PROTEOMICS • FASTA FILES • SEARCH OPTIMIZATION

FilterFasta: a Tool for Generating Keyword Based Bespoke Cross Organism FASTA Files for Proteomics

Jack Chlystek[1], Jonathan Le[2], Jonathan E. Katz[1,2]

[1] Lawrence J. Ellison Institute for Transformative Medicine, Los Angeles, CA
[2] Department of Medicine/Oncology, Keck School of Medicine,
University of Southern California, Los Angeles, California

Correspondence: jonathan@eitm.org
Reprints: https://reprints.katzlab.org/

Background

Mass spectrometry (MS)-based proteomics is often used to perform identification and/or differential quantification on a collection of proteins that are derived from one or more known species of origin. The typical bottom-up proteomics workflow (Figure A) critically depends on the underlying database that is used to drive the search algorithms - making most search algorithms "blind" to proteins or modifications that are not included. One considerably less common workflow is the requirement to identify the species of origin for a protein that has a known function. A major challenge of this is that the straight forward approach of creating a "mega" database which includes all potential species of origin is both cumbersome and often leads to a database size that exceeds the TCI (Threshold of Computational Infeasibility, a term we provide for the enjoyment of the reader) whereby both the False Discovery Rate (FDR) and the computational "clock time" exceed acceptable tolerance (Alves et. al, 2018).

However, in some instances, researchers may already know the identity of one or more proteins expected to be in the sample and would like to determine the organism from which it originated. If only there were a way to create a bespoke search database focused around these putative proteins…

Methods

DATA ACQUISITION: A single stranded DNA binding protein was received from [redacted] someone with the request to identify the species of origin. The protein was denatured (guanidine), reduced (dithiothreitol) and alkylated (iodoacetamide) before digestion with trypsin by standard protocols. The peptides were analyzed by C18 liquid chromatography-tandem mass spectrometry (LC-MS/MS) using an LTQ Orbitrap XL mass spectrometer using standard acquisition parameters. Searches were performed by FragPipe(https://fragpipe.nesvilab.org/, 20 ppm tolerance of parent ion, 2 missed cleavages) using databases specified in figures after the addition of decoys and common contaminants.

FilterFasta DEVELOPMENT: Protein sequences from all reference proteomes were obtained from UniProt using the command:

```
wget https://ftp.uniprot.org/pub/databases/uniprot/current
_release/knowledgebase/reference_proteomes/Reference_Proteomes_2021_04.t
ar.gz
```

The resultant ~60 million protein entries were uncompressed and an index of all the descriptions was created. A C program was written that scanned the descriptions for keyword matches (terms logically ANDed) and returned the complete sequence entry. A PHP wrapper was created to create a web interface through which the program could be accessed.

Results

Our initial attempt at species identification based on 2 putative source organisms of suspected origin led to no matches of significance – only common contaminants were identified as matches in the search (Figure B.3). We considered a search of all proteins within a single domain (e.g., Archaea or Eubacteria) past the TCI. For this particular study, we assign a TCI to be around ~50,000 proteins based on the expected time to compute and FDR. This threshold is roughly double the size of all proteins for *Homo sapiens*. The FilterFasta tool allowed the creation of a database containing only proteins related to SSB across all organisms in the UniProt database. The bespoke FASTA file from a "Single Stranded Binding" keyword search led to a species identification of *Saccharolobus solfataricus* of the Archaea domain as the species from which our protein derived.

Conclusions

This workflow, for us, provided a simple mechanism by which a protein based FASTA database was created and was used to identify the species of origin for a protein of unknown provenance but known activity. We imagine that other workflows could take advantage of this mechanism when tradition identification techniques (such as DNA) are unavailable. While there are many ways to create bespoke FASTA files, including, potentially, with the same criteria that we used here, FilterFasta was able to provide result with considerably increased ease. For example, in this authors hands, the best approximate UniProt search we performed yielded over 100,000 entries with a more convoluted search of "protein_name:"single" protein_name:"stranded" protein_name:"binding"" without a clear path for further refinement.

Highlights

- The success of bottom-up proteomics hinges upon the appropriate selection of search databases, a process typically guided by organism specificity.
- With samples whose species of origin delve into the domain of the unknown, the challenge of selecting an appropriate search database arises.
- Acting as the ultimate enabler, FilterFasta facilitates the creation of bespoke search databases which rely solely on the potential of a putative protein lurking within, undeterred by the enigma surrounding the species of origin - Cases where you know a protein will be present, yet the species remains a mystery!
- FilterFasta emerges as a tool for mass spectrometry-based proteomics, offering not just computational efficiency, but also a path to unlocking the mysteries of unknown protein origins.

References

Alves, G., Wang, G., Ogurtsov, A. Y., Drake, S. K., Gucek, M., Sacks, D. B., & Yu, Y.-K. (2018). Rapid classification and identification of multiple microorganisms with accurate statistical significance via high-resolution tandem mass spectrometry. Journal of the American Society for Mass Spectrometry, 29(8), 1721-1737. https://doi.org/10.1007/s13361-018-1986-y

Ma, Bin; Zhang, Kaizhong; Hendrie, Chris; et al. (2003). PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Molecular & Cellular Proteomics, 3(6), 1154-1165. https://doi.org/10.1074/mcp.M300001-MCP200

Conflicts of Interest

There are no conflicts of interest to report.

A. Bottom-up proteomics workflow depends on the database created to match peptide sequences

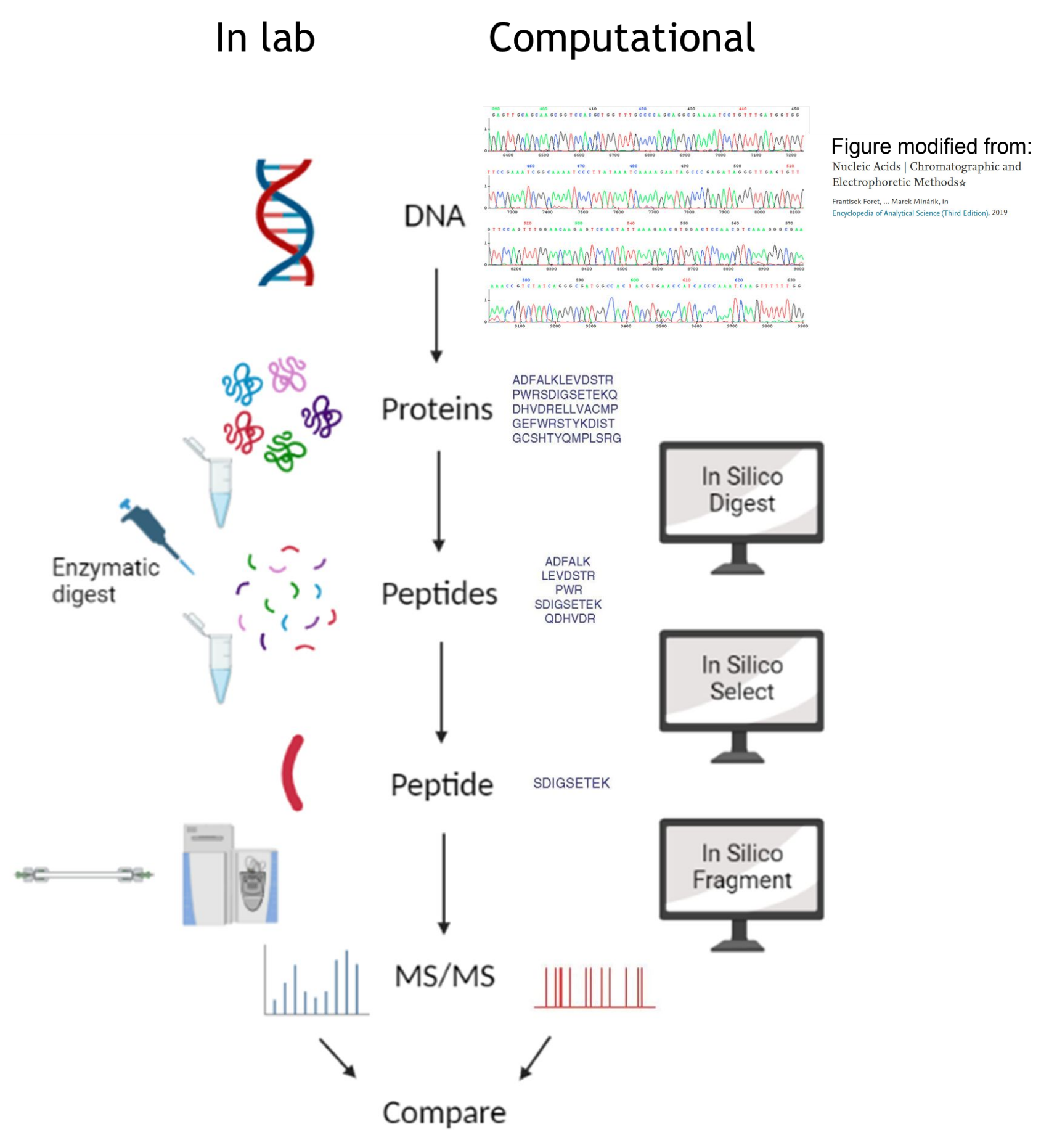
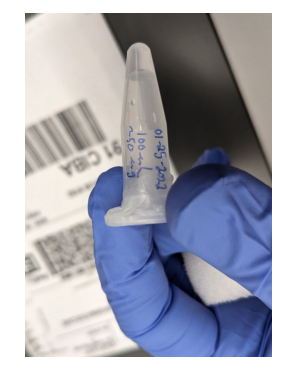


Figure modified from: Nucleic Acids | Chromatographic and Electrophoretic Methods☆ Frantisek Foret, ... Marek Minárik, in Encyclopedia of Analytical Science (Third Edition), 2019

Figure A: Effectiveness of bottom-up workflow is *In Silico* database dependent. A typical bottom-up MS-based proteomics workflow to identify an organism in an unknown sample begins with a tryptic digestion of the sample of interest, followed by analysis via liquid chromatography-tandem mass spectrometry (LC-MS/MS). The peptides identified in the sample are then searched against a reference sequence database that typically contains proteins from sources likely to be in the sample. The effectiveness of this workflow is dependent on the database used to search against.

B. Manual database curation failed to identify species of origin

1. Someone [redacted] asked…. "We have a commercial **SSB, single stranded binding protein**, can you tell us which **organism** is the source of this protein?. We think it is from **Thermotoga maritima** or **Thermatoga neapolitina**



2. >AAD35689.1 single stranded DNA-binding protein, putative [**Thermotoga maritima** MSB8]
MSFFNKIILIGRLVRDPEERYTLSGTPVTTFTIAVDRVPRKNAPDDAQTTDFFRIVTFGRLAEFARTYLT
KGRLVLVEGEMRMRRWETPTGEKRVSPEVVANVVRFMDRKPAETVSETEEELEIPEEDFSSDTFSEDEPPF

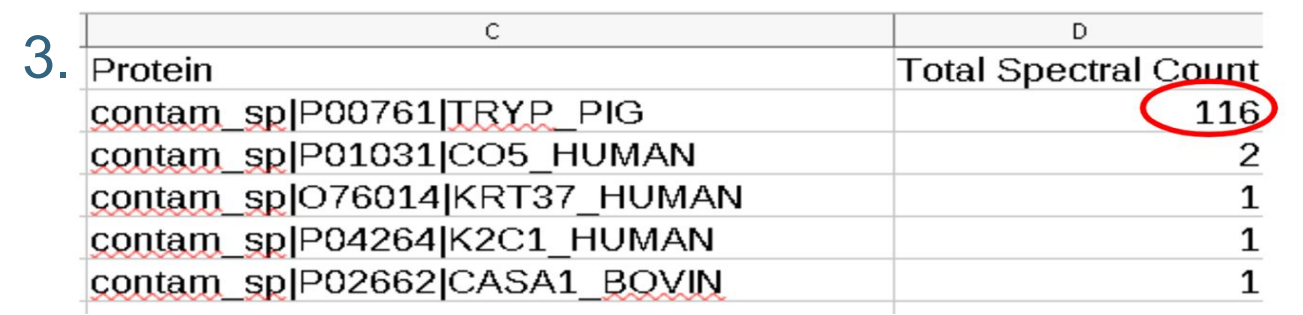
>GU125728.1 **Thermotoga neapolitana** strain DSMZ 4359 single stranded DNA-binding protein (ssb) gene, complete cds
MSFFNRIILIGRLVRDPEERYTLSGTPVTTFTIAVDRVPRKNAP
DDAQTTDFFRVVTFGRLAEFARTYLTKGRLILVEGEMRMRRWETQTGEKRVSPEVVAN
VVRFMDRKPVEMPSEDIEEKLEIPEEDFTDDTFSEDEPPF

>CAA0172152.1 Single-stranded+DNA-binding+protein [**Escherichia coli**]
MASRGVNKVILVGNLGQDPEVRYMPNGGAVANITLATSESWRDKATGEMKEQTEWHRVVLFGKLAEVASE
YLRKGSQVYIEGQLRTRKWTDQSGQDRYTTEVVVNVGGTMQMLGGRQGGGAPAGGNIGGGQPQGGWGQPQ
QPQGGNQFSGGAQSRPQQSAPAAPSNEPPMDFDDDIPF

>AAA36332.1 single stranded DNA binding protein [**Homo sapiens**]
MFRRPVLQVLRQFVRHESETTTSLVLERSLNRVHLLGRVGQDPVLRQVEGKNPVTIFSLATNEMWRSGDS
EVYQLGDVSQKTTWHRISVFRPGLRDVAYQYVKKGSRIYLEGKIDYGEYMDKNNVRRQATTIIADNIIFL
SDQTKEKE

>AAM24086.1 Single-stranded DNA-binding protein [C**aldanaerobacter subterraneus** subsp. tengcongensis MB4]
MAGNFLENNTVTLVGKVFTPLEFSHELYGEKFFNFILEVPRLSETKDYLPITISNRLFEGMNLEVGTRVK
IEGQLRSYNRKSPEEGKNKLILTVFARDISVVPEEEVVKNPNEIFLDGFICKKPVYRTTPLGREITDLLI
AVNRPYNKSDYIPVIAWGRNARFSEKLEVGDRIRLWGRVQSREYQKKLGDEVVTKVAYEVSITRMEVVEK
ELQKTE

3.



C	D
Protein	Total Spectral Count
contam_sp|P00761|TRYP_PIG	116
contam_sp|P01031|CO5_HUMAN	2
contam_sp|O76014|KRT37_HUMAN	1
contam_sp|P04264|K2C1_HUMAN	1
contam_sp|P02662|CASA1_BOVIN	1

Figure B: Custom database from presumed organisms did not yield significant results. 1. A collaborator gave us a sample of their single stranded binding protein (SSB) and asked us to identify the species of origin. **2.** They hypothesized the species was either *Thermatoga maritima* or *Thermatoga neapolitina*, so we created a custom FASTA file containing SSBs from those two species and other commonly found ones. **3.** A bottom-up workflow was used as shown in Figure A, but only common contaminants came back positive for the unknown protein.

C. FilterFasta tool creates bespoke database for searching single type of protein across all organisms

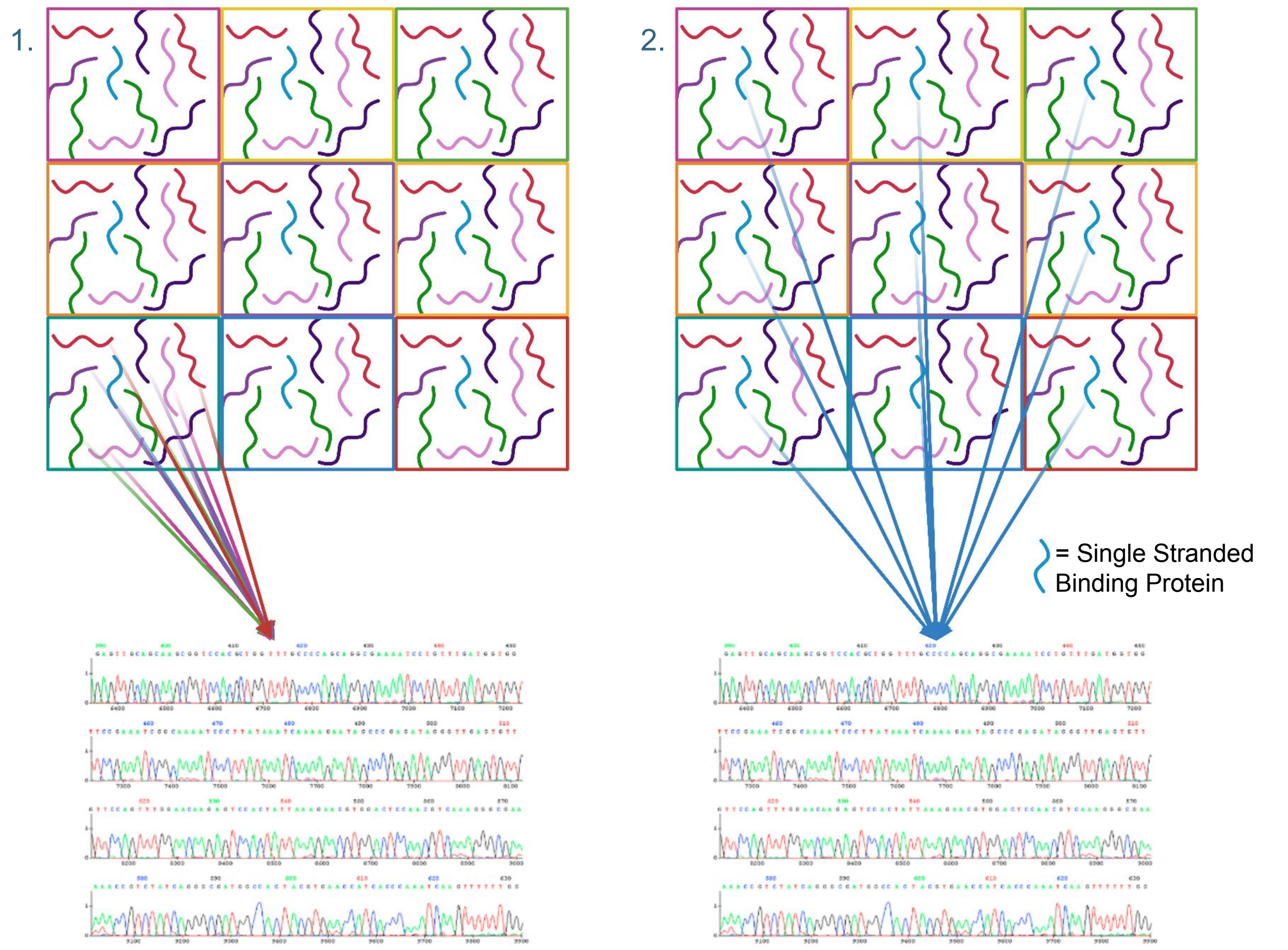
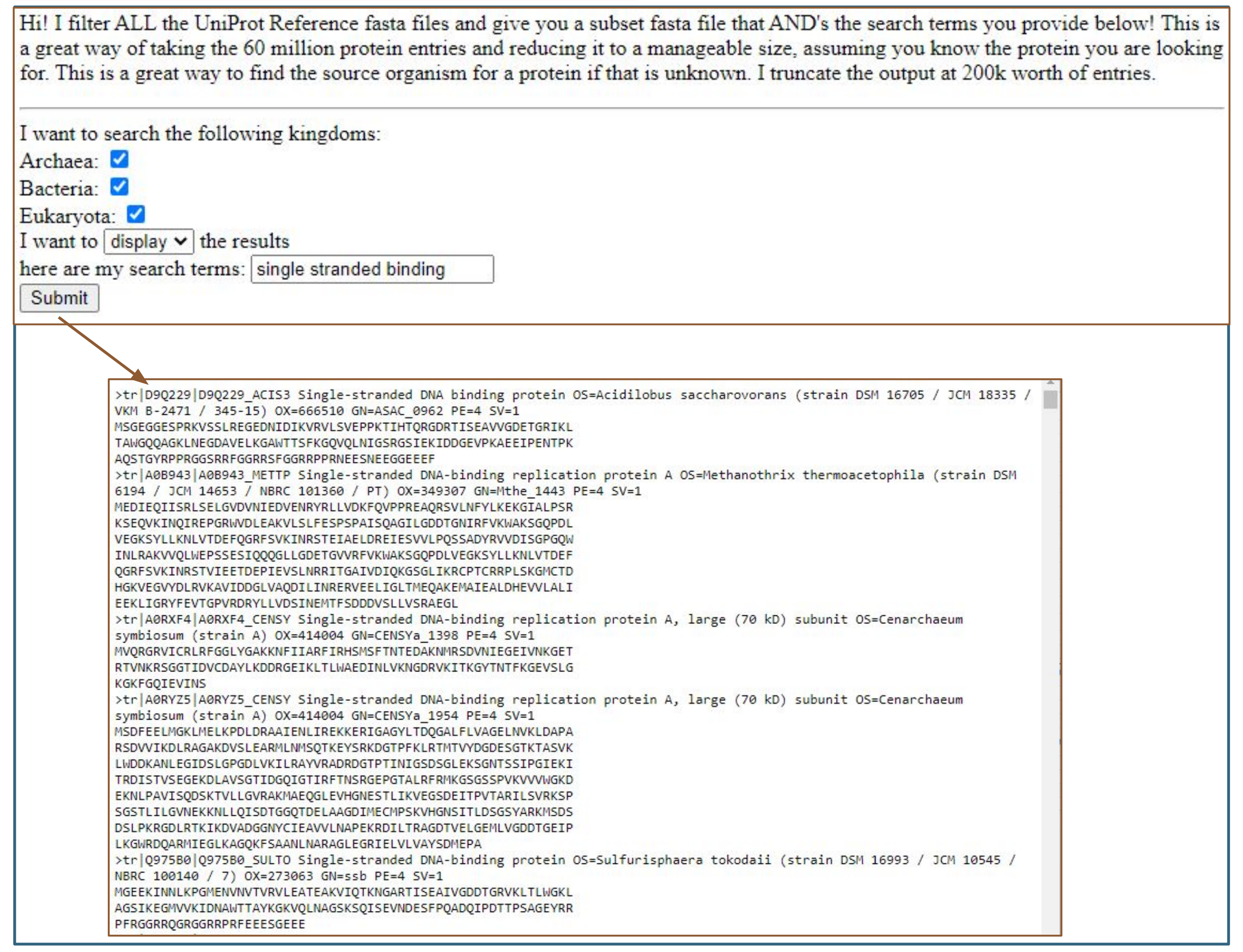


Figure C: Database by Organism vs by Protein Description. The FilterFasta tool operates with a superset database of all entries from UniProt as downloaded in 2021_04 (59,653,900 entries). For the two panels, the boxes are meant to reflect the myriad of organisms and within each box the squiggles reflect the myriad of proteins. **1.** A typical search database would be created by downloading the sequences of all the proteins from a single organism. This is a typical workflow for identifications of proteins from a known organism. **2.** FilterFasta takes keywords and searches all the protein descriptions from all the organisms and produces a single FASTA reflecting all the hits. In this case we used "Single Stranded Binding Protein" represented as the blue squiggle.

D. FilterFasta tool enables generation of FASTA based on keyword search

1. http://filterfasta.katzlab.org/

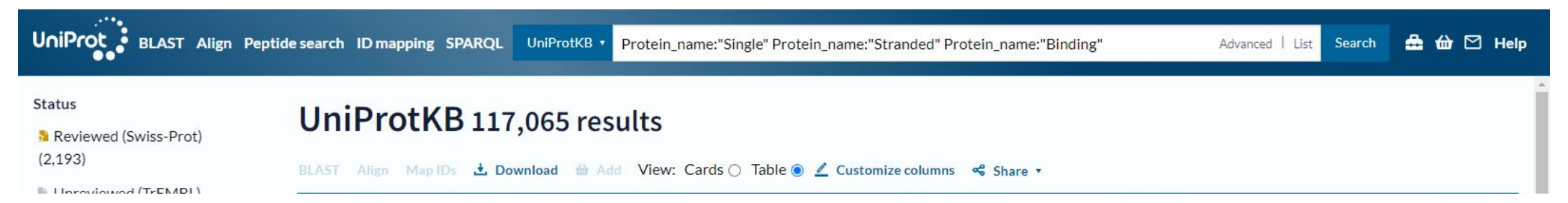


Hi! I filter ALL the UniProt Reference fasta files and give you a subset fasta file that AND's the search terms you provide below! This is a great way of taking the 60 million protein entries and reducing it to a manageable size, assuming you know the protein you are looking for. This is a great way to find the source organism for a protein if that is unknown. I truncate the output at 200k worth of entries.

I want to search the following kingdoms:
Archaea: ☑
Bacteria: ☑
Eukaryota: ☑
I want to display the results
here are my search terms: single stranded binding

```
>tr|D9Q229|D9Q229_ACIS3 Single-stranded DNA binding protein OS=Acidilobus saccharovorans (strain DSM 16705 / JCM 18335 /
VKM B-2471 / 345-15) OX=666510 GN=ASAC_0962 PE=4 SV=1
MSGEGGESPRKVSSLREGEDNIDIKVRVLSVEPPKTIHTQRGDRTISEAVVGDETGRIKL
TAWGQQAGKLNEGDAVELKGAWTTSFKGQVQLNIGSRGSIEKIDDGEVPKAEEIPENTPK
AQSTGYRPPRGGSRRFGGRRSFGGRRPPRNEESNEEGGEEEF
>tr|A0B943|A0B943_METTP Single-stranded DNA-binding replication protein A OS=Methanothrix thermoacetophila (strain DSM
6194 / JCM 14653 / NBRC 101360 / PT) OX=349307 GN=Mthe_1443 PE=4 SV=1
MEDIEQIISRLSELGVDVNIEDVENRYRLLVDKFQVPPREAQRSVLNFYLKEKGIALPSR
KSEQVKINQIREPGRWVDLEAKVLSLFESPSPAISQAGILGDDTGNIRFVKWAKSGQPDL
VEGKSYLLKNLVTDEFQGRFSVKINRSTEIAELDREIESVVLPQSSADYRVVDISGPGQW
INLRAKVVQLWEPSSESIQQQGLLGDETGVVRFVKWAKSGQPDLVEGKSYLLKNLVTDEF
QGRFSVKINRSTVIEETDEPIEVSLNRRITGAIVDIQKGSGLIKRCPTCRRPLSKGMCTD
HGKVEGVYDLRVKAVIDDGLVAQDILINRERVEELIGLTMEQAKEMAIEALDHEVVLALI
EEKLIGRYFEVTGPVRDRYLLVDSINEMTFSDDDVSLLVSRAEGL
>tr|A0RXF4|A0RXF4_CENSY Single-stranded DNA-binding replication protein A, large (70 kD) subunit OS=Cenarchaeum
symbiosum (strain A) OX=414004 GN=CENSYa_1398 PE=4 SV=1
MVQRGRVICRLRFGGLYGAKKNFIIARFIRHSMSFTNTEDAKNMRSDVNIEGEIVNKGET
RTVNKRSGGTIDVCDAYLKDDRGEIKLTLWAEDINLVKNGDRVKITKGYTNTFKGEVSLG
KGKFGQIEVINS
>tr|A0RYZ5|A0RYZ5_CENSY Single-stranded DNA-binding replication protein A, large (70 kD) subunit OS=Cenarchaeum
symbiosum (strain A) OX=414004 GN=CENSYa_1954 PE=4 SV=1
MSDFEELMGKLMELKPLDLDRAAIENLIREKKERIGAGYLTDQGALFLVAGELNVKLDAPA
RSDVVIKDLRAGAKDVSLEARMLNMSQTKEYSRKDGTPFKLRTMTVYDGDESGTKTASVK
LWDDKANLEGIDSLGPGDLVKILRAYVRADRDGTPTINIGSDSGLEKSGNTSSIPGIEKI
TRDISTVSEGEKDLAVSGTIDGQIGTIRFTNSRGEPGTALRFRMKGSGSSPVKVVVWGKD
EKNLPAVISQDSKTVLLGVRAKMAEQGLEVHGNESTLIKVEGSDEITPVTARILSVRKSP
SGSTLILGVNEKKNLLQISDTGGQTDELAAGDIMECMPSKVHGNSITLDSGSYARKMSDS
DSLPKRGDLRTKIKDVADGGNYCIEAVVLNAPEKRDILTRAGDTVELGEMLVGDDTGEIP
LKGWRDQARMIEGLKAGQKFSAANLNARAGLEGRIELVLVAYSDMEPA
>tr|Q975B0|Q975B0_SULTO Single-stranded DNA-binding protein OS=Sulfurisphaera tokodaii (strain DSM 16993 / JCM 10545 /
NBRC 100140 / 7) OX=273063 GN=ssb PE=4 SV=1
MGEEKINNLKPGMENVNVTVRVLEATEAKVIQTKNGARTISEAIVGDDTGRVKLTLWGKL
AGSIKEGMVVKIDNAWTTAYKGKVQLNAGSKSQISEVNDESFPQADQIPDTTPSAGEYRR
PFRGGRRQGRGGRRPRFEEESGEEE
```

2.



Protein_name:"Single" Protein_name:"Stranded" Protein_name:"Binding"

UniProtKB 117,065 results

Figure D: FilterFasta website interface. 1. The user inputs a search term and selects the kingdom(s) of interest to search. Results are either displayed or downloaded as a FASTA file. Output is limited to the first 200,000 entries in the case of very broad search terms. **2.** For comparison we show the additional complexity of performing this as a UniProt search ("single" "stranded" "binding") resulting in a FASTA file that contains 117,065 peptide sequences versus our result of 18,283 sequences – it is unclear in the interface why there is a disparity and how to refine this result to match that of FilterFasta

E. All-inclusive search of proteins is not computationally feasible

1.

Domain	Number of Proteins
Archaea	760,947
Bacteria	30,426,496
Eukaryota	28,466,457
Total Proteins	59,653,900

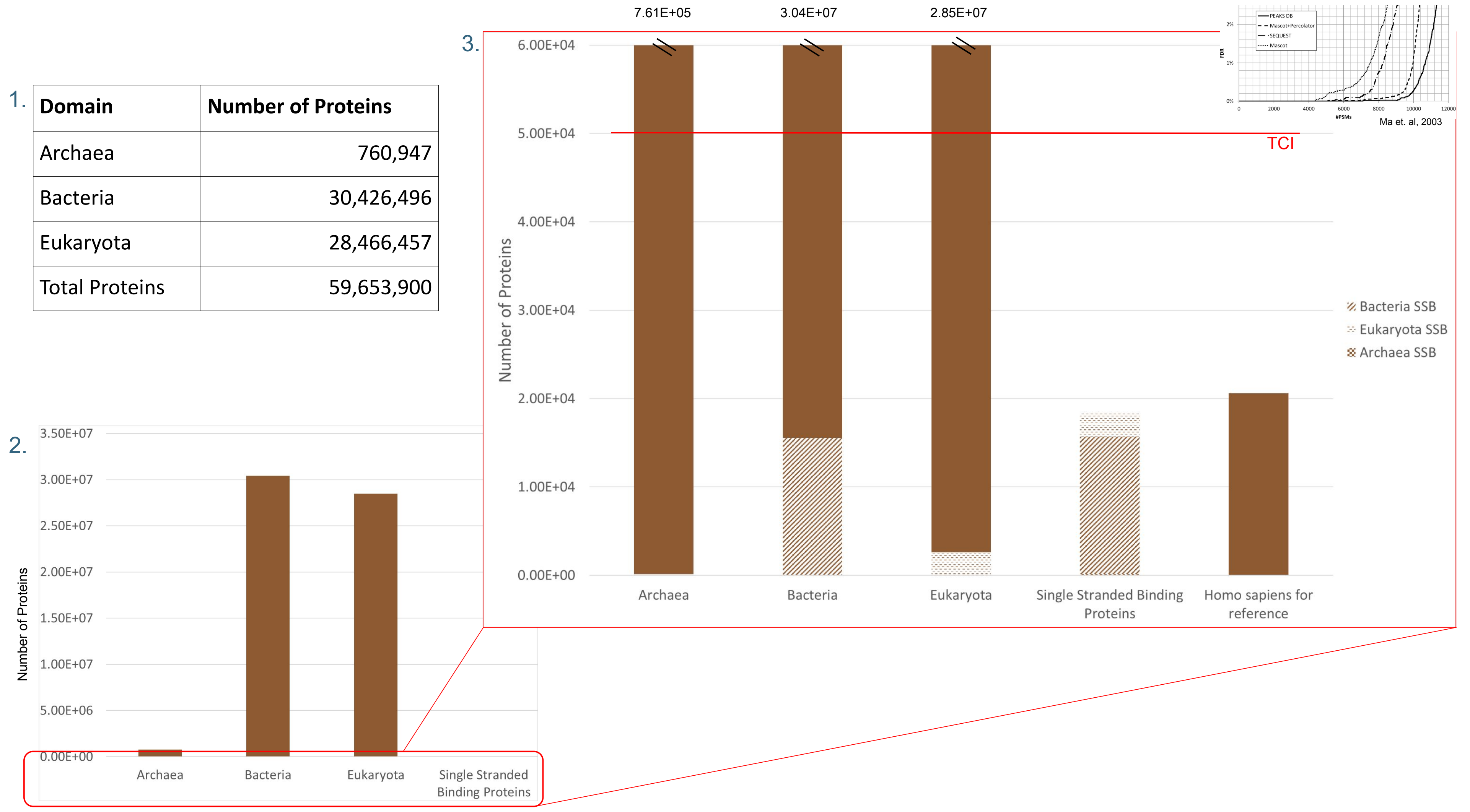
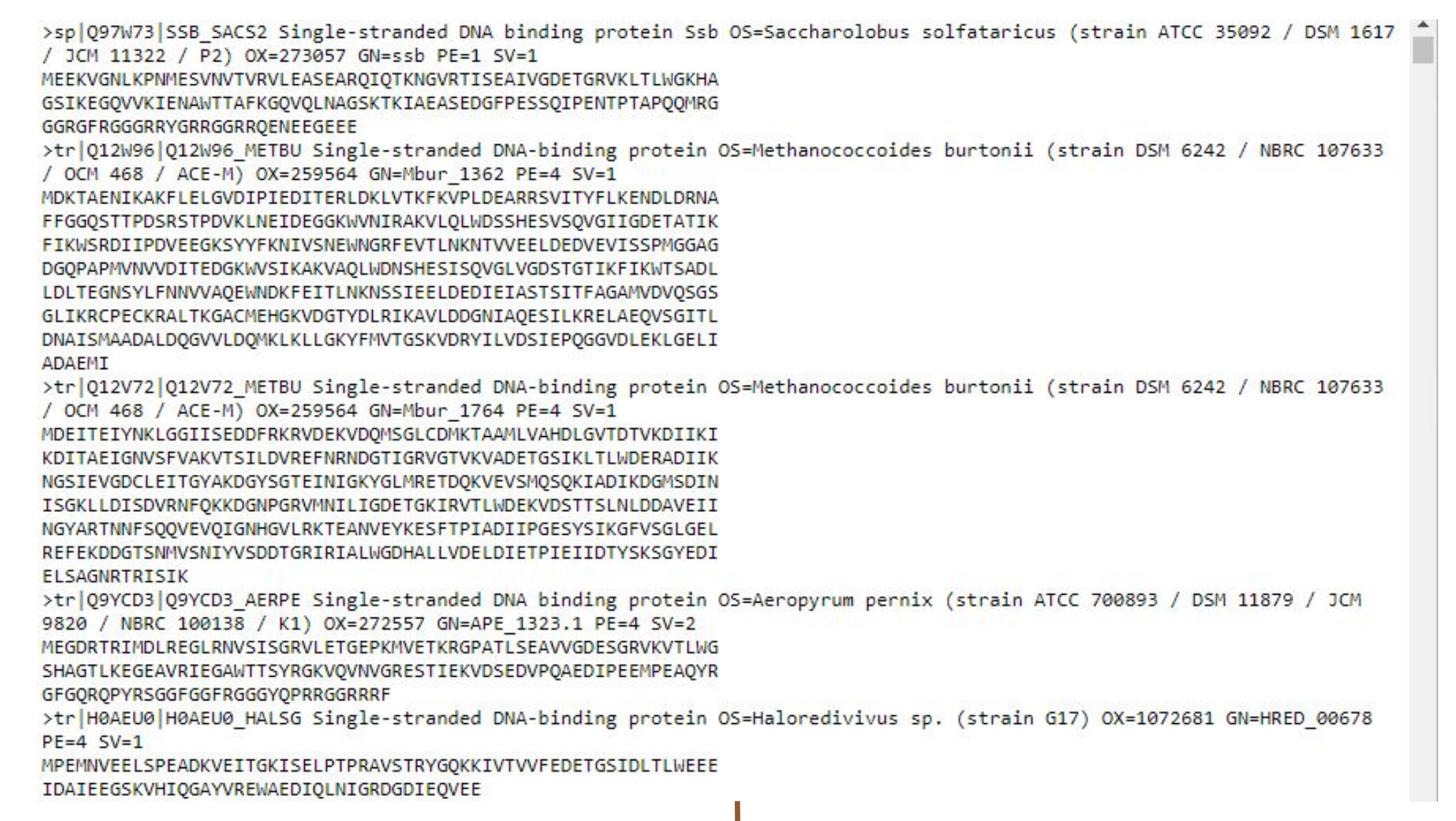
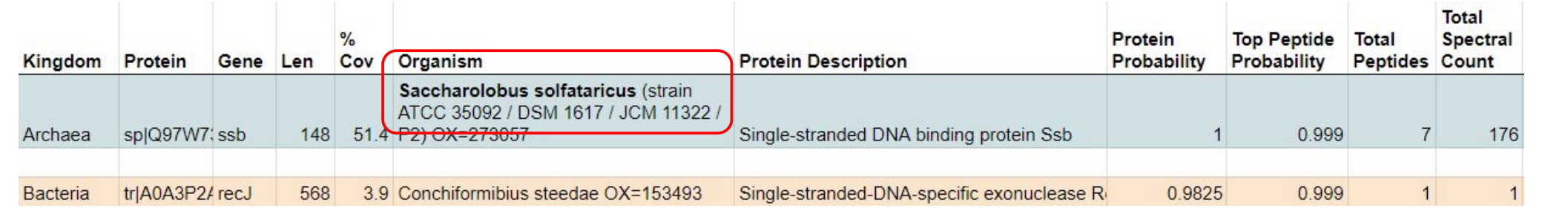


Figure E: Database Size Comparisons. 1. Table showing number of proteins in each domain from the 2021_04 download. **2.** Full graph of number of proteins in each domain, as well as all SSB proteins from our FilterFasta keyword search. **3.** Blow-up of the y-axis of plot 2 to show the SSB overlap within the three domains, with *Homo sapiens* for reference. The red line shows the TCI, which estimates the limit of what is computationally feasible to search based on time to compute and FDR (For our case we set at 50,000 proteins). The plot in the top right corner of (3) shows the FDR vs. peptide spectrum matches for various search programs (Ma et. al, 2003).

F. Bespoke FASTA file led to efficient and facile identification of species of origin



```
>sp|Q97W73|SSB_SACS2 Single-stranded DNA binding protein Ssb OS=Saccharolobus solfataricus (strain ATCC 35092 / DSM 1617
/ JCM 11322 / P2) OX=273057 GN=ssb PE=1 SV=1
MEEKVGNLKPNMESVNVTVRVLEASEARQIQTKNGVRTISEAIVGDETGRVKLTLWGKHA
GSIKEGQVVKIENAWTTAFKGQVQLNAGSKTKIAEASEDGFPESSQIPENTPTAPQQMRG
GGRGFRGGGRRYGRRGGRRQENEEGEEE
>tr|Q12W96|Q12W96_METBU Single-stranded DNA-binding protein OS=Methanococcoides burtonii (strain DSM 6242 / NBRC 107633
/ OCM 468 / ACE-M) OX=259564 GN=Mbur_1362 PE=4 SV=1
MDKTAENIKAKFLELGVDIPIEDITERLDKLVTKFKVPLDEARRSVITYFLKENDLDRNA
FFGGQSTTPDSRSTPDVKLNEIDEGGKWVNIRAKVLQLWDSSHESVSQVGIIGDETATIK
FIKWSRDIIPDVEEGKSYYFKNIVSNEWNGRFEVTLNKNTVVEELDEDVEVISSPMGGAG
DGQPAPMVNVVDITEDGKWVSIKAKVAQLWDNSHESISQVGLVGDSTGTIKFIKWTSADL
LDLTEGNSYLFNNVVAQEWNDKFEITLNKNSSIEELDEDIEIASTSITFAGAMVDVQSGS
GLIKRCPECKRALTKGACMEHGKVDGTYDLRIKAVLDDGNIAQESILKRELAEQVSGITL
DNAISMAADALDQGVVLDQMKLKLLGKYFMVTGSKVDRYILVDSIEPQGGVDLEKLGELI
ADAEMI
>tr|Q12V72|Q12V72_METBU Single-stranded DNA-binding protein OS=Methanococcoides burtonii (strain DSM 6242 / NBRC 107633
/ OCM 468 / ACE-M) OX=259564 GN=Mbur_1764 PE=4 SV=1
MDEITEIYNKLGGIISEDDFRKRVDEKVDQMSGLCDMKTAAMLVAHDLGVTDTVKDIIKI
KDITAEIGNVSFVAKVTSILDVREFNRNDGTIGRVGTVKVADETGSIKLTLWDERADIIK
NGSIEVGDCLEITGYAKDGYSGTEINIGKYGLMRETDQKVEVSMQSQKIADIKDGMSDIN
ISGKLLDISDVRNFQKKDGNPGRVMNILIGDETGKIRVTLWDEKVDSTTSLNLDDAVEII
NGYARTNNFSQQVEVQIGNHGVLRKTEANVEYKESFTPIADIIPGESYSIKGFVSGLGEL
REFEKDDGTSNMVSNIYVSDDTGRIRIALWGDHALLVDELDIETPIEIIDTYSKSGYEDI
ELSAGNRTRISIK
>tr|Q9YCD3|Q9YCD3_AERPE Single-stranded DNA binding protein OS=Aeropyrum pernix (strain ATCC 700893 / DSM 11879 / JCM
9820 / NBRC 100138 / K1) OX=272557 GN=APE_1323.1 PE=4 SV=2
MEGDRTRIMDLREGLRNVSISGRVLETGEPKMVETKRGPATLSEAVVGDESGRVKVTLWG
SHAGTLKEGEAVRIEGAWTTSYRGKVQVNVGRESTIEKVDSEDVPQAEDIPEEMPEAQYR
GFGQRQPYRSGGFGGFRGGGYQPRRGGRRRF
>tr|H0AEU0|H0AEU0_HALSG Single-stranded DNA-binding protein OS=Haloredivivus sp. (strain G17) OX=1072681 GN=HRED_00678
PE=4 SV=1
MPEMNVEELSPEADKVEITGKISELPTPRAVSTRYGQKKIVTVVFEDETGSIDLTLWEEE
IDAIEEGSKVHIQGAYVREWAEDIQLNIGRDGDIEQVEE
```



Kingdom	Protein	Gene	Len	% Cov	Organism	Protein Description	Protein Probability	Top Peptide Probability	Total Peptides	Total Spectral Count
Archaea	sp|Q97W7:	ssb	148	51.4	**Saccharolobus solfataricus** (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2) OX=273057	Single-stranded DNA binding protein Ssb	1	0.999	7	176
Bacteria	tr|A0A3P2/	recJ	568	3.9	Conchiformibius steedae OX=153493	Single-stranded-DNA-specific exonuclease R	0.9825	0.999	1	1

Figure F: FilterFasta generated FASTA file led to species ID for SSB protein. Snippet of 18,283 entry FASTA file generated from "Single Stranded Binding" keyword search across the three kingdoms. Results from searching this FASTA file led to a species ID of the SSB protein of *Saccharolobus solfataricus* which was confirmed to be correct.

Ellison Institute