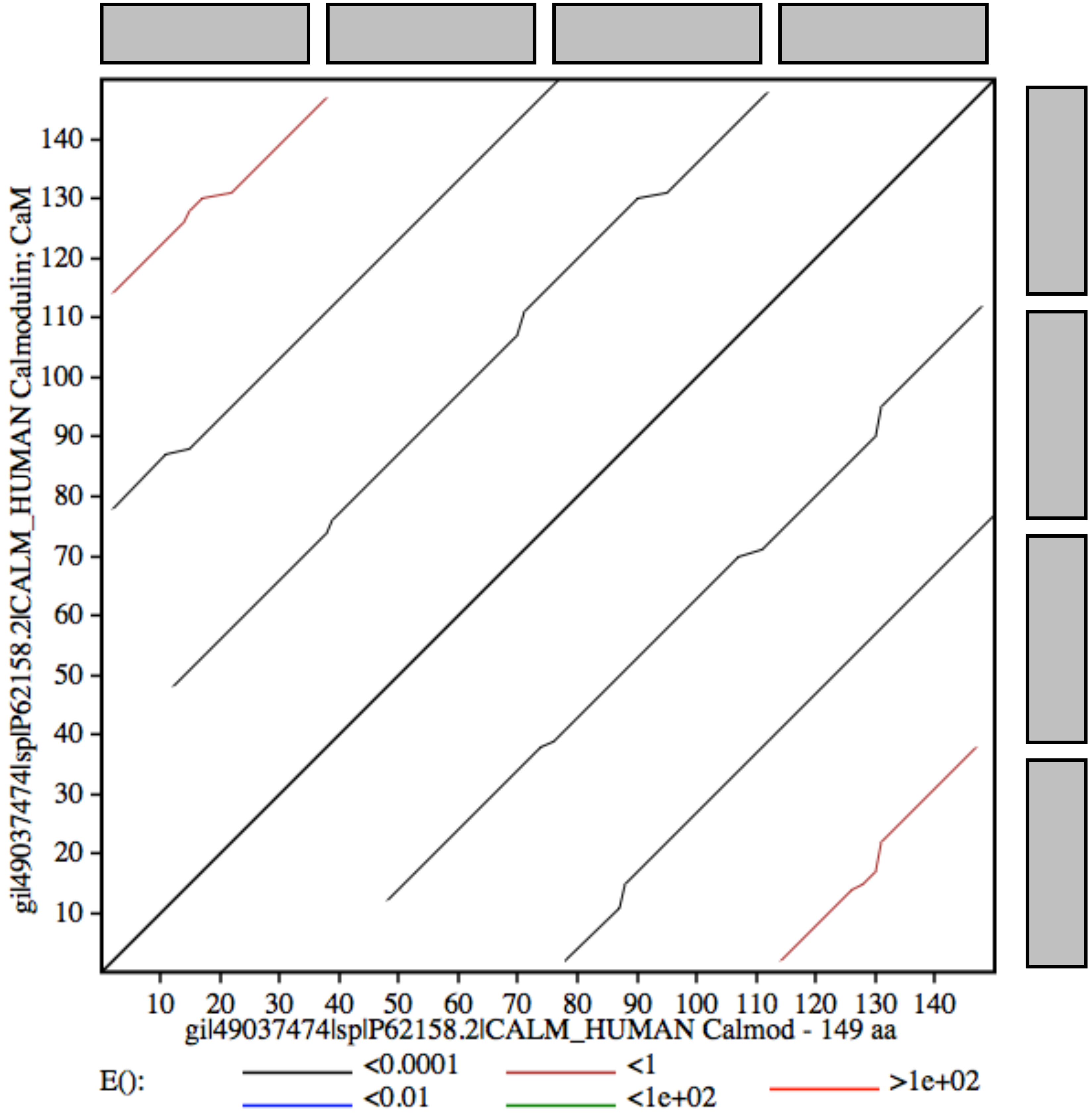
Finding Repeated Domains
Local Alignments

Calmodulin



```
>>>gi|49037474|sp|P62158.2|CALM_HUMAN, 149 aa vs TMP.q2 library

>>sp|P62158.2|CALM_HUMAN Calmodulin;  CaM
 Waterman-Eggert score: 220;  50.8 bits; E(1) <  1.1e-11
46.1% identity (73.7% similar) in 76 aa overlap (1-76:77-149)
Entrez Lookup  Re-search database  General re-search
               10        20        30        40        50        60        70
gi|490 MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARK
       : :  .::.:   .::: .:::::.: :.. ::  :: .::.. :. :...:: :.: ::.: ... ::. ::. :
sp|P62 MKDTDSEEEI---REAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK
         80          90       100       110       120       130       140

 Waterman-Eggert score: 181;  42.6 bits; E(1) <  3.2e-09
34.3% identity (64.8% similar) in 105 aa overlap (11-111:47-147)
Entrez Lookup  Re-search database  General re-search
               20        30         40        50        60           70        80
gi|490 AEFKEAFSLFDKDGDGTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEF---LTMMARKMKDTDSEEEI
       ::... ..  : ::.:::   :. :.: :.. .. .: :... .   : :::: :.  :.   .: ...:. : . .: :
sp|P62 AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMI
         50        60        70        80        90       100       110       120

         90       100       110
gi|490 REAFRVFDKDGNGYISAAELRHVMT
       :::    : ::.: ..  :. ..::
sp|P62 REA----DIDGDGQVNYEEFVQMMT
             130       140

 Waterman-Eggert score: 64;  18.2 bits; E(1) <  0.07
34.2% identity (71.1% similar) in 38 aa overlap (1-37:113-146)
Entrez Lookup  Re-search database  General re-search
               10         20        30
gi|490 MADQLTEEQIAEF-KEAFSLFDKDGDGTITTKELGTVM
       ....::.:.. :. .::    : :::: .. .:.  .:
sp|P62 LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
            120           130       140
```

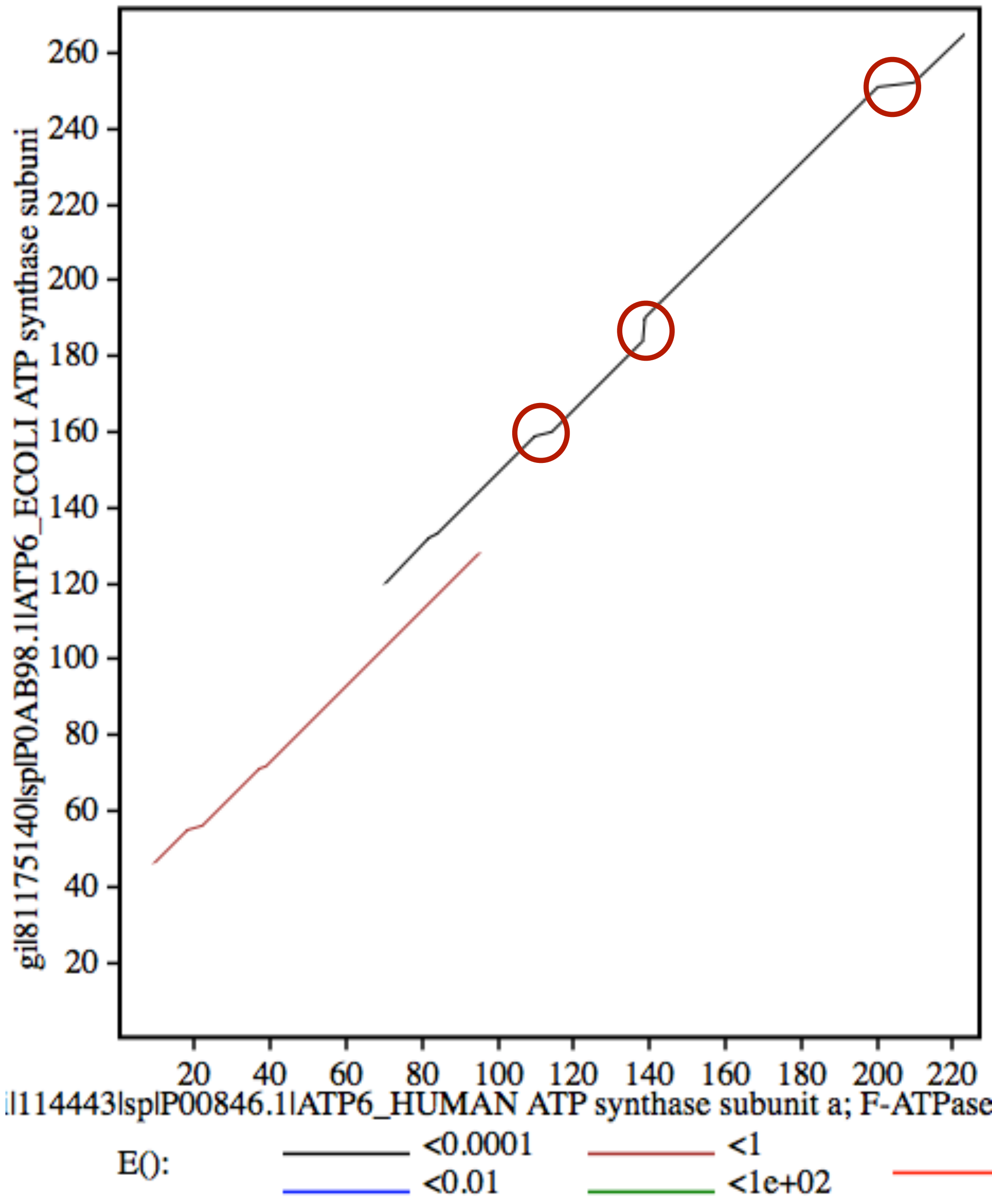
Finding Domains
Local Alignments

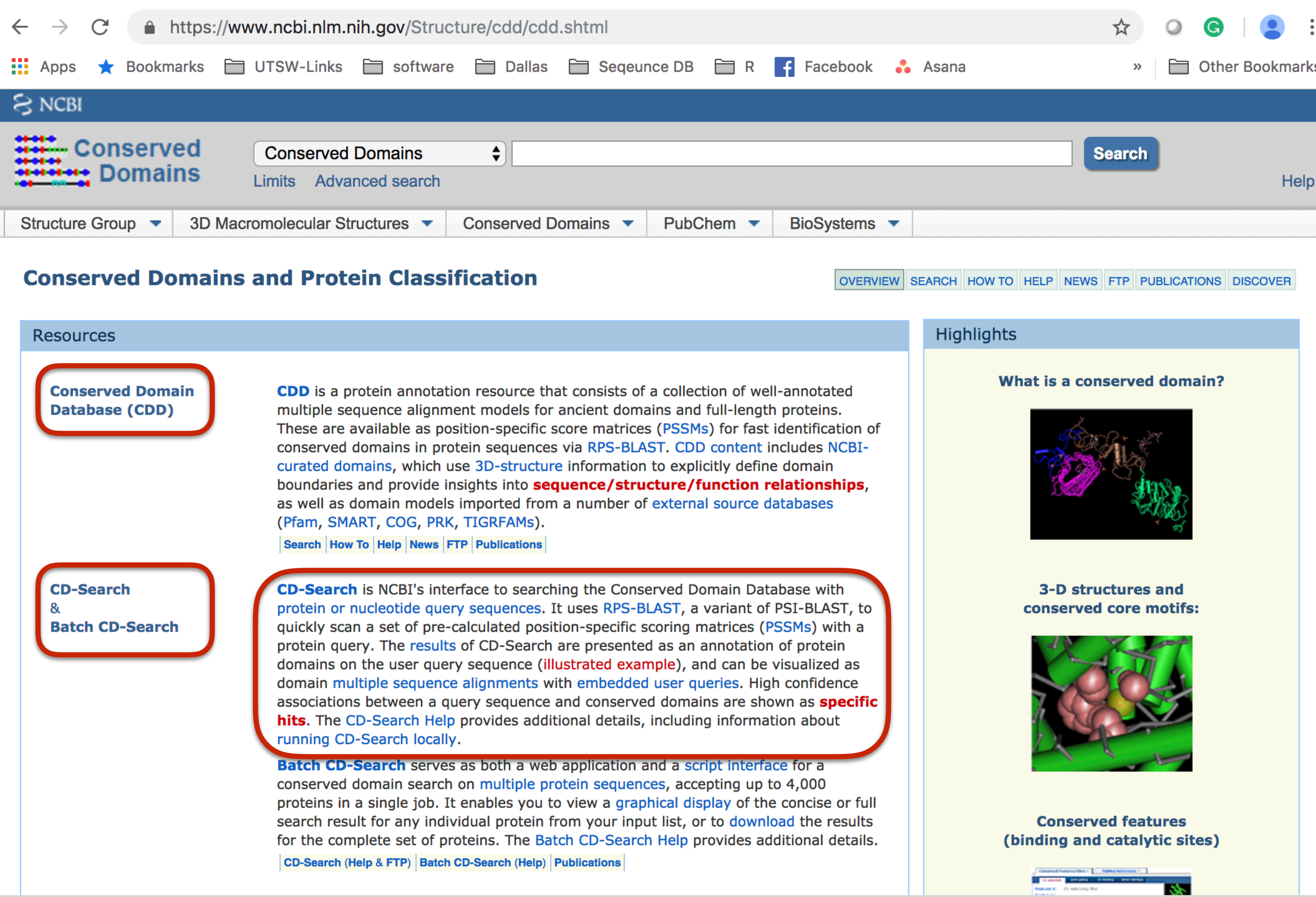


Southwestern Medical Center
Lyda Hill Department of Bioinformatics

Local Alignments

```
        10        20        30        40        50        60        70        80
gi|114 SFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLMLVSLIIFIATTNLLGL
       :..  ..:::   ....::  .   ...  . ... :. . .: ... ..  :.. :..  . . .......   ::. :
sp|P0A SMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVFLMNLMDL
           50           60        70         80        90       100       110       120

        90
gi|114 LPHSFTP
       :: .. :
sp|P0A LPIDLLP


       70        80        90       100       110       120       130            140
gi|114 SLMLVSLIIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPL-----IPMLVI
       .:. ..:. .::  ..:::     .:.......:.::. ..   ... : :   .... :  . :  :.     ::. .:
sp|P0A DLLPIDLLPYIAE-HVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLI
      120       130        140       150          160       170       180       190

           150       160       170       180       190       200       210       220
gi|114 IETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYL
       .: .::: .:..:..:: .:. ::.:.. ::..     :   :     :: :::.          .::..: .:. .::
sp|P0A LEGVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYL
          200       210       220       230       240       250                260
```





Conserved Domains and Protein Classification

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAMs).

Search | How To | Help | News | FTP | Publications

CD-Search
&
Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**. The CD-Search Help provides additional details, including information about running CD-Search locally.

Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 4,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

CD-Search (Help & FTP) | Batch CD-Search (Help) | Publications

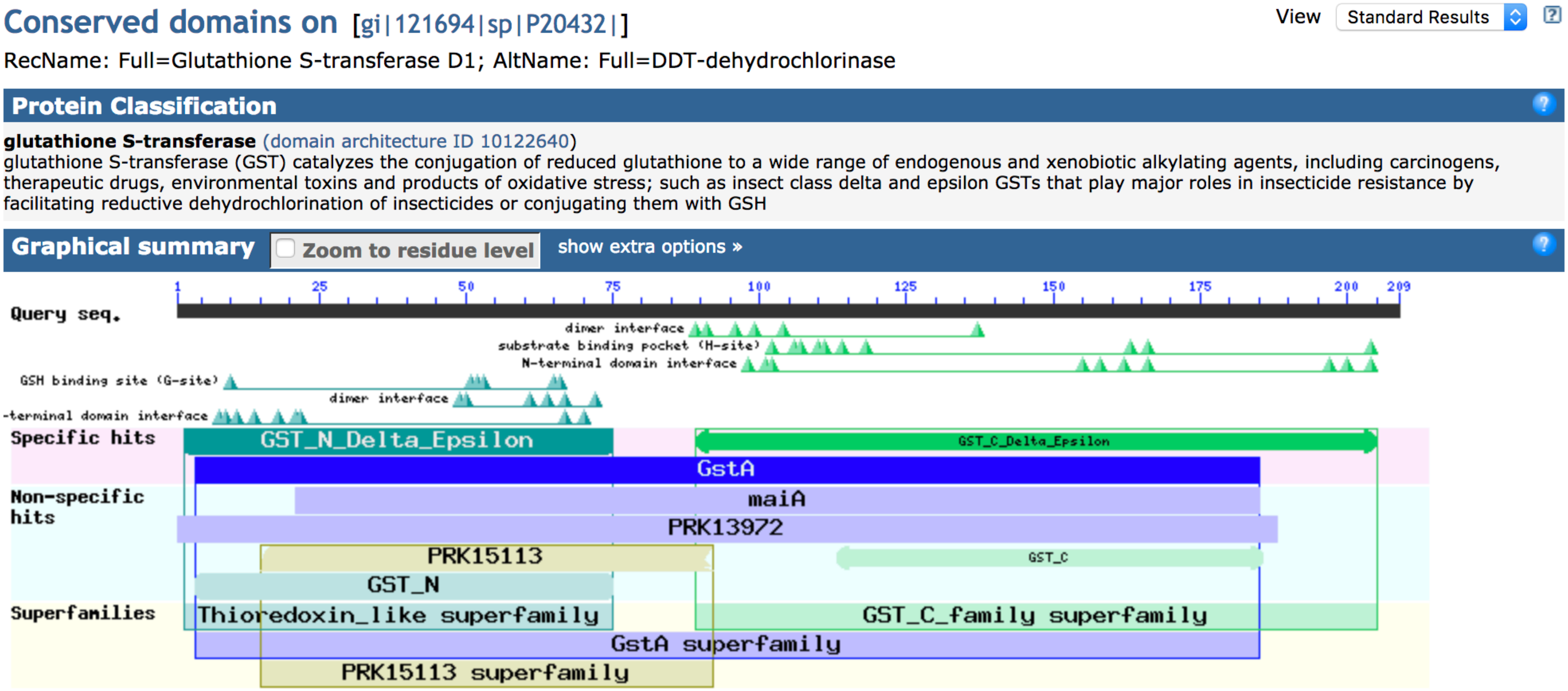
Highlights

What is a conserved domain?

3-D structures and conserved core motifs:

Conserved features (binding and catalytic sites)

Conserved Domains Database



Conserved domains on [gi|121694|sp|P20432|]

View Standard Results

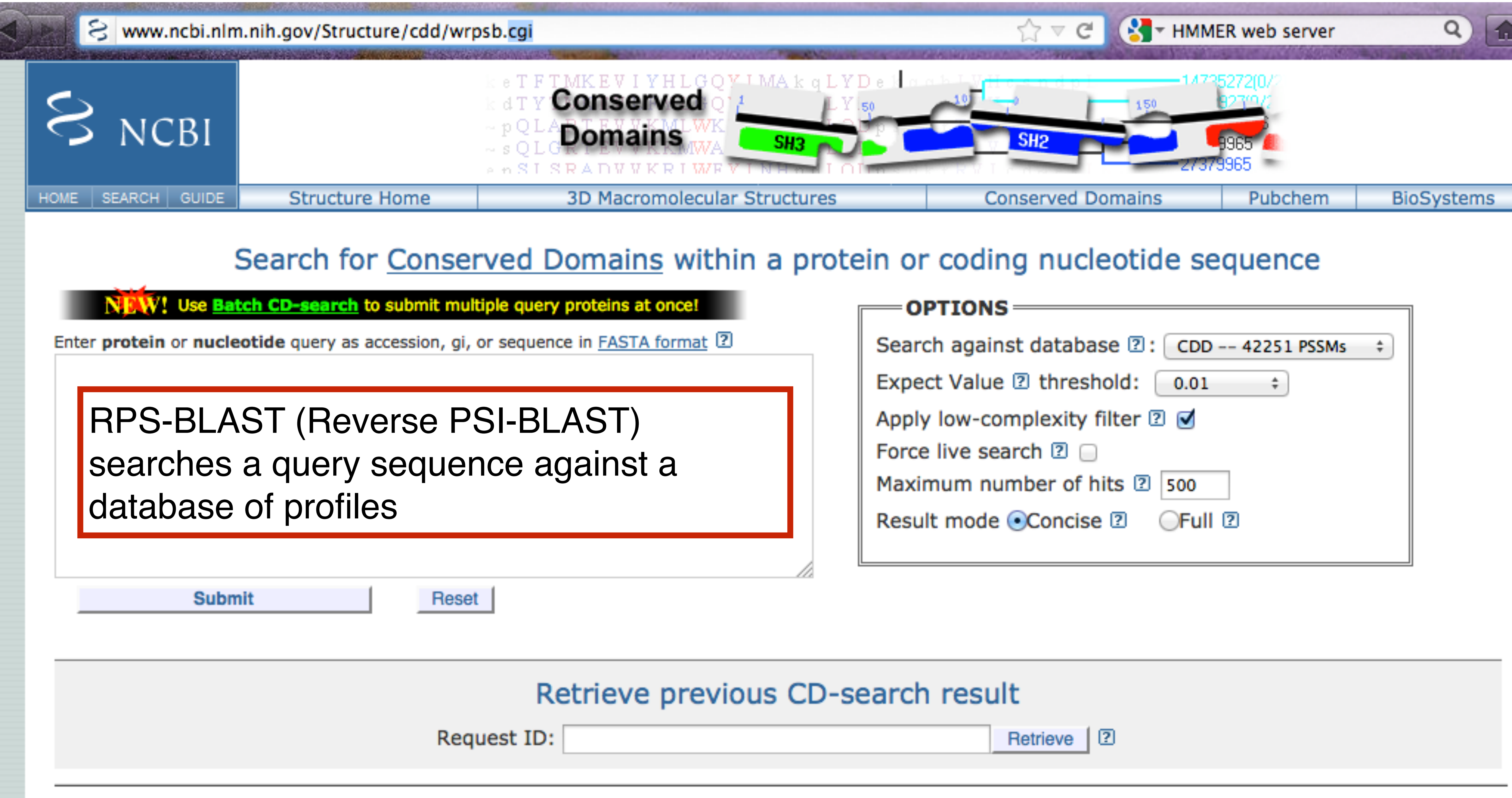
RecName: Full=Glutathione S-transferase D1; AltName: Full=DDT-dehydrochlorinase

Protein Classification

glutathione S-transferase (domain architecture ID 10122640)
glutathione S-transferase (GST) catalyzes the conjugation of reduced glutathione to a wide range of endogenous and xenobiotic alkylating agents, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress; such as insect class delta and epsilon GSTs that play major roles in insecticide resistance by facilitating reductive dehydrochlorination of insecticides or conjugating them with GSH

Graphical summary □ Zoom to residue level show extra options »

CD Search



Search for Conserved Domains within a protein or coding nucleotide sequence

NEW! Use Batch CD-search to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in FASTA format

RPS-BLAST (Reverse PSI-BLAST) searches a query sequence against a database of profiles

OPTIONS

Search against database: CDD -- 42251 PSSMs

Expect Value threshold: 0.01

Apply low-complexity filter ☑

Force live search □

Maximum number of hits 500

Result mode ● Concise ○ Full

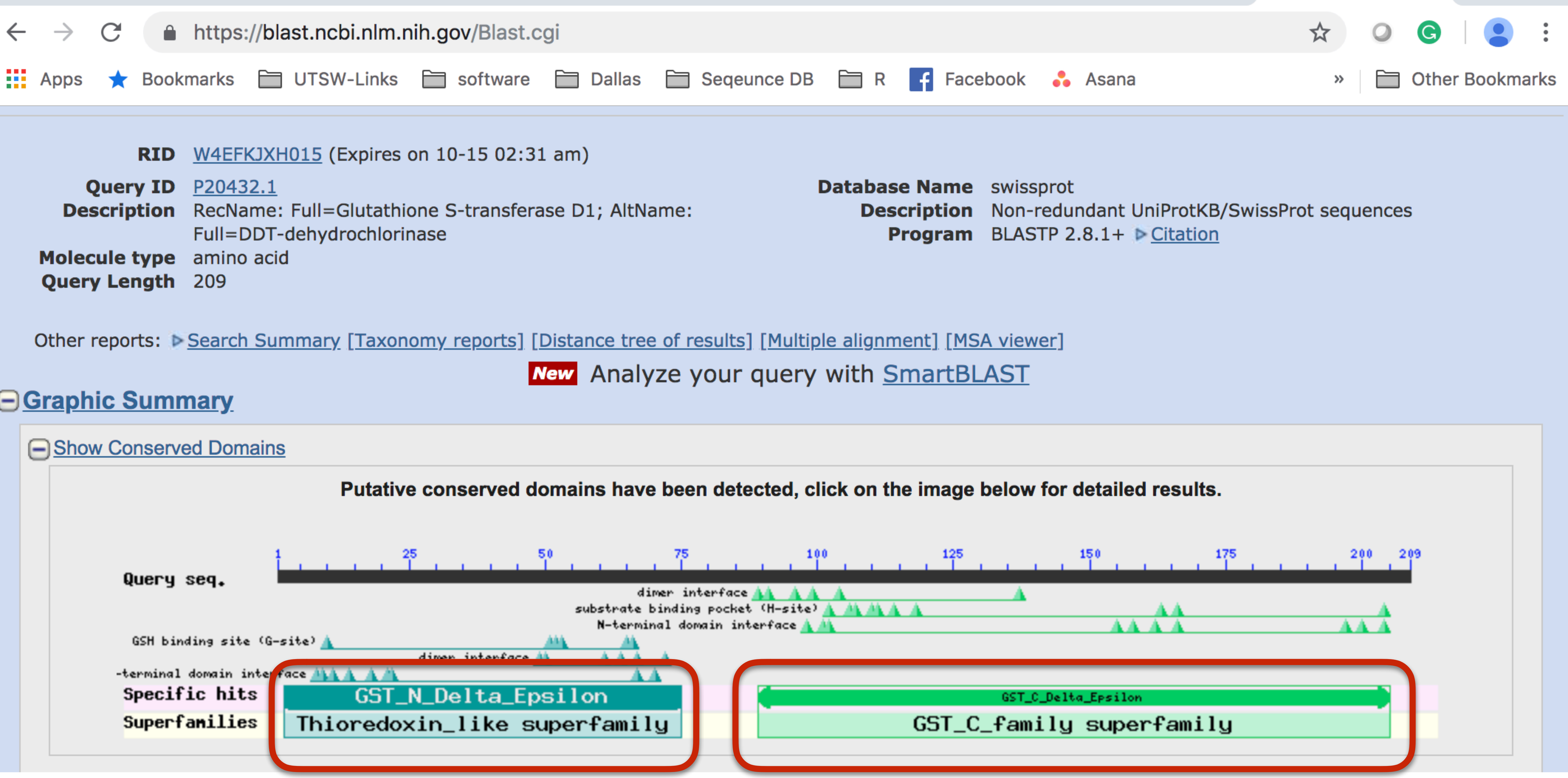
Submit Reset

Retrieve previous CD-search result

Request ID: Retrieve

UTSouthwestern Medical Center
Lyda Hill Department of Bioinformatics

Domain Search is Run with Web BLAST



RID W4EFKJXH015 (Expires on 10-15 02:31 am)

Query ID P20432.1
Description RecName: Full=Glutathione S-transferase D1; AltName: Full=DDT-dehydrochlorinase
Molecule type amino acid
Query Length 209

Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.8.1+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

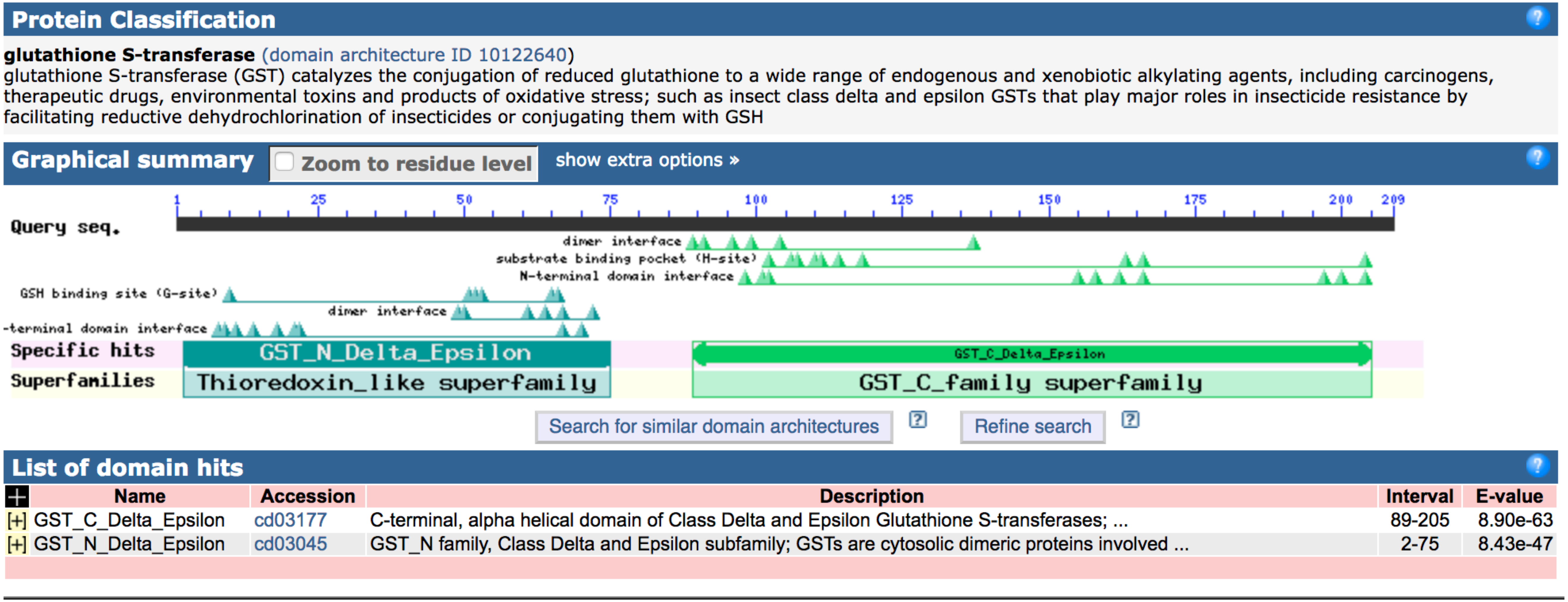
⊖Graphic Summary

⊖Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

UTSouthwestern
Medical Center
Lyda Hill Department of Bioinformatics

CD Search



Protein Classification

glutathione S-transferase (domain architecture ID 10122640)
glutathione S-transferase (GST) catalyzes the conjugation of reduced glutathione to a wide range of endogenous and xenobiotic alkylating agents, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress; such as insect class delta and epsilon GSTs that play major roles in insecticide resistance by facilitating reductive dehydrochlorination of insecticides or conjugating them with GSH

Graphical summary

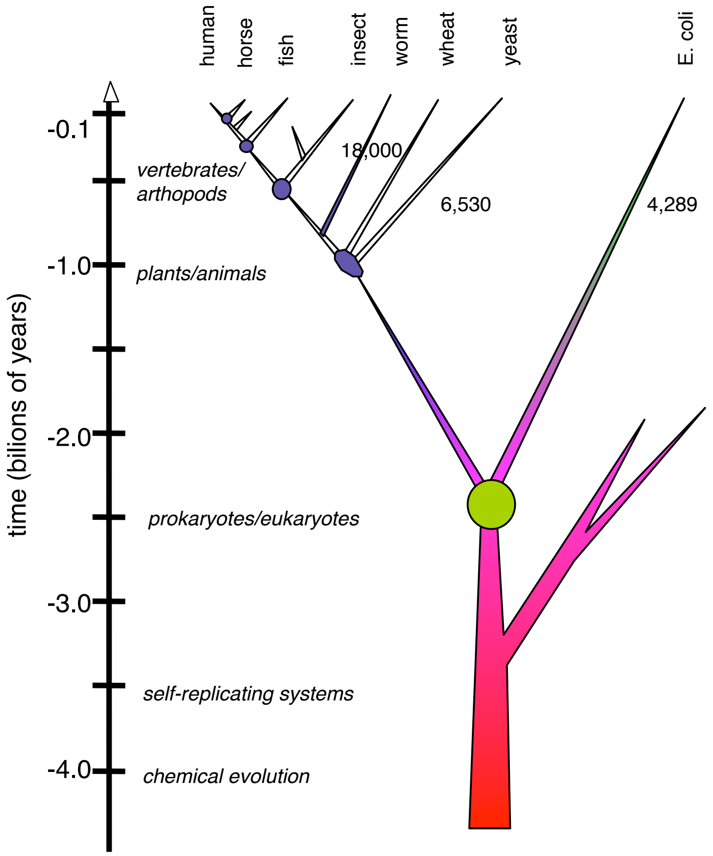
List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	GST_C_Delta_Epsilon	cd03177	C-terminal, alpha helical domain of Class Delta and Epsilon Glutathione S-transferases; ...	89-205	8.90e-63
[+]	GST_N_Delta_Epsilon	cd03045	GST_N family, Class Delta and Epsilon subfamily; GSTs are cytosolic dimeric proteins involved ...	2-75	8.43e-47

Homology through Transitivity

- What is a point specific scoring matrix?
- How can we use PSSMs in order to identify distance family members?

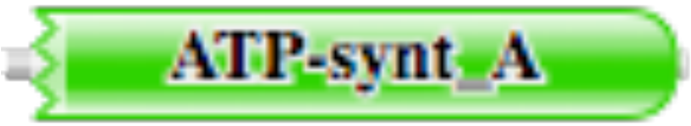
Homology through Transitivity

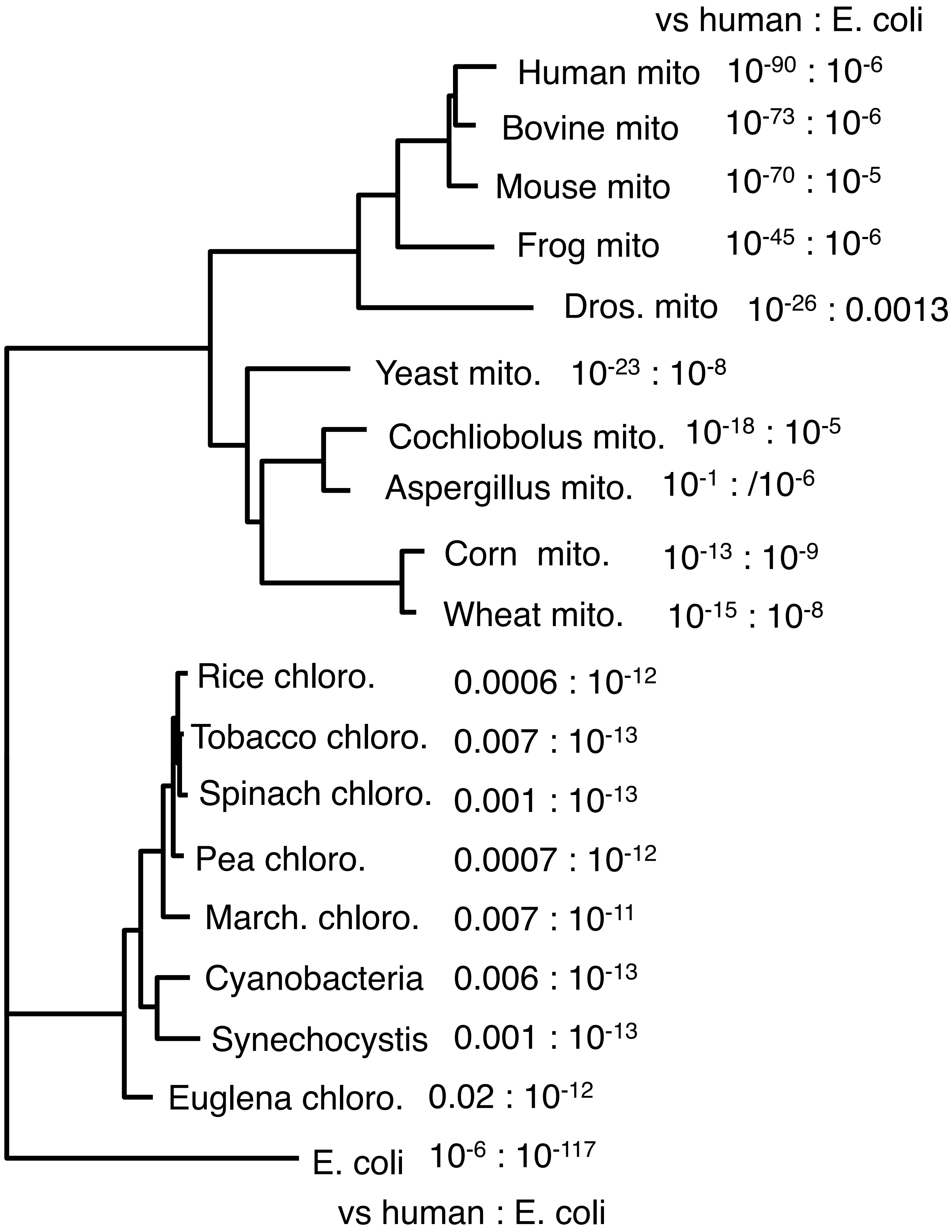


Protein A is Homologous to Proteins B
Protein B is Homologous to Protein C

Therefore:
Protein C is Homologous to Protein A

Homology is Transitive (in Protein Domains)

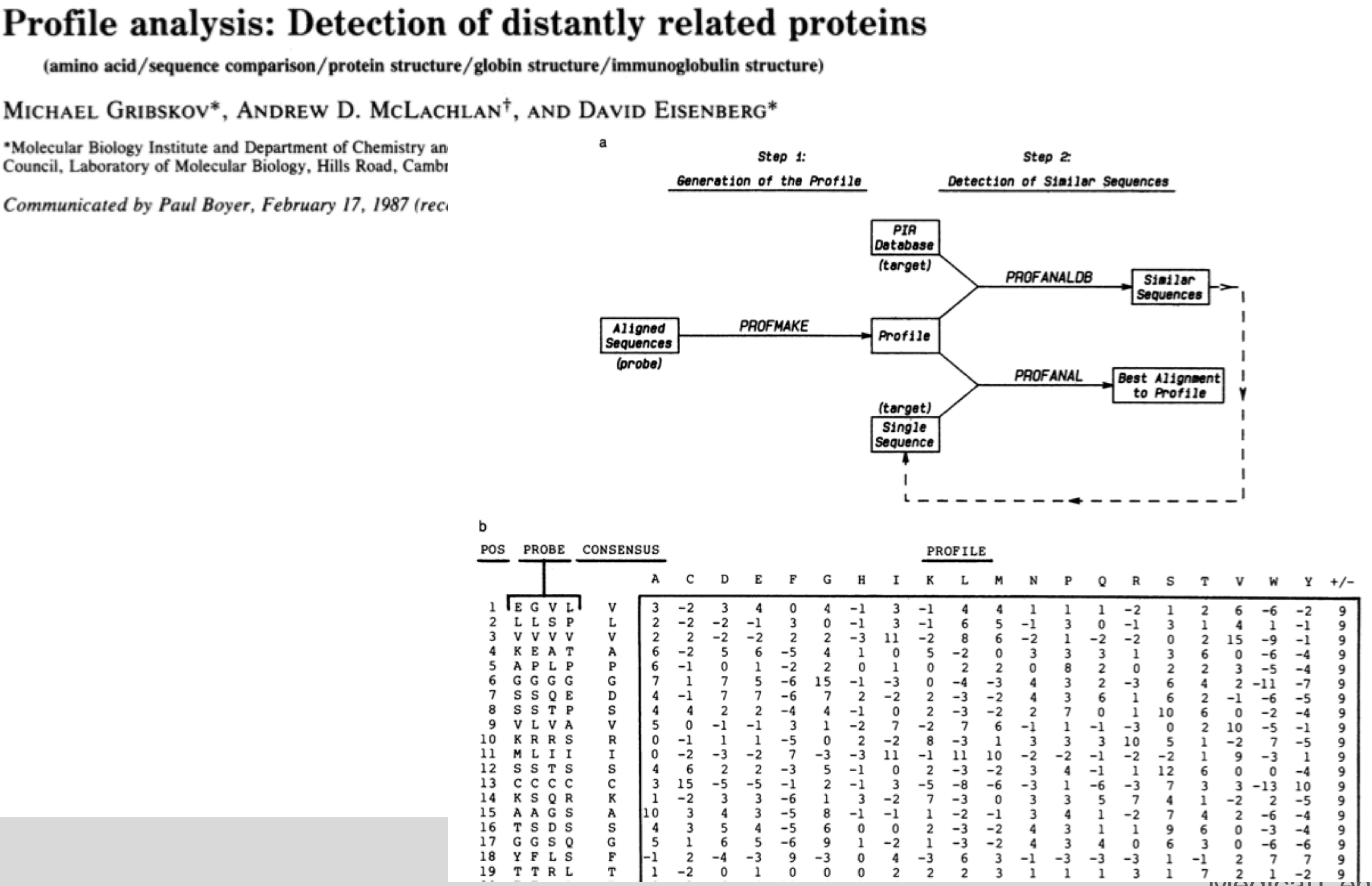




	vs human : E. coli
Human mito	10^{-90} : 10^{-6}
Bovine mito	10^{-73} : 10^{-6}
Mouse mito	10^{-70} : 10^{-5}
Frog mito	10^{-45} : 10^{-6}
Dros. mito	10^{-26} : 0.0013
Yeast mito.	10^{-23} : 10^{-8}
Cochliobolus mito.	10^{-18} : 10^{-5}
Aspergillus mito.	10^{-1} : /10^{-6}
Corn mito.	10^{-13} : 10^{-9}
Wheat mito.	10^{-15} : 10^{-8}
Rice chloro.	0.0006 : 10^{-12}
Tobacco chloro.	0.007 : 10^{-13}
Spinach chloro.	0.001 : 10^{-13}
Pea chloro.	0.0007 : 10^{-12}
March. chloro.	0.007 : 10^{-11}
Cyanobacteria	0.006 : 10^{-13}
Synechocystis	0.001 : 10^{-13}
Euglena chloro.	0.02 : 10^{-12}
E. coli	10^{-6} : 10^{-117}

vs human : E. coli

PSSM for detecting distance relationships



Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*

*Molecular Biology Institute and Department of Chemistry an
Council, Laboratory of Molecular Biology, Hills Road, Cambr

Communicated by Paul Boyer, February 17, 1987 (rec

a

Step 1:
Generation of the Profile

Step 2:
Detection of Similar Sequences

Aligned Sequences
(probe)

PROFMAKE

Profile

PIR Database
(target)

PROFANALDB

Similar Sequences

PROFANAL

Best Alignment to Profile

(target)
Single Sequence

b

POS	PROBE	CONSENSUS	PROFILE																				
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	−2	3	4	0	4	−1	3	−1	4	4	1	1	1	−2	1	2	6	−6	−2	9
2	L L S P	L	2	−2	−2	−1	3	0	−1	3	−1	6	5	−1	3	0	−1	3	1	4	1	−1	9
3	V V V V	V	2	2	−2	−2	2	2	−3	11	−2	8	6	−2	1	−2	−2	0	2	15	−9	−1	9
4	K E A T	A	6	−2	5	6	−5	4	1	0	5	−2	0	3	3	3	1	3	6	0	−6	−4	9
5	A P L P	P	6	−1	0	1	−2	2	0	1	0	2	2	0	8	2	0	2	2	3	−5	−4	9
6	G G G G	G	7	1	7	5	−6	15	−1	−3	0	−4	−3	4	3	2	−3	6	4	2	−11	−7	9
7	S S Q E	D	4	−1	7	7	−6	7	2	−2	2	−3	−2	4	3	6	1	6	2	−1	−6	−5	9
8	S S T P	S	4	4	2	2	−4	4	−1	0	2	−3	−2	2	7	0	1	10	6	0	−2	−4	9
9	V L V A	V	5	0	−1	−1	3	1	−2	7	−2	7	6	−1	1	−1	−3	0	2	10	−5	−1	9
10	K R R S	R	0	−1	1	1	−5	0	2	−2	8	−3	1	3	3	3	10	5	1	−2	7	−5	9
11	M L I I	I	0	−2	−3	−2	7	−3	−3	11	−1	11	10	−2	−2	−1	−2	−2	1	9	−3	1	9
12	S S T S	S	4	6	2	2	−3	5	−1	0	2	−3	−2	3	4	−1	1	12	6	0	0	−4	9
13	C C C C	C	3	15	−5	−5	−1	2	−1	3	−5	−8	−6	−3	1	−6	−3	7	3	3	−13	10	9
14	K S Q R	K	1	−2	3	3	−6	1	3	−2	7	−3	0	3	3	5	7	4	1	−2	2	−5	9
15	A A G S	A	10	3	4	3	−5	8	−1	−1	1	−2	−1	3	4	1	−2	7	4	2	−6	−4	9
16	T S D S	S	4	3	5	4	−5	6	0	0	2	−3	−2	4	3	1	1	9	6	0	−3	−4	9
17	G G S Q	G	5	1	6	5	−6	9	1	−2	1	−3	−2	4	3	4	0	6	3	0	−6	−6	9
18	Y F L S	F	−1	2	−4	−3	9	−3	0	4	−3	6	3	−1	−3	−3	−3	1	−1	2	7	7	9
19	T T R L	T	1	−2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	−2	9

Simple PSSM

TACGAT

TATAAT

TATAAT

GATACT

TATGAT

TATGTT

	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

TATACT

5	6	5	3	1	6	26

PSSMs

```
sp|O74706|EGLB_ASPNG      MKFQSTL--LLAAAAGSALAV-----------------PHGSGHKKRASVFEWFGSNESG
sp|Q96WQ8|EGLB_ASPKA      MKFQSTL--LLAAAAGSALAV-----------------PHGPGHKKRASVFEWFGSNESG
sp|P51529|MANA_STRLI      MR---NARSTLITTAGMAFAVLGLLFALAGPSAGRAEAAAGGIHVSNGRVVE--GNGSAF
sp|P22533|MANB_CALSA      MRLKTKIRKKWLSVLCTVVFLLNILFI-----ANVTILPKVGAATSNDGVVKI----DTS
                          *.   .      :.   .. :                 .      ..  *.:     .:

sp|O74706|EGLB_ASPNG      AEFGTNIPGVWGTDYIFPDPST--ISTLIGKGMNFFRVQFMMERLLPDSMTGSYDEEYLA
sp|Q96WQ8|EGLB_ASPKA      AEFGTNIPGVWGTDYIFPDPSA--ISTLIDKGMNFFRVQFMMERLLPDSMTGSYDEEYLA
sp|P51529|MANA_STRLI      VMRGVNHAYTW-----YPDRTGS-IADIAAKGANTVRVVL---------SSGGRWTKTSAS
sp|P22533|MANB_CALSA      TLIGTNHAHCW-----YRDRLDTALRGIRSWGMNSVRVVL---------SNGYRWTKIPAS
                          .  *.* .  *     : *     :  :  .* * .** :        *    : :   :
```

Score

% at Position

```
          A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V   A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
 1 M     -1 -2 -2 -3 -2 -1 -2 -3 -2  1  2 -2  6  0 -3 -2 -1 -2 -1  1    0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0  0.43 inf
 2 K     -1  5  0 -1 -3  1  0 -2 -1 -3 -2  4 -1 -3 -2 -1 -1 -3 -2 -3    0  58   0   0   0   0   0   0   0   0   0  42   0   0   0   0   0   0   0   0  0.60 inf
 3 F     -1 -2 -2 -2 -1 -2 -2 -2 -1  0  2 -2  1  4 -2 -1 -1  0  2  0    2   1   1   2   1   1   2   2   1   1  31   2   1  44   1   2   2   0   1   2  0.22 inf
 4 Q     -1  1  0  0 -2  4  1 -1  0 -2 -2  3 -1 -2 -1  0  0 -2 -1 -2    2   1   1   2   1  44   2   2   1   1   3  30   1   1   1   2   2   0   1   2  0.30 inf
 5 S      1 -1  0  0 -1  0  0 -1 -1 -1 -1  0 -1 -2  0  3  3 -2 -1 -1    2   1   1   2   1   1   2   2   1   1   3   2   1   1   1  45  30   0   1   2  0.24 inf
 6 T     -1  0  3  0 -2  0  0 -1 -1 -2 -2  2 -1 -3 -1  1  3 -3 -2 -1    0   0  29   0   0   0   0   0   0   0   0  29   0   0   0   0  42   0   0   0  0.32 inf
 7 L      1 -2 -3 -3 -1 -2 -2 -2 -3  2  3 -2  1  0 -2 -1 -1 -2 -1  1   29   0   0   0   0   0   0   0   0  29  42   0   0   0   0   0   0   0   0   0  0.21 inf
 8 L     -1  0 -1 -2 -2  0 -1 -2 -2  0  2  2  1 -1 -2  0  2 -2 -2  0    0   0   0   0   0   0   0   0   0   0  42  29   0   0   0   0  29   0   0   0  0.15 inf
 9 L     -2 -2 -4 -4 -2 -2 -3 -3 -3  1  3 -3  1  1 -3 -3 -2  7  0  0    0   0   0   0   0   0   0   0   0   0  71   0   0   0   0   0   0  29   0   0  0.68 inf
10 A      2 -2 -3 -3 -1 -2 -2 -2 -2  2  2 -2  1 -1 -2  0 -1 -2 -2  1   42   0   0   0   0   0   0   0   0  29  29   0   0   0   0   0   0   0   0   0  0.18 inf
11 A      3 -1  0 -1 -1 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  2  3 -3 -2 -1   42   0   0   0   0   0   0   0   0   0   0   0   0   0   0  29  29   0   0   0  0.32 inf
12 A      2 -2 -1 -2 -1 -1 -1 -1 -2  0 -1 -1  0 -2 -1  1  2 -3 -2  2   42   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  29   0   0  29  0.21 inf
13 A      3 -2 -2 -2 -1 -1 -1 -1 -2  0  1 -1  0 -1 -1  0  0 -2 -2  0   71   0   0   0   0   0   0   0   0   0  29   0   0   0   0   0   0   0   0   0  0.24 inf
14 G      0 -3 -1 -2  5 -2 -3  5 -2 -3 -3 -2 -2 -3 -2 -1 -1 -3 -3 -2    0   0   0   0  29   0   0  71   0   0   0   0   0   0   0   0   0   0   0   0  0.79 inf
15 S      0 -1  0 -1 -1  0 -1 -1 -1 -1  0 -1  3 -2 -1  3  3 -2 -2  0    0   0   0   0   0   0   0   0   0   0   0   0  29   0   0  42  29   0   0   0  0.23 inf
16 A      3 -2 -2 -2 -1 -1 -1 -1 -2  0 -1 -1  0 -2 -1  1  0 -3 -2  2   71   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  29  0.27 inf
17 L     -1 -3 -3 -4 -1 -3 -3 -4 -2  2  3 -3  1  3 -3 -2 -1 -1  1  2    0   0   0   0   0   0   0   0   0   0  42   0   0  29   0   0   0   0   0  29  0.31 inf
18 A      3 -2 -2 -2 -1 -1 -1 -1 -2 -1 -1 -1 -1  3 -2  0 -1 -1  0  0   71   0   0   0   0   0   0   0   0   0   0   0   0  29   0   0   0   0   0   0  0.27 inf
19 V     -1 -3 -3 -3 -1 -2 -3 -3 -3  2  2 -2  1  0 -3 -2  0 -3 -1  3    0   0   0   0   0   0   0   0   0   0  29   0   0   0   0   0   0   0   0  71  0.33 inf
 20 P     2 -2 -2 -2 -2 -1 -1 -1 -2 -2 -3 -1 -2 -3  7  0 -1 -4 -3 -2   29   0   0   0   0   0   0   0   0   0   0   0   0   0  71   0   0   0   0   0  1.33 i
```

UTSouthwestern Medical Center
Lyda Hill Department of Bioinformatics

Where Pairwise Scores Come From

“probability of A given an A”" the observed probability of seeing an A" aligned to an A in real alignments"

$$\text{score}(AA)=\log\frac{P(A|A)}{f(A)}$$

frequency of A” the expected frequency of A in any sequence

$$Sc(AA) = \log_2\frac{0.64}{0.04} = +4$$

$$Sc(AE) = \log_2\frac{0.01}{0.04} = -2$$

Where Profile Scores Should Come From

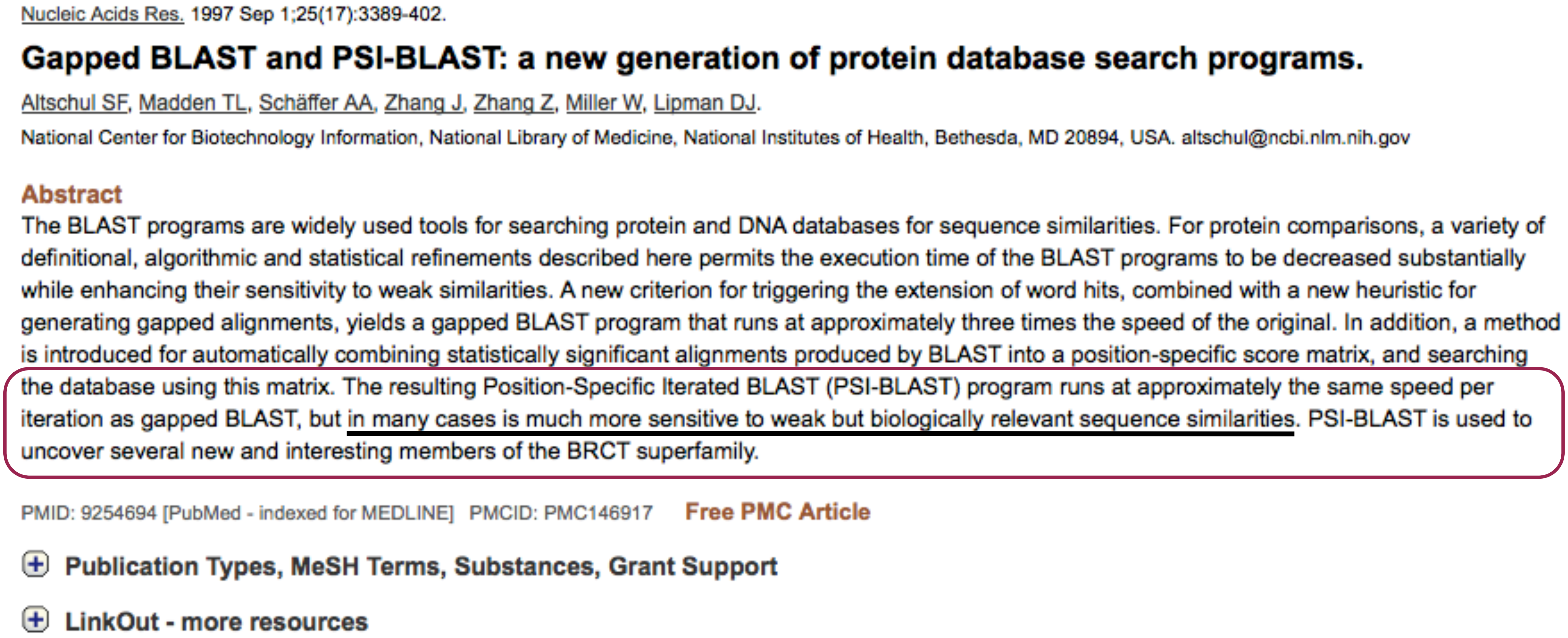
$$\text{score}(A|x) = \log \frac{P(A|\text{position } x)}{f(A)}$$

“probability of A at position x”" the observed probability of seeing an A in the consensus column X

$$Sc(A|6) = \log_2 \frac{1.00}{0.04} = +4.6 \qquad Sc(A|5) = \log_2 \frac{0.04}{0.04} = 0$$

$$Sc(N|6) = \log_2 \frac{0.00}{0.06} = \text{-inf} \qquad Sc(N|5) = \log_2 \frac{0.06}{0.06} = 0$$

what about position-specific gap penalties?
how to estimate parameters from small numbers of observations?



Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. altschul@ncbi.nlm.nih.gov

Abstract

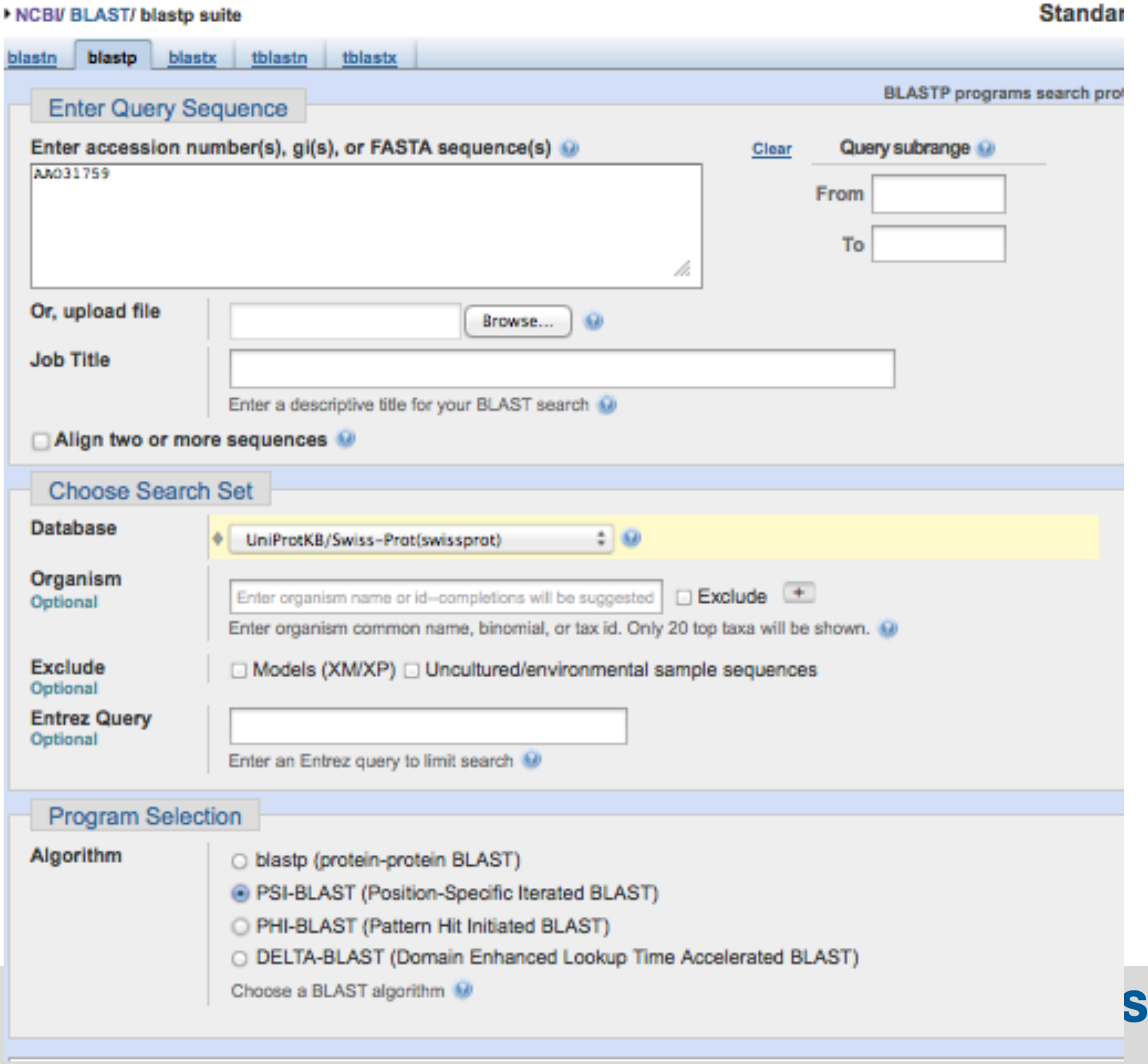
The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the BLAST programs to be decreased substantially while enhancing their sensitivity to weak similarities. A new criterion for triggering the extension of word hits, combined with a new heuristic for generating gapped alignments, yields a gapped BLAST program that runs at approximately three times the speed of the original. In addition, a method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as gapped BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. PSI-BLAST is used to uncover several new and interesting members of the BRCT superfamily.

PMID: 9254694 [PubMed - indexed for MEDLINE] PMCID: PMC146917 Free PMC Article

⊕ Publication Types, MeSH Terms, Substances, Grant Support

⊕ LinkOut - more resources

PSI-BLAST uses PSSMs to Find Distant Homologs



NCBI/ BLAST/ blastp suite

Standar

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search pro

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
AAO31759
```

Clear

Query subrange

From

To

Or, upload file [ ] Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

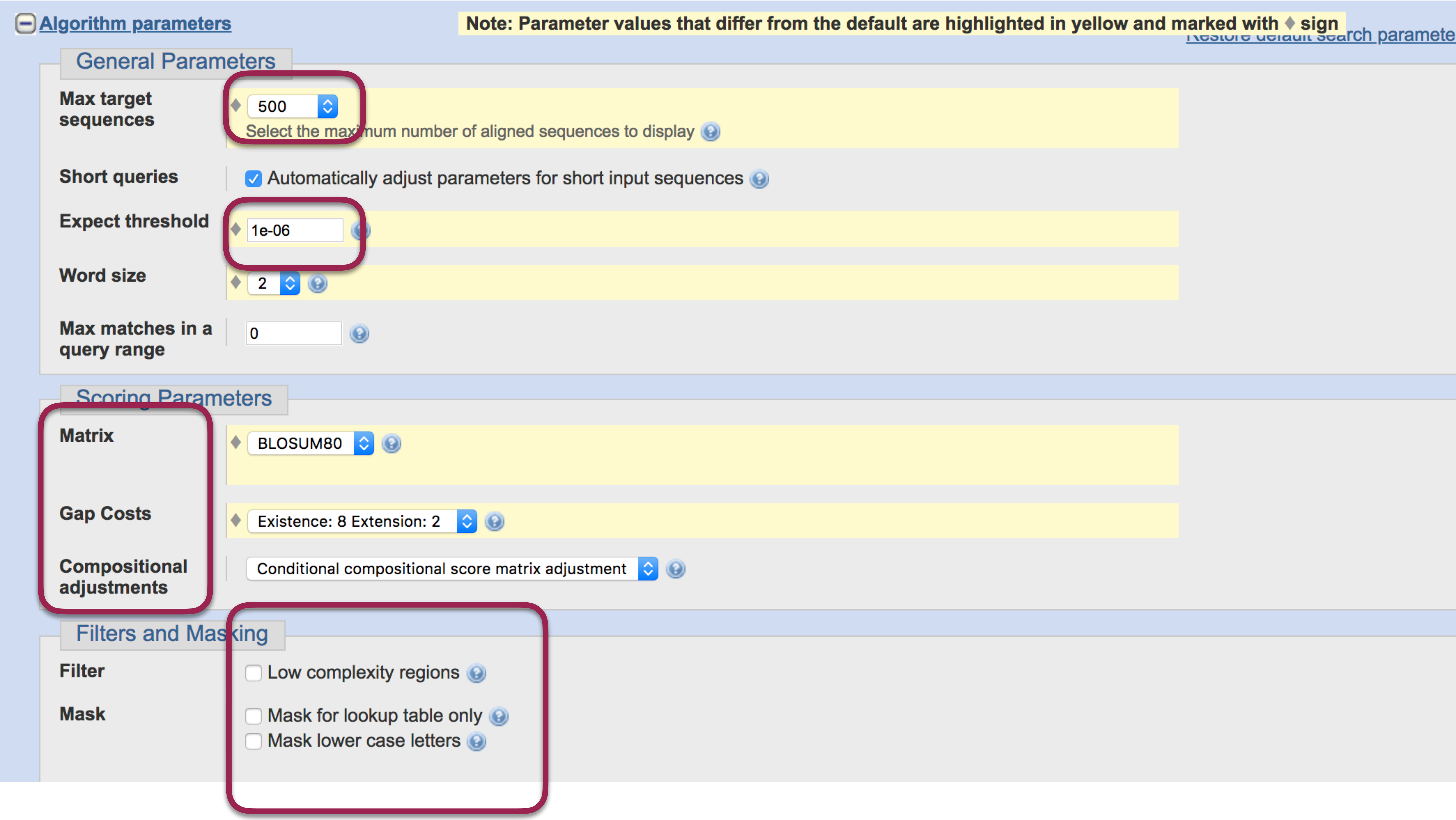
○ blastp (protein-protein BLAST)

● PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm



⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Restore default search paramete

General Parameters

Max target sequences: ♦ 500
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: ♦ 1e-06

Word size: ♦ 2

Max matches in a query range: 0

Scoring Parameters

Matrix: ♦ BLOSUM80

Gap Costs: ♦ Existence: 8 Extension: 2

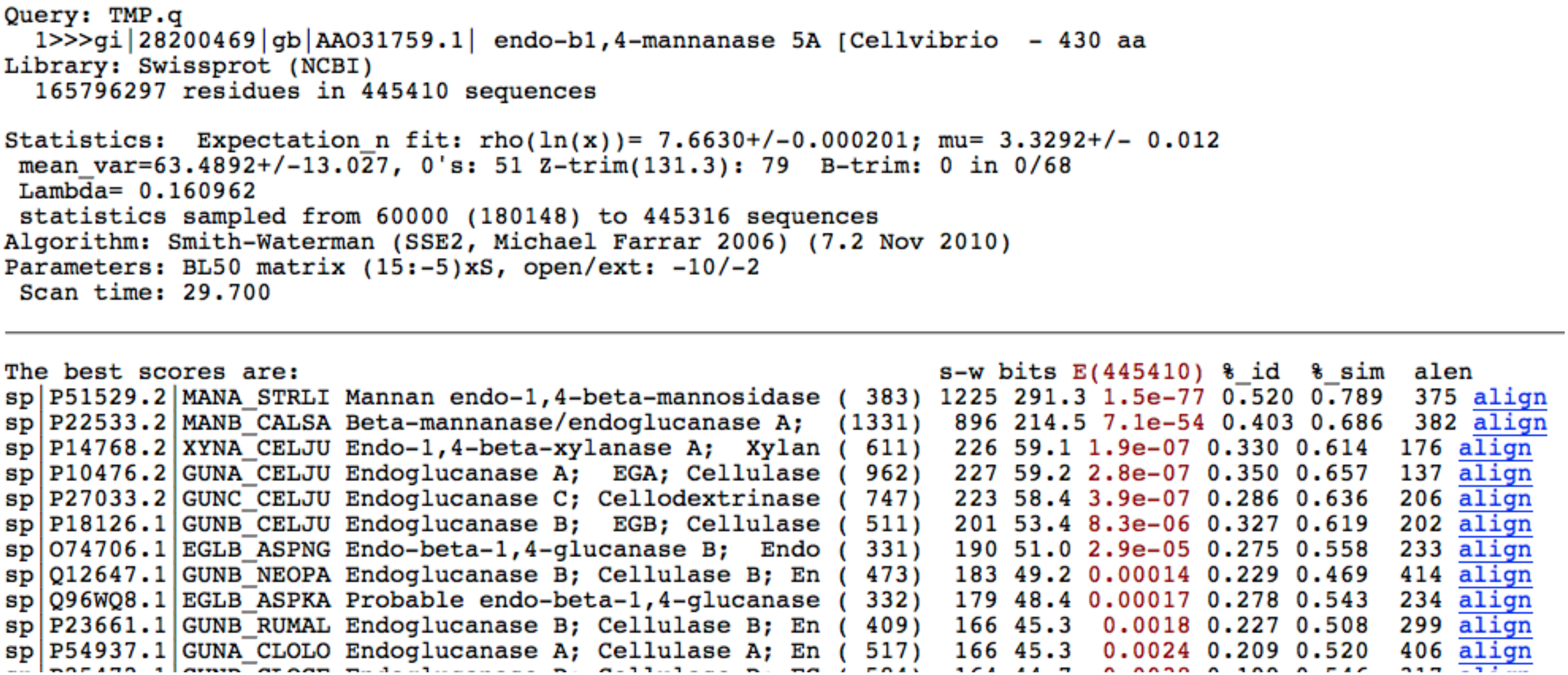
Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

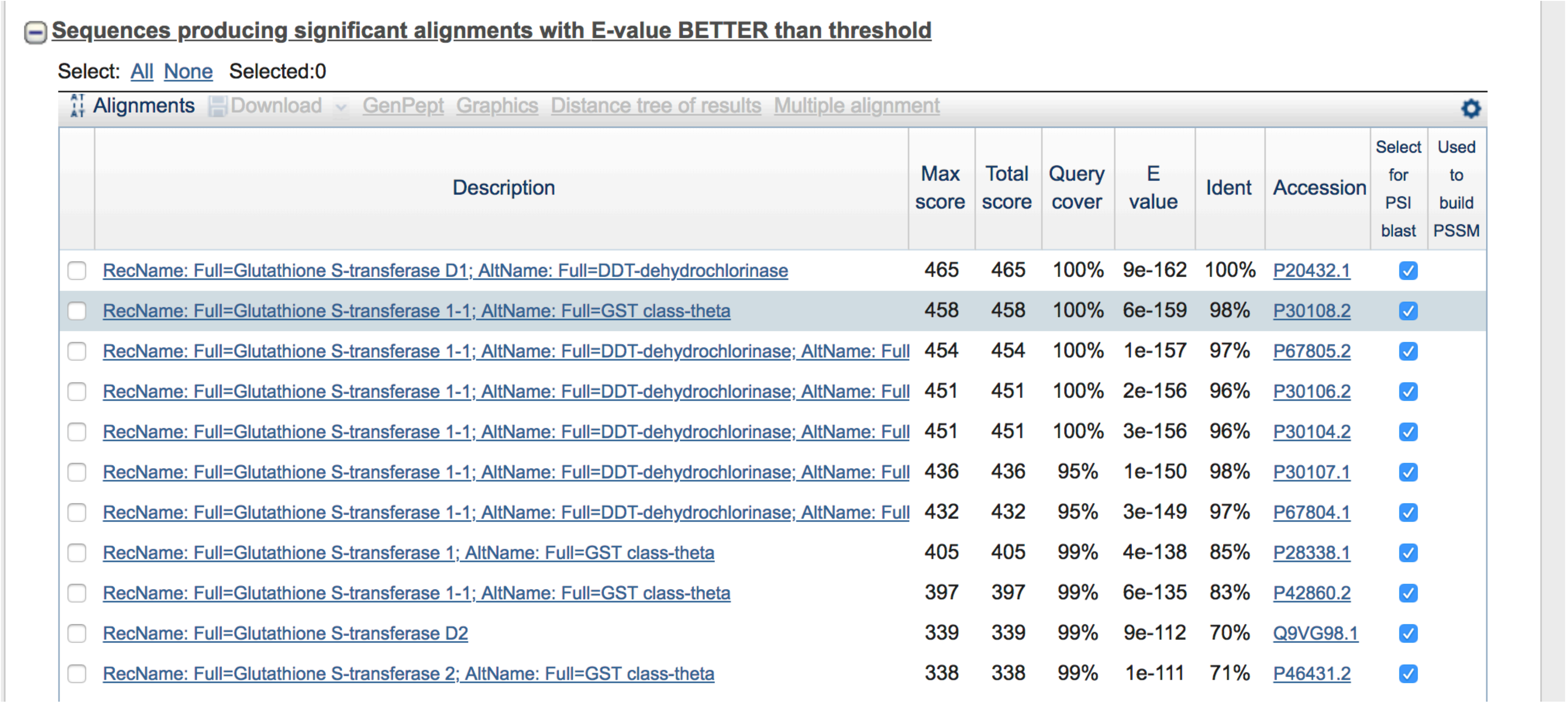
A SmithWaterman Search



```
Query: TMP.q
  1>>>gi|28200469|gb|AAO31759.1| endo-b1,4-mannanase 5A [Cellvibrio  - 430 aa
Library: Swissprot (NCBI)
  165796297 residues in 445410 sequences

Statistics:  Expectation_n fit: rho(ln(x))= 7.6630+/-0.000201; mu= 3.3292+/- 0.012
 mean_var=63.4892+/-13.027, 0's: 51 Z-trim(131.3): 79  B-trim: 0 in 0/68
 Lambda= 0.160962
 statistics sampled from 60000 (180148) to 445316 sequences
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2010)
Parameters: BL50 matrix (15:-5)xS, open/ext: -10/-2
 Scan time: 29.700

The best scores are:                                          s-w bits E(445410) %_id  %_sim  alen
sp|P51529.2|MANA_STRLI Mannan endo-1,4-beta-mannosidase ( 383) 1225 291.3 1.5e-77 0.520 0.789  375 align
sp|P22533.2|MANB_CALSA Beta-mannanase/endoglucanase A;  (1331)  896 214.5 7.1e-54 0.403 0.686  382 align
sp|P14768.2|XYNA_CELJU Endo-1,4-beta-xylanase A;  Xylan ( 611)  226 59.1 1.9e-07 0.330 0.614  176 align
sp|P10476.2|GUNA_CELJU Endoglucanase A;  EGA; Cellulase ( 962)  227 59.2 2.8e-07 0.350 0.657  137 align
sp|P27033.2|GUNC_CELJU Endoglucanase C; Cellodextrinase ( 747)  223 58.4 3.9e-07 0.286 0.636  206 align
sp|P18126.1|GUNB_CELJU Endoglucanase B;  EGB; Cellulase ( 511)  201 53.4 8.3e-06 0.327 0.619  202 align
sp|O74706.1|EGLB_ASPNG Endo-beta-1,4-glucanase B;  Endo ( 331)  190 51.0 2.9e-05 0.275 0.558  233 align
sp|Q12647.1|GUNB_NEOPA Endoglucanase B; Cellulase B; En ( 473)  183 49.2 0.00014 0.229 0.469  414 align
sp|Q96WQ8.1|EGLB_ASPKA Probable endo-beta-1,4-glucanase ( 332)  179 48.4 0.00017 0.278 0.543  234 align
sp|P23661.1|GUNB_RUMAL Endoglucanase B; Cellulase B; En ( 409)  166 45.3  0.0018 0.227 0.508  299 align
sp|P54937.1|GUNA_CLOLO Endoglucanase A; Cellulase A; En ( 517)  166 45.3  0.0024 0.209 0.520  406 align
```

UTSouthwestern Medical Center
Lyda Hill Department of Bioinformatics

A PSI-BLAST First Iteration

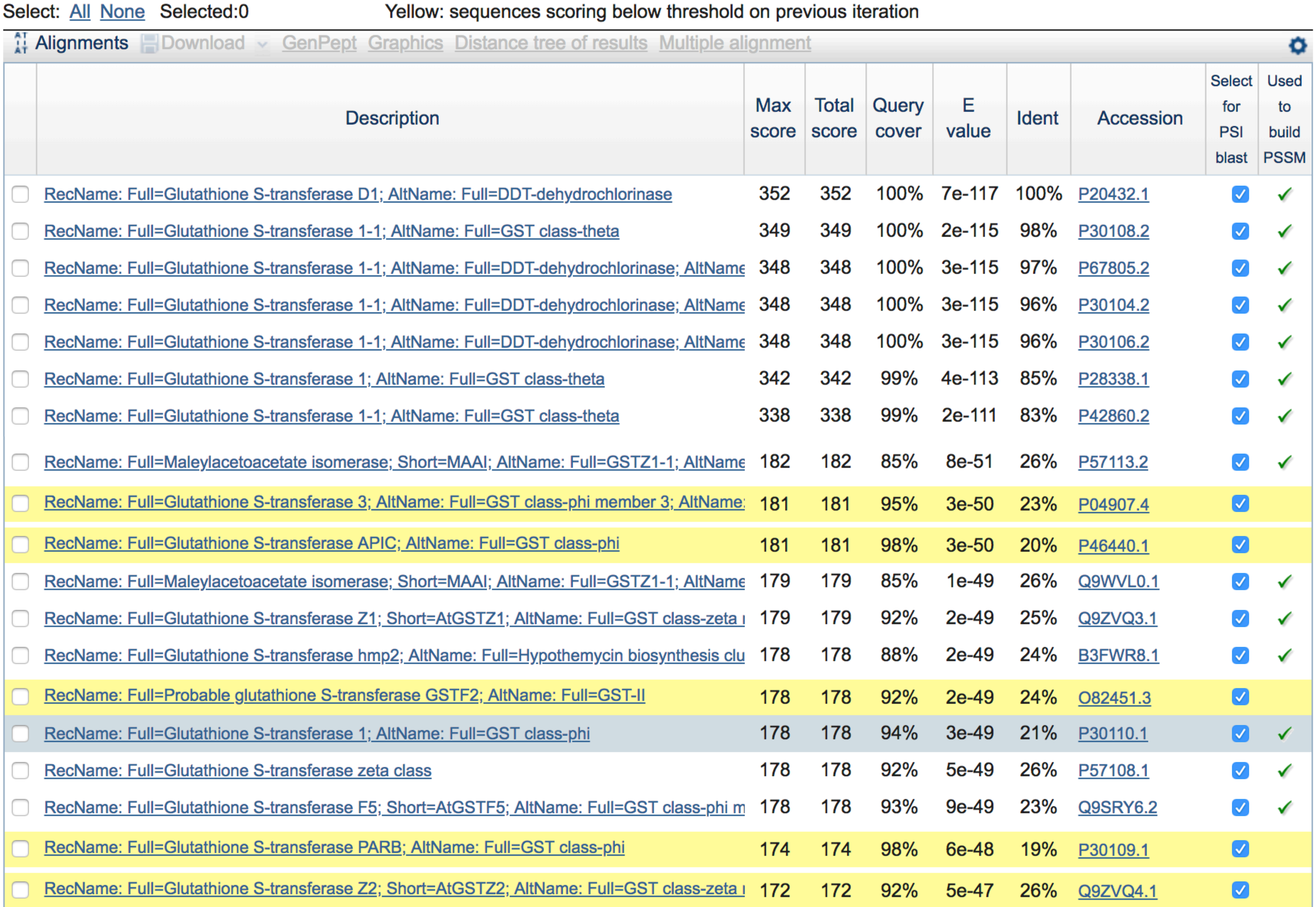


Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
□	RecName: Full=Glutathione S-transferase D1; AltName: Full=DDT-dehydrochlorinase	465	465	100%	9e-162	100%	P20432.1	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=GST class-theta	458	458	100%	6e-159	98%	P30108.2	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName: Full	454	454	100%	1e-157	97%	P67805.2	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName: Full	451	451	100%	2e-156	96%	P30106.2	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName: Full	451	451	100%	3e-156	96%	P30104.2	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName: Full	436	436	95%	1e-150	98%	P30107.1	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName: Full	432	432	95%	3e-149	97%	P67804.1	☑	
□	RecName: Full=Glutathione S-transferase 1; AltName: Full=GST class-theta	405	405	99%	4e-138	85%	P28338.1	☑	
□	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=GST class-theta	397	397	99%	6e-135	83%	P42860.2	☑	
□	RecName: Full=Glutathione S-transferase D2	339	339	99%	9e-112	70%	Q9VG98.1	☑	
□	RecName: Full=Glutathione S-transferase 2; AltName: Full=GST class-theta	338	338	99%	1e-111	71%	P46431.2	☑	

PSI-BLAST Second Iteration

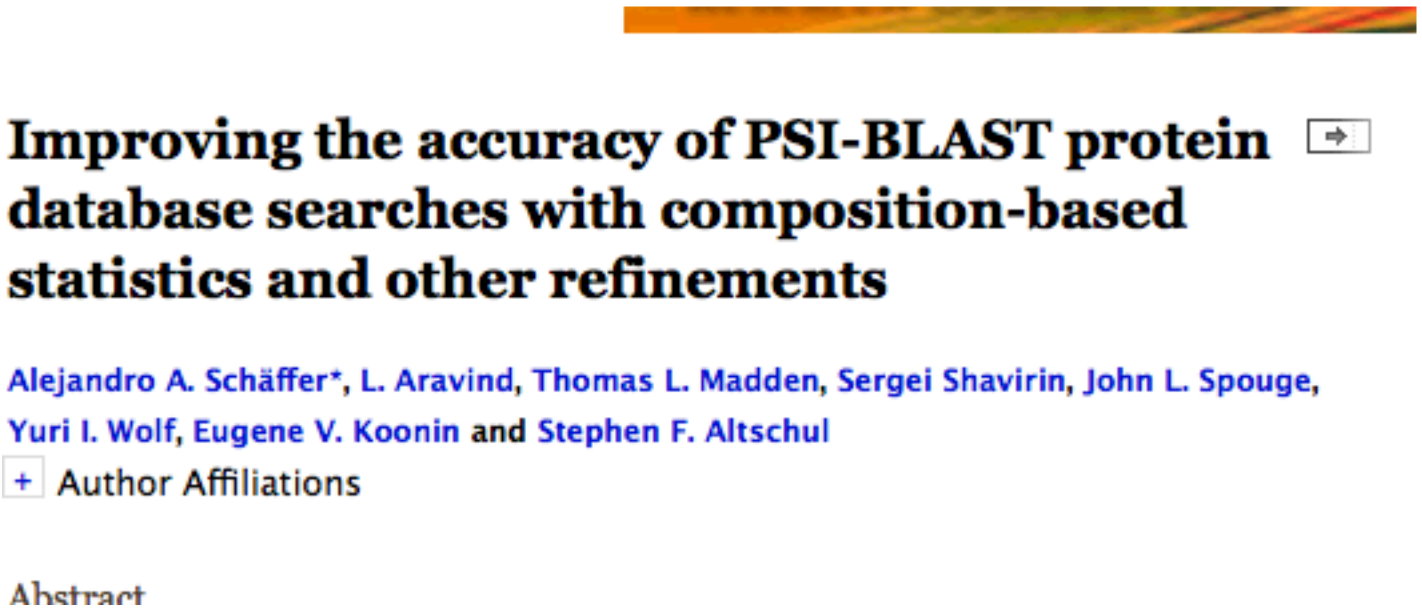


Select: All None Selected:0 Yellow: sequences scoring below threshold on previous iteration

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☐	RecName: Full=Glutathione S-transferase D1; AltName: Full=DDT-dehydrochlorinase	352	352	100%	7e-117	100%	P20432.1	☑	✓
☐	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=GST class-theta	349	349	100%	2e-115	98%	P30108.2	☑	✓
☐	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName	348	348	100%	3e-115	97%	P67805.2	☑	✓
☐	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName	348	348	100%	3e-115	96%	P30104.2	☑	✓
☐	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=DDT-dehydrochlorinase; AltName	348	348	100%	3e-115	96%	P30106.2	☑	✓
☐	RecName: Full=Glutathione S-transferase 1; AltName: Full=GST class-theta	342	342	99%	4e-113	85%	P28338.1	☑	✓
☐	RecName: Full=Glutathione S-transferase 1-1; AltName: Full=GST class-theta	338	338	99%	2e-111	83%	P42860.2	☑	✓
☐	RecName: Full=Maleylacetoacetate isomerase; Short=MAAI; AltName: Full=GSTZ1-1; AltName	182	182	85%	8e-51	26%	P57113.2	☑	✓
☐	RecName: Full=Glutathione S-transferase 3; AltName: Full=GST class-phi member 3; AltName:	181	181	95%	3e-50	23%	P04907.4	☑	
☐	RecName: Full=Glutathione S-transferase APIC; AltName: Full=GST class-phi	181	181	98%	3e-50	20%	P46440.1	☑	
☐	RecName: Full=Maleylacetoacetate isomerase; Short=MAAI; AltName: Full=GSTZ1-1; AltName	179	179	85%	1e-49	26%	Q9WVL0.1	☑	✓
☐	RecName: Full=Glutathione S-transferase Z1; Short=AtGSTZ1; AltName: Full=GST class-zeta ı	179	179	92%	2e-49	25%	Q9ZVQ3.1	☑	✓
☐	RecName: Full=Glutathione S-transferase hmp2; AltName: Full=Hypothemycin biosynthesis clu	178	178	88%	2e-49	24%	B3FWR8.1	☑	✓
☐	RecName: Full=Probable glutathione S-transferase GSTF2; AltName: Full=GST-II	178	178	92%	2e-49	24%	O82451.3	☑	
☐	RecName: Full=Glutathione S-transferase 1; AltName: Full=GST class-phi	178	178	94%	3e-49	21%	P30110.1	☑	✓
☐	RecName: Full=Glutathione S-transferase zeta class	178	178	92%	5e-49	26%	P57108.1	☑	✓
☐	RecName: Full=Glutathione S-transferase F5; Short=AtGSTF5; AltName: Full=GST class-phi m	178	178	93%	9e-49	23%	Q9SRY6.2	☑	✓
☐	RecName: Full=Glutathione S-transferase PARB; AltName: Full=GST class-phi	174	174	98%	6e-48	19%	P30109.1	☑	
☐	RecName: Full=Glutathione S-transferase Z2; Short=AtGSTZ2; AltName: Full=GST class-zeta ı	172	172	92%	5e-47	26%	Q9ZVQ4.1	☑	

Improving Accuracy



Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements

Alejandro A. Schäffer*, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin and Stephen F. Altschul

+ Author Affiliations

Abstract

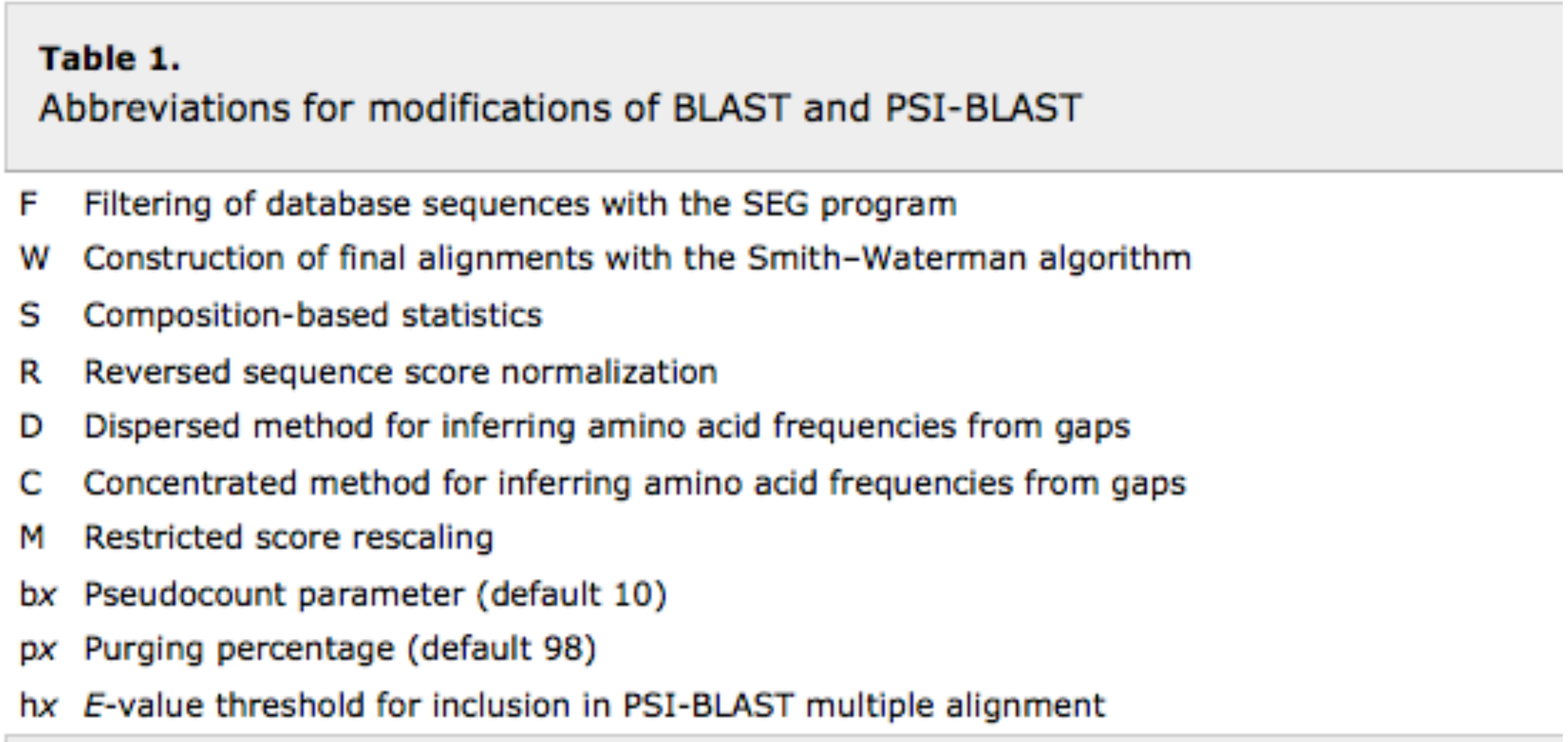
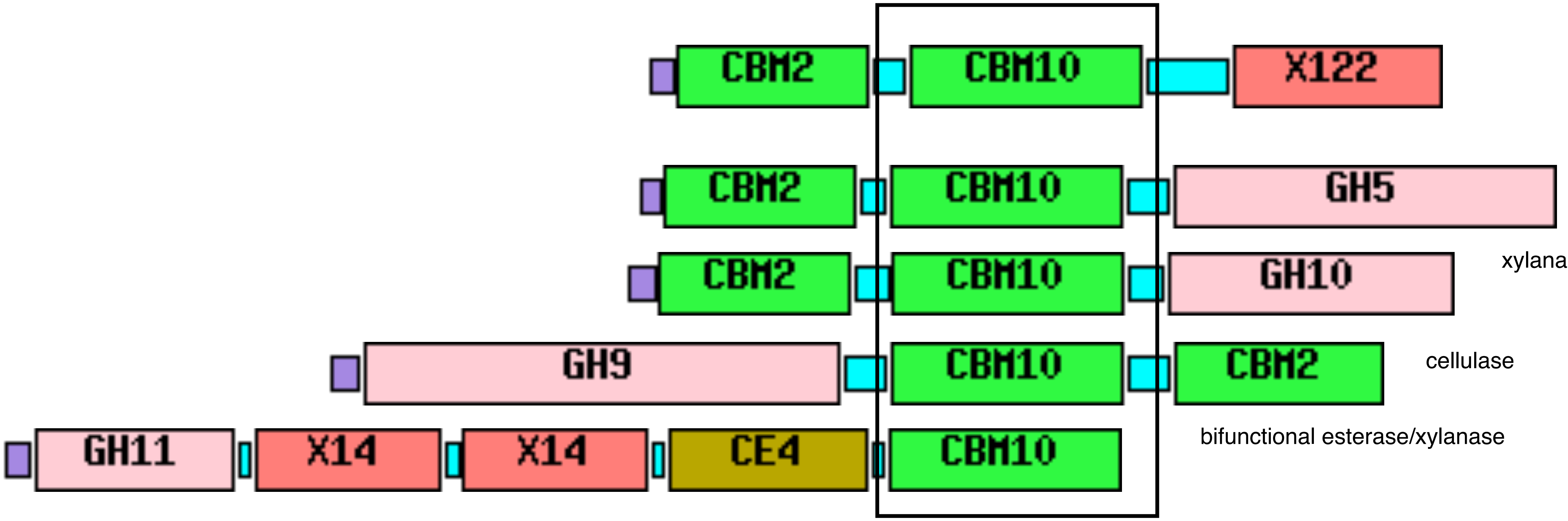


Table 1.
Abbreviations for modifications of BLAST and PSI-BLAST

F	Filtering of database sequences with the SEG program
W	Construction of final alignments with the Smith–Waterman algorithm
S	Composition-based statistics
R	Reversed sequence score normalization
D	Dispersed method for inferring amino acid frequencies from gaps
C	Concentrated method for inferring amino acid frequencies from gaps
M	Restricted score rescaling
b*x*	Pseudocount parameter (default 10)
p*x*	Purging percentage (default 98)
h*x*	*E*-value threshold for inclusion in PSI-BLAST multiple alignment

Error in Profile Searches



b-1,3-glucanase

xylanase

cellulase

bifunctional esterase/xylanase

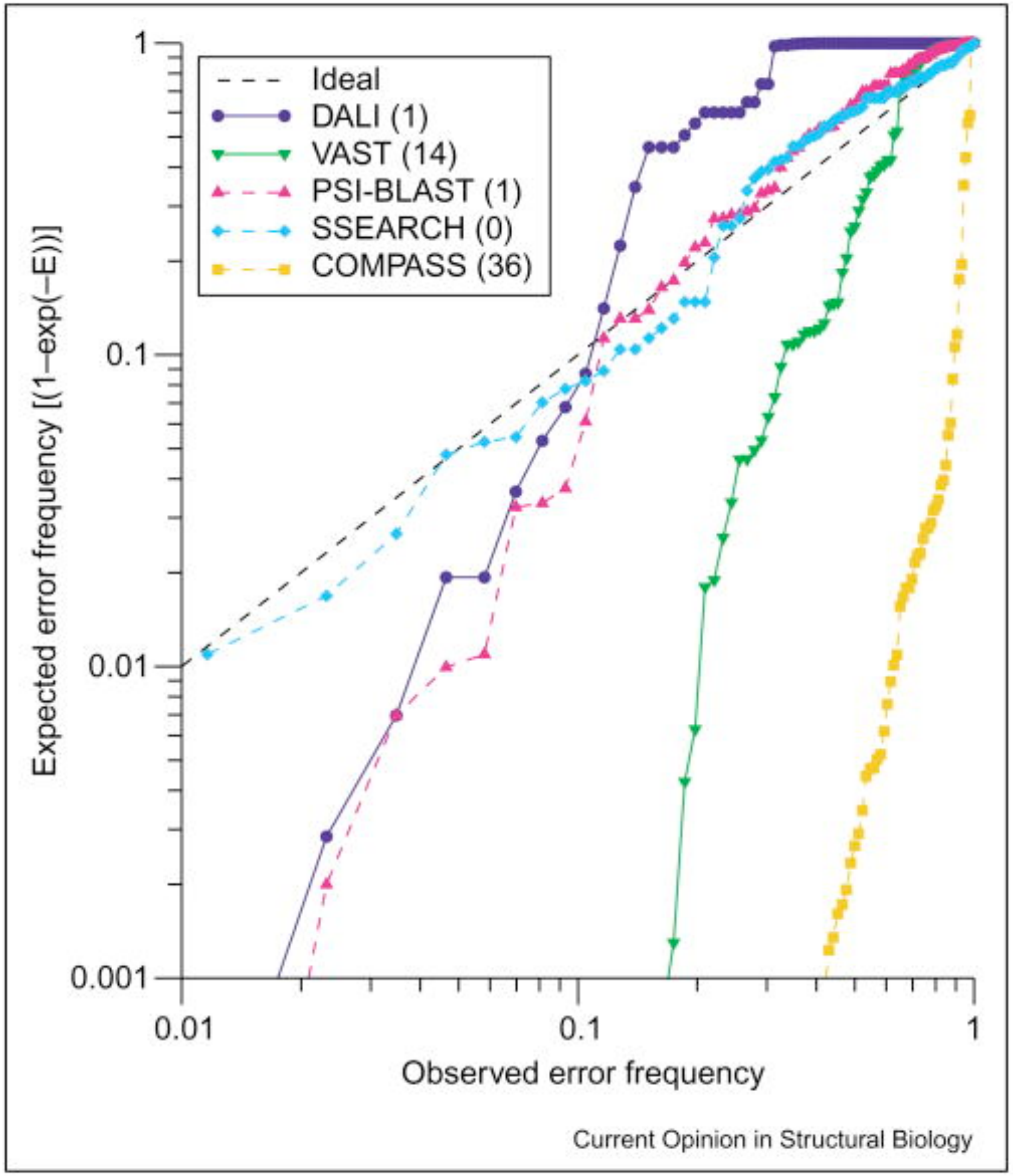
Homologous Over-Extension

Drawbacks to PSI-Search

- Hard to compare 2 profiles
- With few input sequences it's hard to create an accurate profile
- Including a non-homolog will capture "it's friends"

Error in Profile Searches

More Errors than Expected in PSI-BLAST vs SSEARCH



Curr Opin Struct Biol. 2005 Jun;15(3):254-60.

The limits of protein sequence comparison?

Pearson WR, Sierk ML.

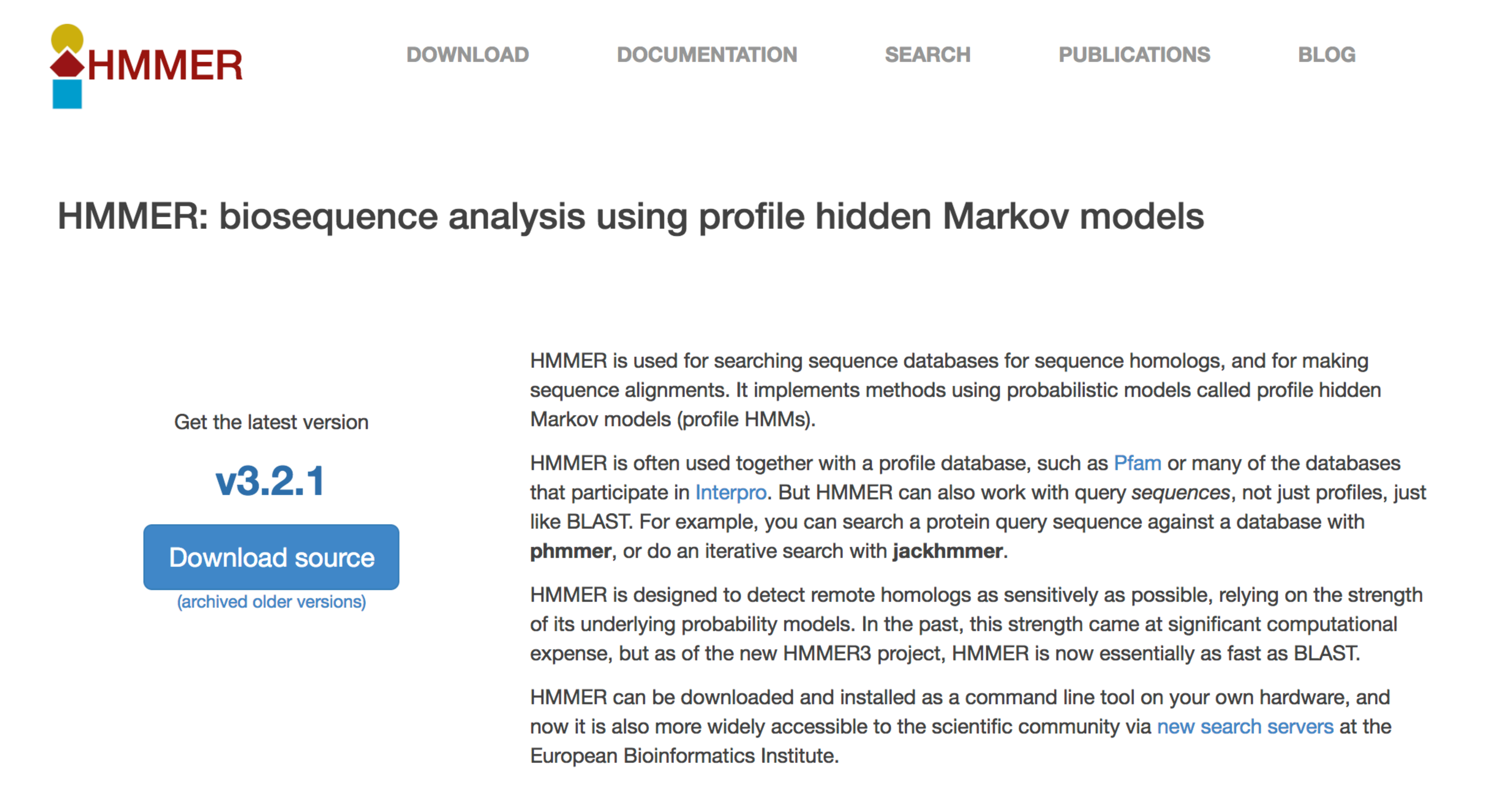
Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA. wrp@virginia.edu

HMMER

- phmmer
  - Compares a protein sequence against a protein sequence database
- hmmscan
  - Compares a protein sequence to a profile HMM
- hmmsearch
  - Compares a profile HMM again a protein sequence database
- jackhammer
  - interactive hmmsearch

HMMER

It detects homology by comparing a profile-HMM to either a single sequence or a database of sequences.



HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

v3.2.1

(archived older versions)

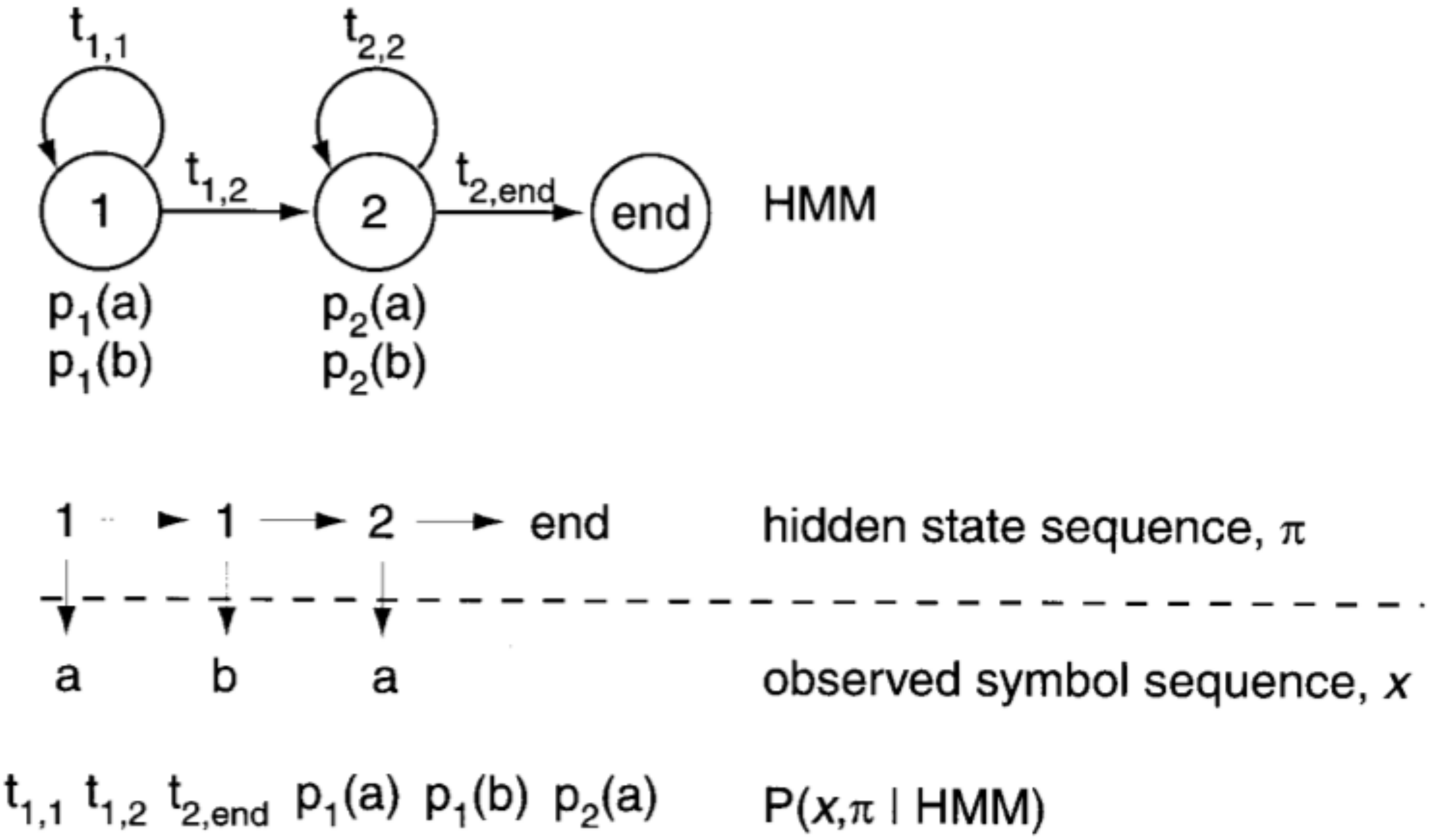
HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query *sequences*, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with **phmmer**, or do an iterative search with **jackhmmer**.

HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.

HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute.

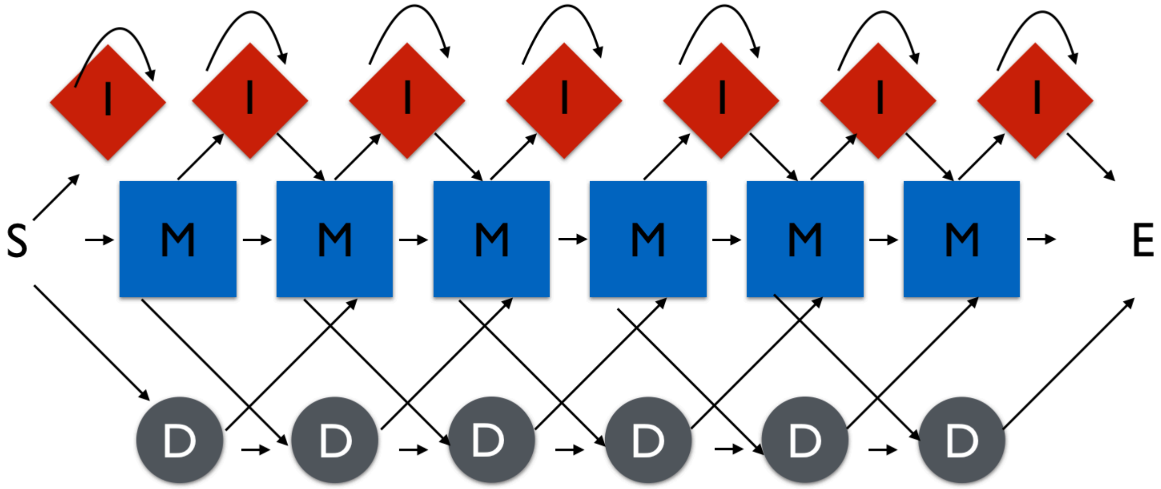
Model HMM



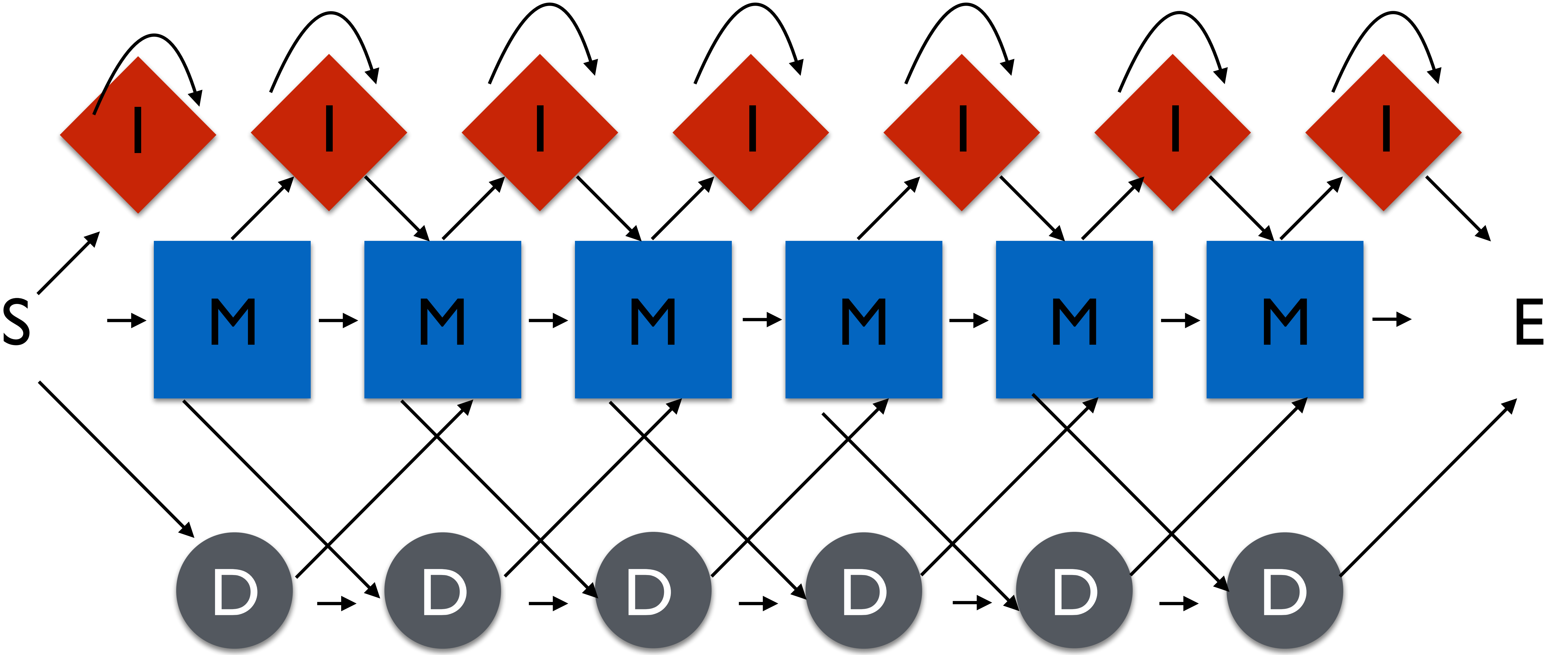
HMM, modeling sequences of as and bs as 2 regions of potentially different residue composition

Profile HMM

- HMM describes the probabilities of each state transitions
- M_i to I_i, I_i to itself, I_i to M_{i+1}
- M_i to M_{i+1}
- M_i to D_{i+1}, D_i to D_{i+1}, D_i to M_{i+1}



Profile HMM



```
AT-GTTAT
TACGT-AC
MMIMMDMM
```

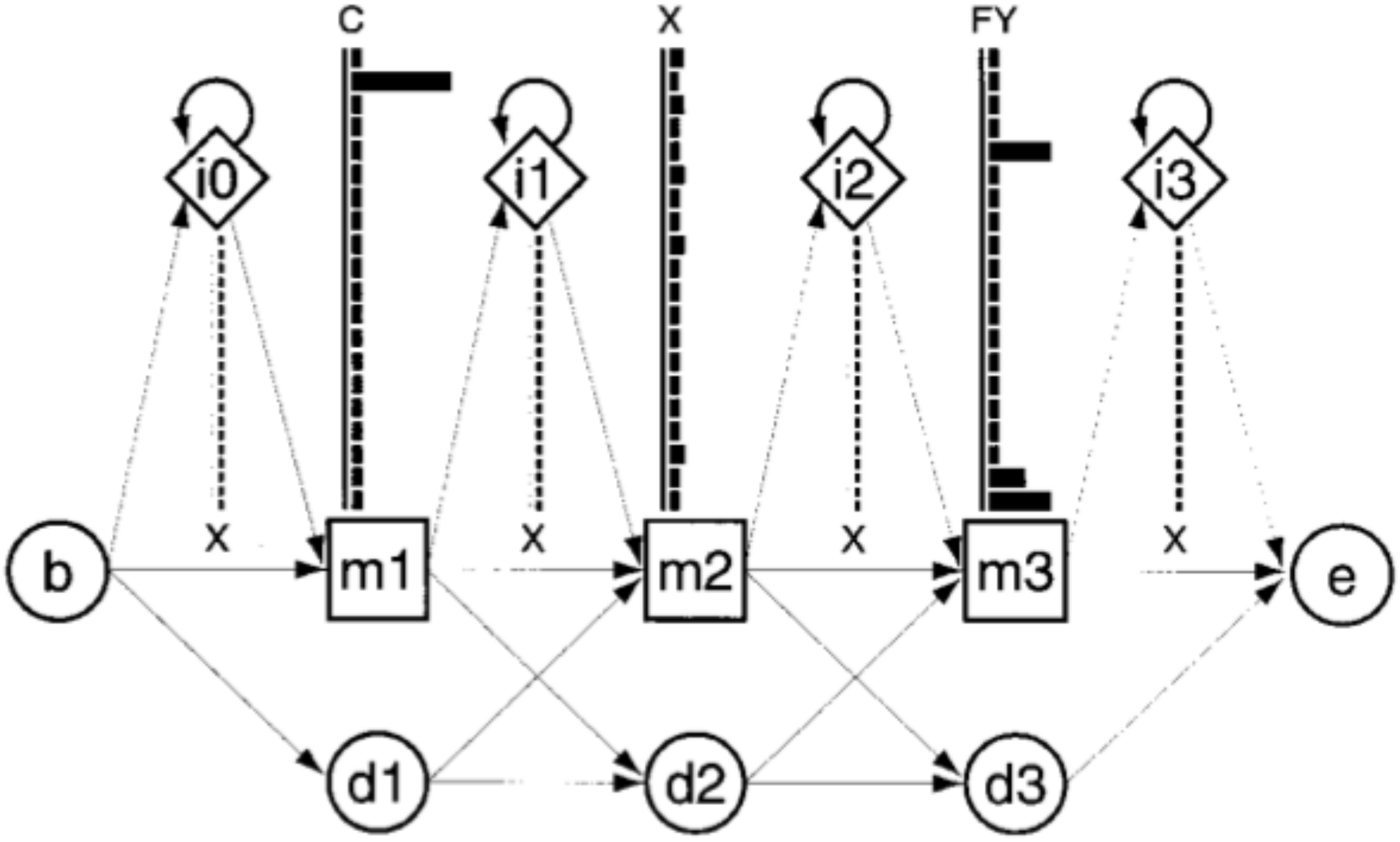
Derive HMMs from Multiple Sequence Alignment

Profile HMMs represents the consensus for the alignment of sequence from the same family and are built using a multiple sequence alignment

```
sp|O74706|EGLB_ASPNG      MKFQSTL--LLAAAAGSALAV-----------------PHGSGHKKRASVFEWFGSNESG
sp|Q96WQ8|EGLB_ASPKA      MKFQSTL--LLAAAAGSALAV-----------------PHGPGHKKRASVFEWFGSNESG
sp|P51529|MANA_STRLI      MR---NARSTLITTAGMAFAVLGLLFALAGPSAGRAEAAAGGIHVSNGRVVE--GNGSAF
sp|P22533|MANB_CALSA      MRLKTKIRKKWLSVLCTVVFLLNILFI-----ANVTILPKVGAATSNDGVVKI----DTS
                          *.   .      :.   .. :                 .      ..  *.:     .:

sp|O74706|EGLB_ASPNG      AEFGTNIPGVWGTDYIFPDPST--ISTLIGKGMNFFRVQFMMERLLPDSMTGSYDEEYLA
sp|Q96WQ8|EGLB_ASPKA      AEFGTNIPGVWGTDYIFPDPSA--ISTLIDKGMNFFRVQFMMERLLPDSMTGSYDEEYLA
sp|P51529|MANA_STRLI      VMRGVNHAYTW-----YPDRTGS-IADIAAKGANTVRVVL--------SSGGRWTKTSAS
sp|P22533|MANB_CALSA      TLIGTNHAHCW-----YRDRLDTALRGIRSWGMNSVRVVL--------SNGYRWTKIPAS
                          .  *.* .  *     : *     :  :  .* * .** :        *    : :   :
```

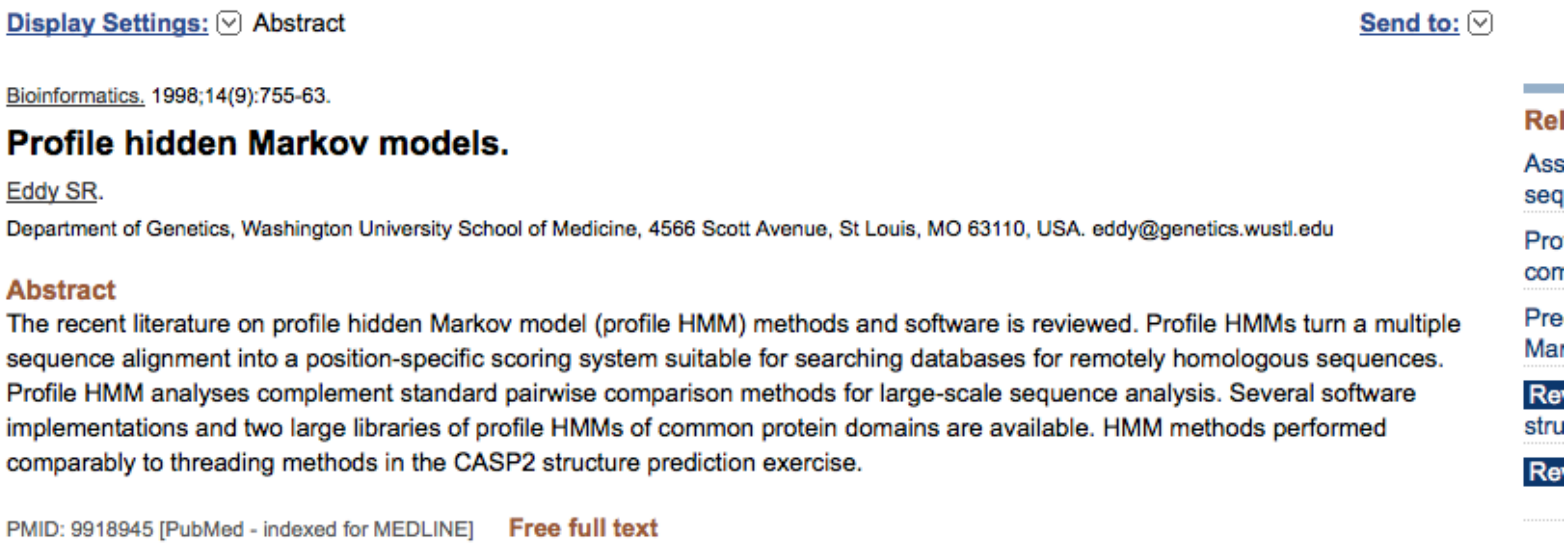
profile HMM

1	2	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	Y



represents a short multiple alignment of 5 sequences with 3 consensus colums

profile HMMs



Bioinformatics. 1998;14(9):755-63.

Profile hidden Markov models.

Eddy SR.

Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St Louis, MO 63110, USA. eddy@genetics.wustl.edu

Abstract

The recent literature on profile hidden Markov model (profile HMM) methods and software is reviewed. Profile HMMs turn a multiple sequence alignment into a position-specific scoring system suitable for searching databases for remotely homologous sequences. Profile HMM analyses complement standard pairwise comparison methods for large-scale sequence analysis. Several software implementations and two large libraries of profile HMMs of common protein domains are available. HMM methods performed comparably to threading methods in the CASP2 structure prediction exercise.

PMID: 9918945 [PubMed - indexed for MEDLINE] Free full text

- Takes the “standard” profiles and uses HMM based “standard” mathematics to solve two problems
- Profile-HMM scores are comparable (sort of)
- Sets gap costs

How to build a profile HMMs

1. Collect the protein sequences from the same protein family
2. Generate a multiple in one of the following formats:
    1. Stockholm, aligned FASTA, Clustal, PSI-BLAST, SELEX and PHYLIP.
3. Use hmmbuild to create a profile HMM
4. This profile can be used to identify distant family members

Multiple Sequence Alignment Tools

https://www.ebi.ac.uk/Tools/msa/

- Clustal Omega
- T-Coffee
- Muscle

Protein Domain Summary

- Protein Domains are independent structural entities that are found with various partners.
- Protein divergence is not uniform over a protein - some parts are more conserved than others
- Position specific scoring matrices can capture the specific patterns of conservation at different sites in a protein
- PSI-BLAST combines searching, multiple alignment, and PSSMs
- Statistical estimates are difficult with PSSMs, use PSI- SEARCH and PSI-PRSS
- HMMER3 creates HMM models of a protein family from a multiple sequence alignment
- Iterative PSSM/HMM searches may be contaminated by Homologous Overextension
- Single models cannot capture diverse families (PFAM Clans)
- Protein domains can be identified using RPS-BLAST or CDD searching

Workshop
Time

https://bcantarel.github.io/cshl_homology_workshop2