

Assessing and Interpreting Protein Identifications

Sean L. Seymour, Ph.D.

ABRF2010 Sacramento, CA



Three Keys to Improving Your Protein ID

- Use a Protein Inference Tool
- Choosing the Right Database to Search
- Robust Comparison

Key#1: Use a Protein Inference Tool

- Once you have your peptide identifications, the job is not done.
- The process of inferring proteins from these IDs is not trivial, and failure to do this rigorously resulted in many reports of inflated numbers of proteins in the recent past.
- Formal protein inference algorithms are now commonly used to address this problem.
- Examples include ProteinProphet, the Pro Group™ Algorithm, and Scaffold.
- There are two main concerns in reporting protein IDs. Are you reporting:
 - **The right number of detected proteins?**
 - **Ambiguity among multiple accessions appropriate for each detection?**
- The ambiguity point may not seem that important until you or someone else tries to do anything with the list of proteins that you reported.

Example: One Detected Protein with 4-fold Ambiguity

Protein Group 15

Proteins in Group				
Unused	Total	Accession #	Name	Species
12.10	12.10	spt P48644	Aldehyde dehydrogenase...	Bos taurus
0.00	12.10	rf NP_7766...	aldehyde dehydrogenase...	Bos taurus
0.00	10.10	spt P51977	Aldehyde dehydrogenase 1...	Ovis aries
0.00	10.10	gb AAA854...	aldehyde dehydrogenase	Ovis aries

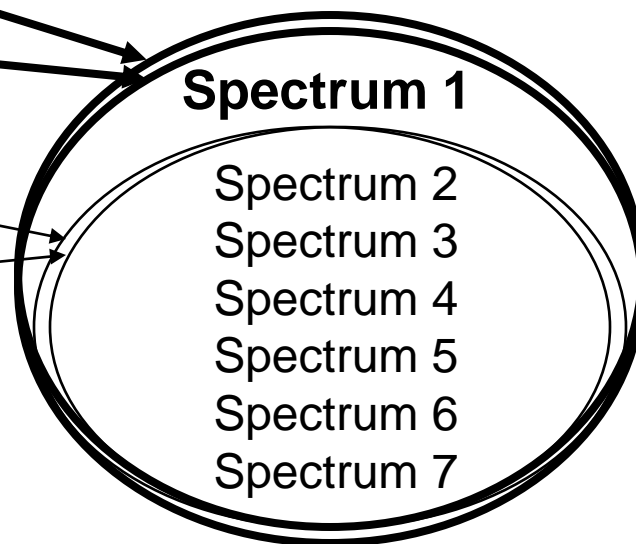
Peptides in Group										
Contrib	Conf	Sequence	Modifications	Cleavages	ΔMass	Prec MW	z	Sc	Spectrum	
2.00	99	DHLLLA TME AMNGGK			0.0155	1599.7904	3	14	1.1.1.1640.2	
2.00	99	IFINNEWHS SVSGK			-0.0012	1616.7936	3	14	1.1.1.1439.3	
2.00	99	LCEVEE GDKED VDK	Carboxamidomethyl(C)@2		-0.0100	1663.7147	3	14	1.1.1.1153.3	
2.00	99	YV LGNP LTP GVSQG PQ IDKEQ ...			-0.0073	2659.3420	3	13	1.1.1.1595.4	
1.70	98	LFVEES IYDEFVR			0.0183	1644.8218	2	10	1.1.1.1825.2	
1.30	95	KFPVFN PATEEK		missed K-F...	0.0147	1405.7389	3	9	1.1.1.1412.3	
1.10	92	ELGEYG FHEYTEVK			0.0027	1699.7758	3	9	1.1.1.1482.2	
0.00	63	LCEVEE GDKED VDK	Carboxamidomethyl(C)@2		0.0033	1663.7281	3	9	1.1.1.1156.3	

Equivalent Winner Proteins

spt|P48644
rf|NP_776664.1

Competitor Subset Proteins

spt|P51977
gb|AAA85435.1



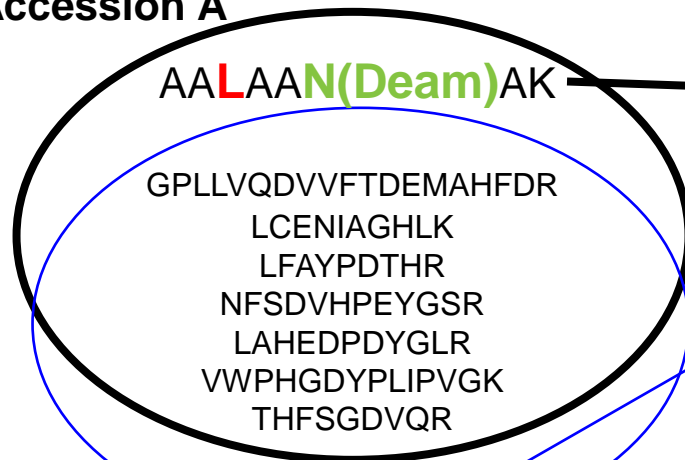
Differences between Protein Inference Tools

- Some find the minimal protein set explaining observed peptide sequences, some minimize vs. the observed spectra.
- Some keep only one answer per spectrum, some preserve peptide ambiguity.
- Some threshold on peptide confidence, some do not.
- Some report larger amounts of accession ambiguity than others.
- Several approaches to ranking reported protein groups.

Sequence-Centric

Spectrum-Centric (Better solution)

Accession A

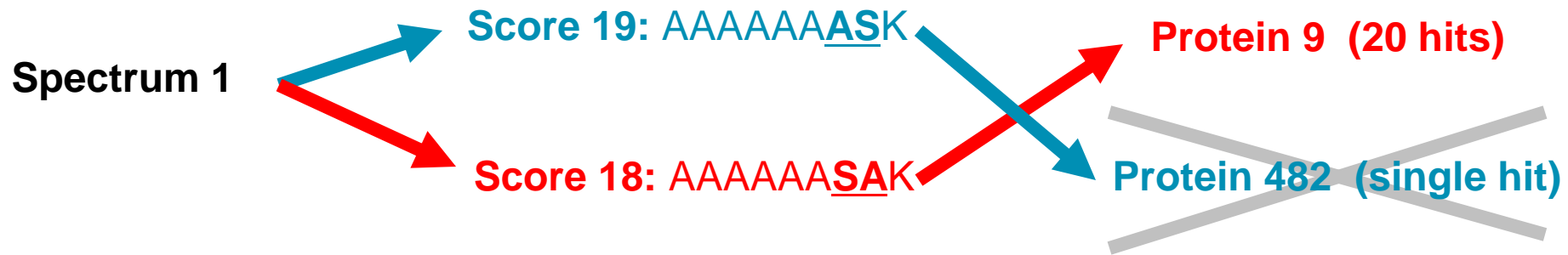


AA|IAADAK

Accession B

spectrum 1849.2
spectrum 1081.3
spectrum 1287.2
spectrum 1318.3
spectrum 1274.2
spectrum 1247.2
spectrum 1081.3
spectrum 1081.3
spectrum 1081.3

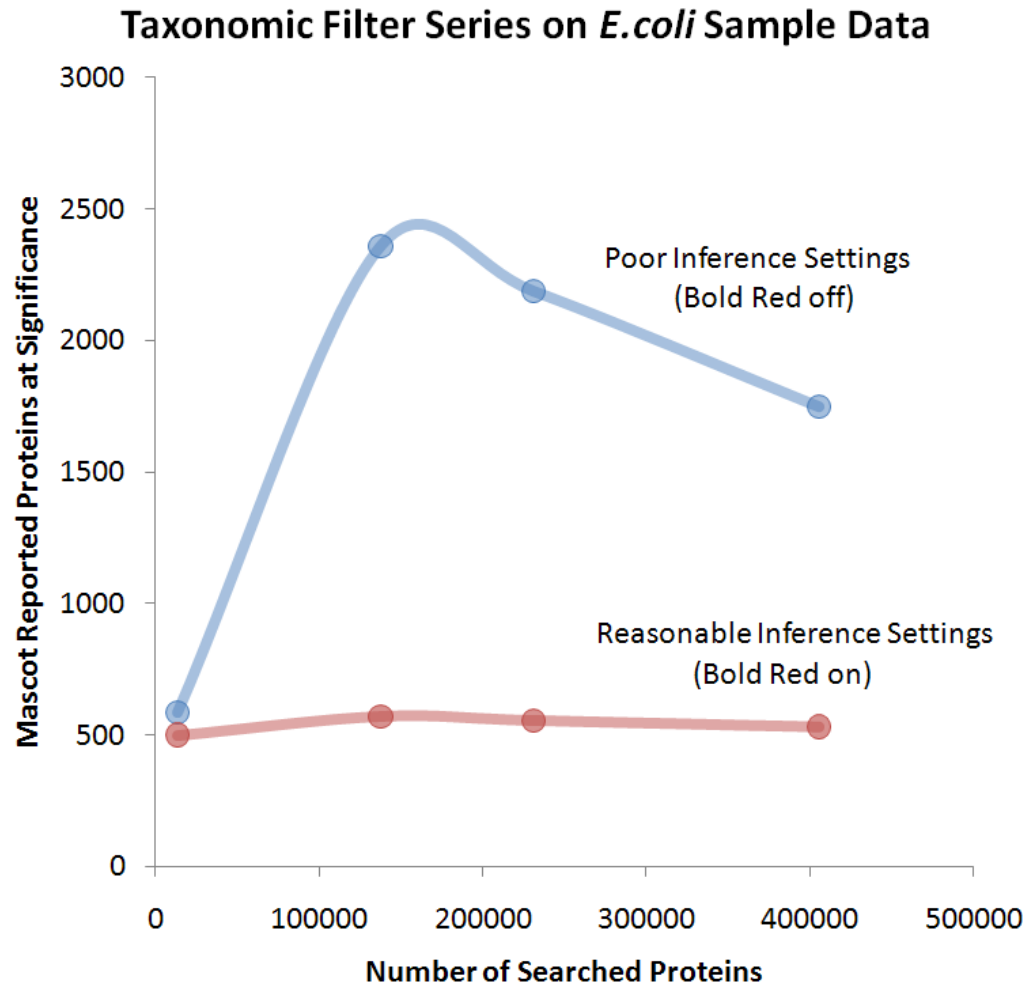
Thresholding Peptides Leads to Protein Errors



- Leads to reporting a wrong or unnecessary protein, thus overall inflation.
- The point is that you can't decide the peptide level until you see the protein level!
- There are many sources of very similar scoring answers pointing to different proteins to cause this: deamidation, switched residues, mutations, tryptic vs. semitryptic context, or simply random hits.
- This problem will be particularly bad with large data sets and/or large databases.

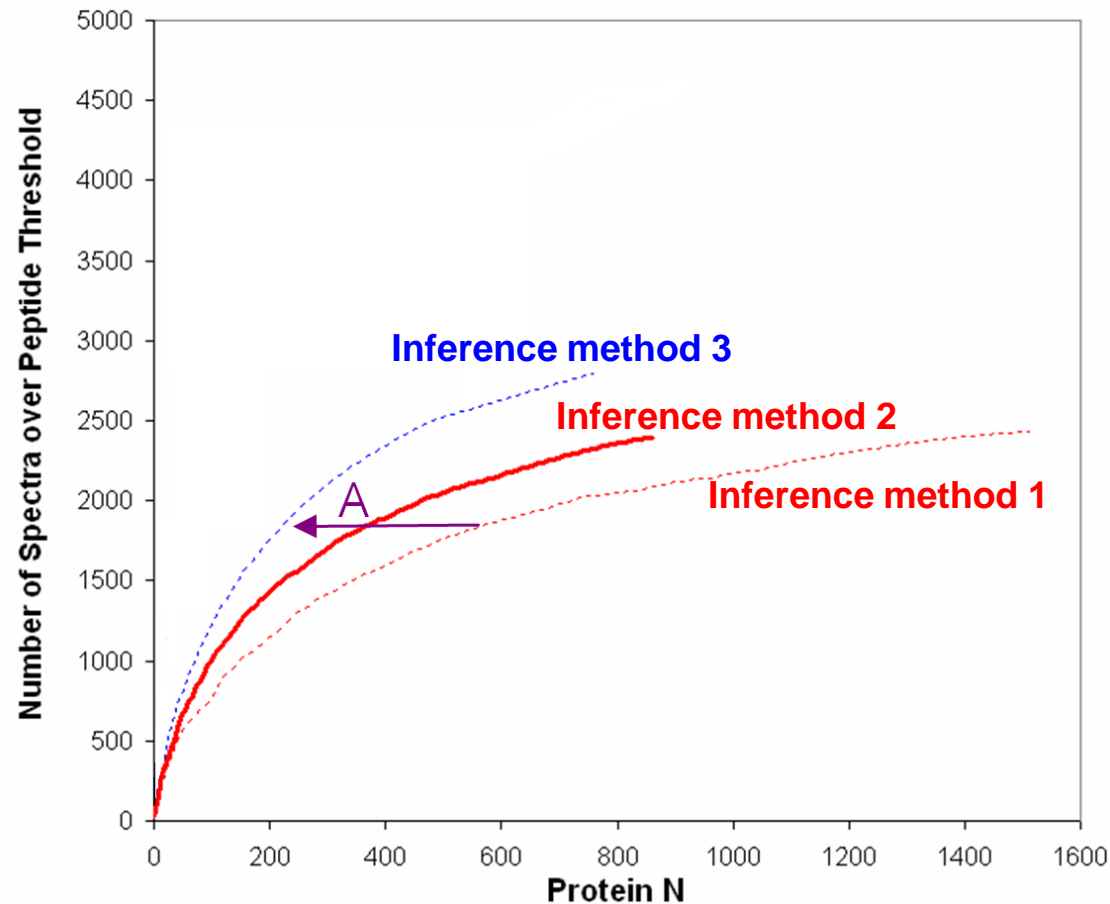
Detecting Gross Protein Inference Problems

- If protein inference is really bad or simply not done, the number of reported proteins can depend strongly on the size of the database.
- Example: *E.coli* data searched against UniProtKB/Swiss-Prot database with species filter progression:
 - *E.coli* (has most the right answers)
 - All Proteobacteria
 - All bacteria
 - All species
- When inference is reasonable, the trend is quite flat. (red line)
- Not so when the proper constraint is not in place. (blue line)



How to Compare Protein Inference Methods

- To compare two protein inference methods that don't have gross problems, a slightly more refined approach is needed.
- Given two searches with identical peptide search space and different inference approaches:
 - Fix a peptide ID quality threshold – ex. 5% local FDR.
 - Starting with the highest ranked protein, count the cumulative number of good spectra (not peptides) explained by at least this threshold ID quality.
 - The best inference method is the curve closest to the top left of the plot because it needs fewer proteins to explain a given number of spectra.



Key#2: Choosing the Right Database to Search

- Main factors:
 - Provider of the database – EBI, NCBI, species consortia, etc.
 - Species constraints – Exact species, similar species, or everything?
 - Elaboration of sequences – Include isoforms, unnecessarily redundant?
- Maximize sensitivity and specificity of the search by:
 - Including high numbers of proteins that could be in the sample
 - Not including high numbers of proteins not likely to be in the sample.
- The best database for you will depend on the organism and the specifics of your research.

Create Tailored Databases – UniProt.org Example

Part 1 – Select a set of proteins

Species **Macaca mulatta (Rhesus macaque)** ★

UniProtKB (3,599) | Taxonomy help

Mnemonic	MACMU
Taxon identifier	9544
Scientific name	Macaca mulatta
Common name	Rhesus macaque
Synonym	-
Other names	<ul style="list-style-type: none"> rhesus macaques rhesus monkey rhesus monkeys
Rank	Species
Lineage	<ul style="list-style-type: none"> cellular organisms Eukaryota Fungi/Metazoa group Metazoa Eumetazoa Bilateria Coelomata Deuterostomia Chordata Craniata Vertebrata Gnathostomata Teleostomi Euteleostomi Sarcopterygii Tetrapoda Amniota Mammalia Theria Eutheria Euarctontoglires Primates Haplorrhini Simiiformes Catarrhini Cercopithecoidea Cercopithecidae Cercopithecinae Macaca

View Taxonomic Lineage for Term to Broaden Search

Search in: Protein Knowledgebase (UniProtKB)

Query: []

Field: Taxonomy [OC] | Term: catarrh

Options: Catarrhini [9526], Veronica catarractae [74690], Moraxella catarrhalis [480], Deer malignant catarrhal fever virus [104223]

120,505 results for taxonomy:catarrhini in UniProtKB

Field: Reviewed (yes/no) | Choose: yes

26,094 results for taxonomy:"Catarrhini [9526]" AND reviewed:yes in UniProtKB

Very complex sets can be created via AND/OR

Search in: Protein Knowledgebase (UniProtKB)

Query: (taxonomy:catarrhini AND reviewed:yes) OR taxonomy:macaca

1 - 100 of 35,798 results for (taxonomy:catarrhini AND reviewed:yes) OR taxonomy:macaca in

Complex Sets of Proteins Can Be Selected

- Examples of studies where complex databases may be appropriate:
 - Infection: human + specific virus
 - G.I. track: human + bacteria.
- Search engines don't have this kind of control but carefully tailored databases can give better results.

Search in: Protein Knowledgebase (UniProtKB) | Query: ((taxonomy:"Homo sapiens [9606]" AND reviewed:yes) OR taxonomy:"Dengue virus group [11052]"

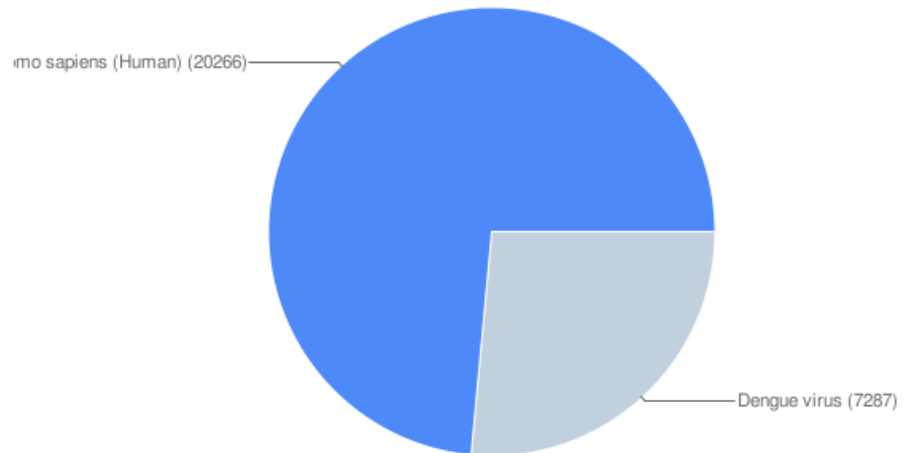
1 - 100 of 27,553 results for (taxonomy:"Homo sapiens (Human) [9606]" AND reviewed:yes) OR taxonomy:"Dengue virus group [11052]" in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%



27,553 results for ((taxonomy:"Homo sapiens [9606]" AND reviewed:yes) OR taxonomy:"Dengue virus group [11052]" in uniprot browsing by taxonomy

- + Dengue virus 7,287
- + Homo sapiens (Human) 20,266



Create Tailored Databases – UniProt.org Example

Part 2 – Download the database

- Very easy to download a FASTA database for the set of proteins you have selected.
- Choice between canonical vs. canonical and isoforms.
- The UniProtKB/Swiss-Prot you download from the ftp does not have isoforms.

FASTA

20,265 proteins



Canonical sequence data in FASTA format.

[\[Download \(10 MB*\) | Open | Open first 10 \]](#)

35,773 proteins



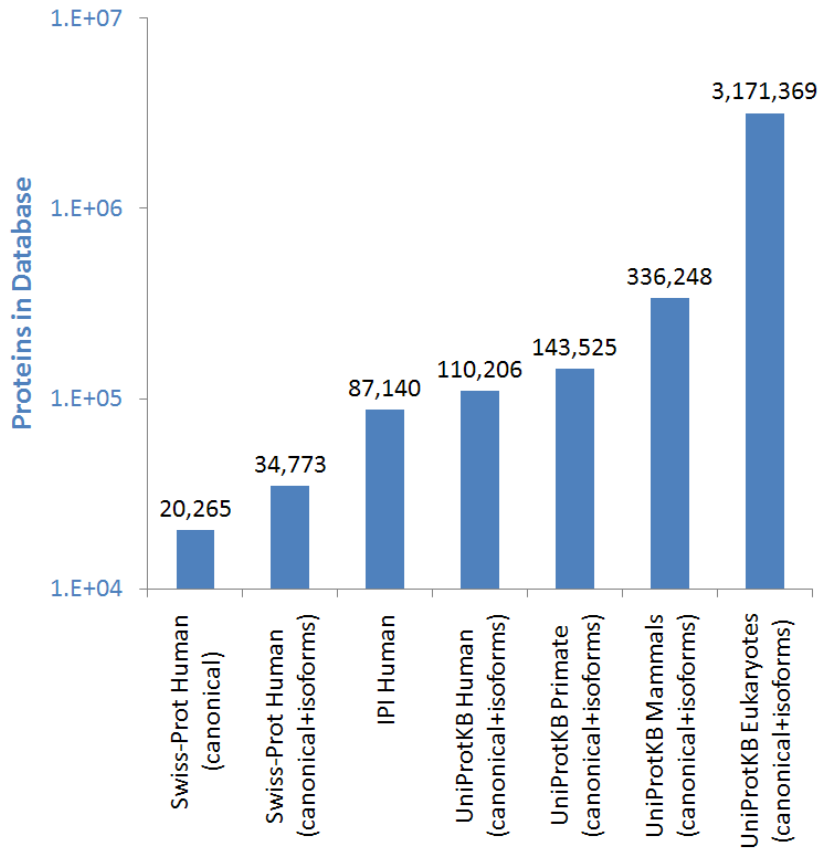
Canonical and isoform sequence data in FASTA format.

[\[Download \(10 MB*\) | Open | Open first 10 \]](#)

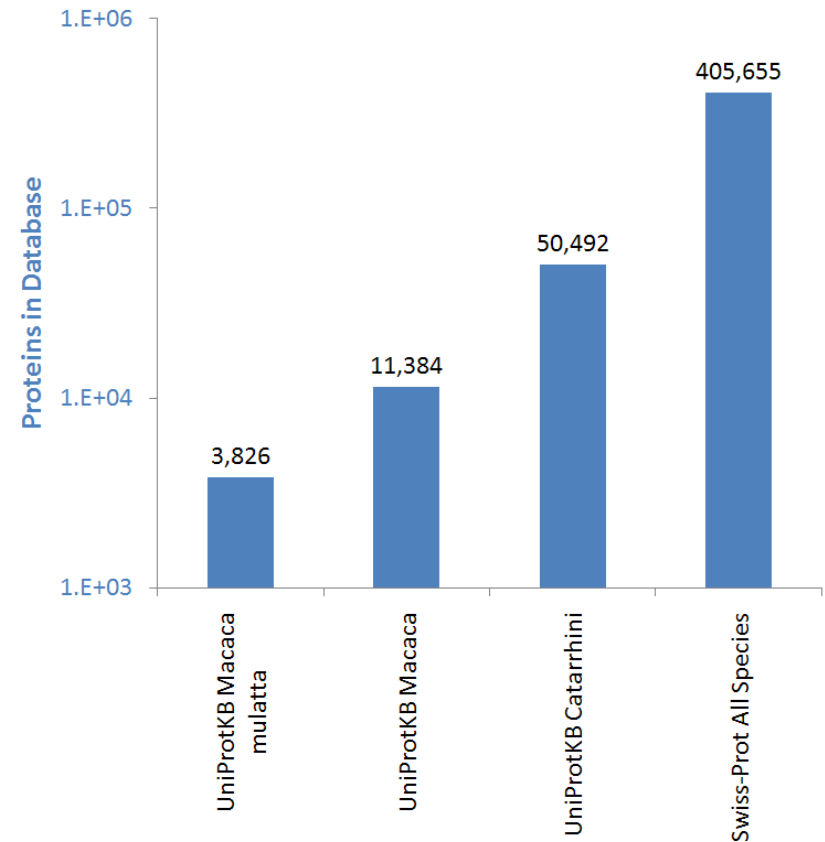
Many Options, but What Works Best?

- With tools like UniProt.org, you can easily get any subset or complex combination of proteins with various constraints.
- What's in the database you usually download? Is there a better approach? How to decide?

Database Size Progression for Human Samples



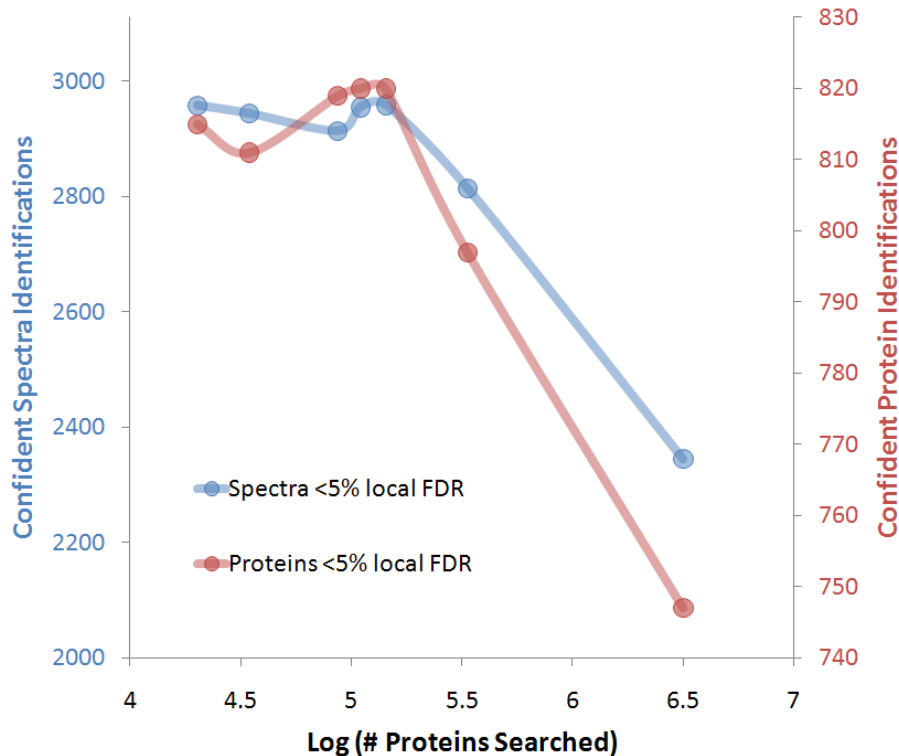
Database Size Progression for Monkey Samples



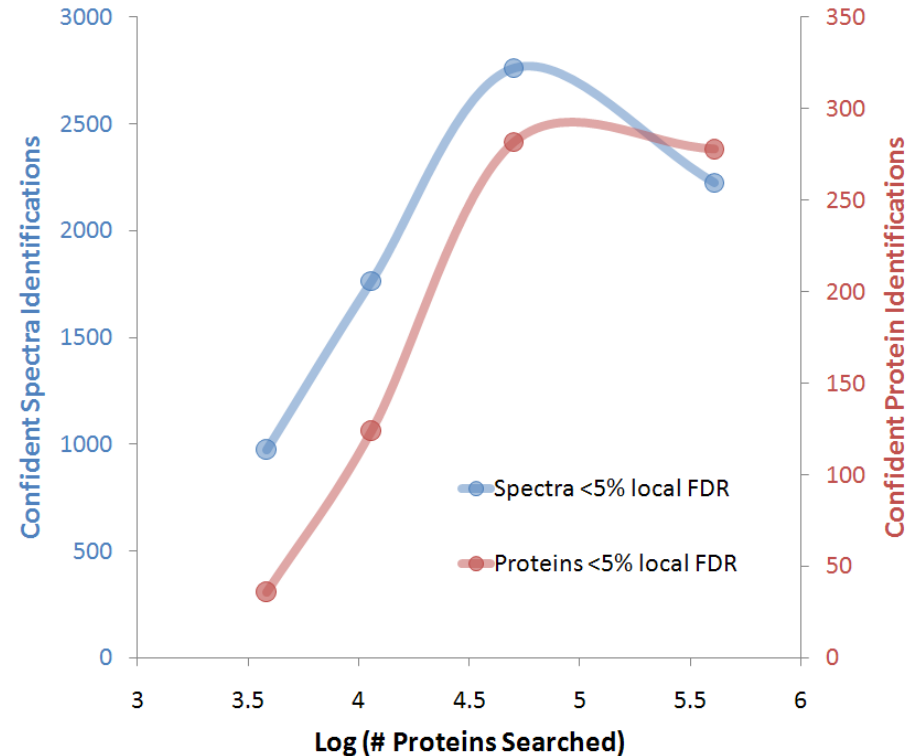
Species-Dependent Impact of Broader Search

- Taking an example data set from these two species and running searches against each database in the appropriate series, we get the results shown below. (Paragon™ Algorithm Rapid mode)
- Human is well represented in the small DB limit, while the monkey species is not, and thus benefits from searching broader species sets.
- Decline occurs in both when additional sequences only add noise and few new correct sequences.

Taxonomic Filter Series on Human Sample

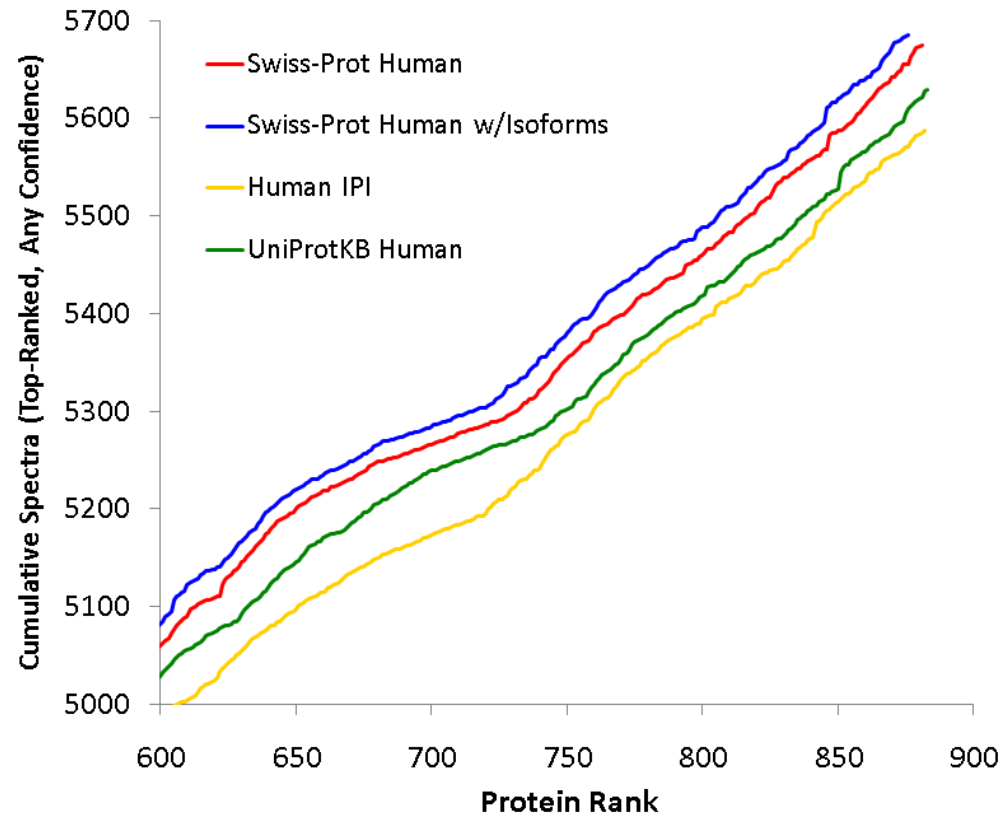


Taxonomic Filter Series on Monkey Sample



Fine Comparison to Select Optimum Database

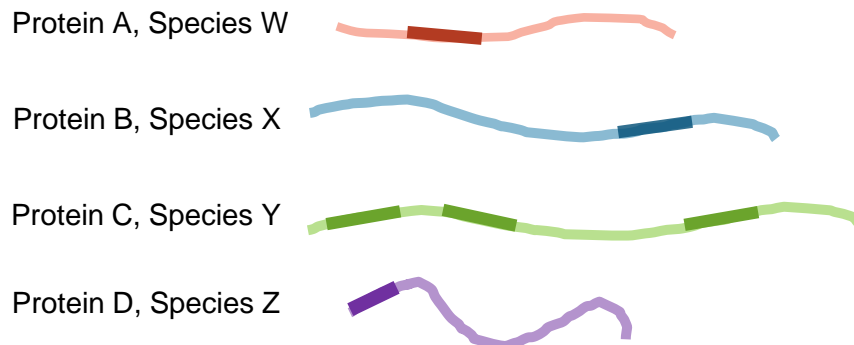
- How do we decide which of the four similar all-human databases is best to search data from our sample?
- Maximal proteins not necessarily the best.
- A higher portion of correct isoforms would explain more spectra per protein.
- Count cumulative spectra accounted for, regardless of confidence, to avoid the database size affect on confidence threshold.
- Plot this vs. protein rank to avoid the issue of different numbers of confident proteins in each final list.
- This suggests that Swiss-Prot with the isoform component is explaining more spectra per protein, despite yielding a smaller final list.



Protein Inference Can Be Meaningless

- If the sequence database is relatively incomplete for the organism of interest, this generally necessitates searching related species as well, as with the monkey sample data we just saw.
- If substantial additional identifications come extra-species, the protein inference results should not be considered meaningful – particularly the number of reported proteins.
- When this occurs, maximize confident spectra identifications.

Multi-species search result against present-day database, which lacks the true protein sequence:



How these same 6 peptides will be identified in the future we have this protein in our database:



Append a Contaminant Database

- Add sequences of common contaminants to your database
 - Ex. pig/cow trypsin, human keratins, BSA, etc.
- By correctly accounting for the spectra acquired on these proteins, you:
 - Avoid wrong answers to these spectra from the expected species
 - Achieve higher spectral utilization (how gain much depends on sample).
- Make sure your filter/parsing settings allow search of these proteins.
- Searching all species is a terrible way to get this additional handful of proteins. Appending a contaminant database preserves search specificity.
- Several places to get contaminant databases:
 - cRAP database – <http://www.thegpm.org/crap/index.html>
 - Provided with software – ex. ProteinPilot™ Software Help folder
 - Search setting in search engine (ex. GPM)

Key#3: Robust Comparison

- The practical goal of identifying proteins is to do something with the results and that usually means comparison in some form.
- Important example: creation of a protein feature table in a biomarker study by alignment of many results.
- This is why reporting accession ambiguity matters!
- A major point of the iPRG2008 study was to assess if people were reporting accession ambiguity.

Correct Comparison Requires Accession Ambiguity

Simplistic First Protein Comparison

Sample 1	≠	Sample 2
Protein Group #8: plectin 1 IPI:00398775.3 (isoform 2)		Protein Group #3: plectin 1 IPI:00186711.3 (isoform 6)

Incorrectly conclude we have found different proteins.

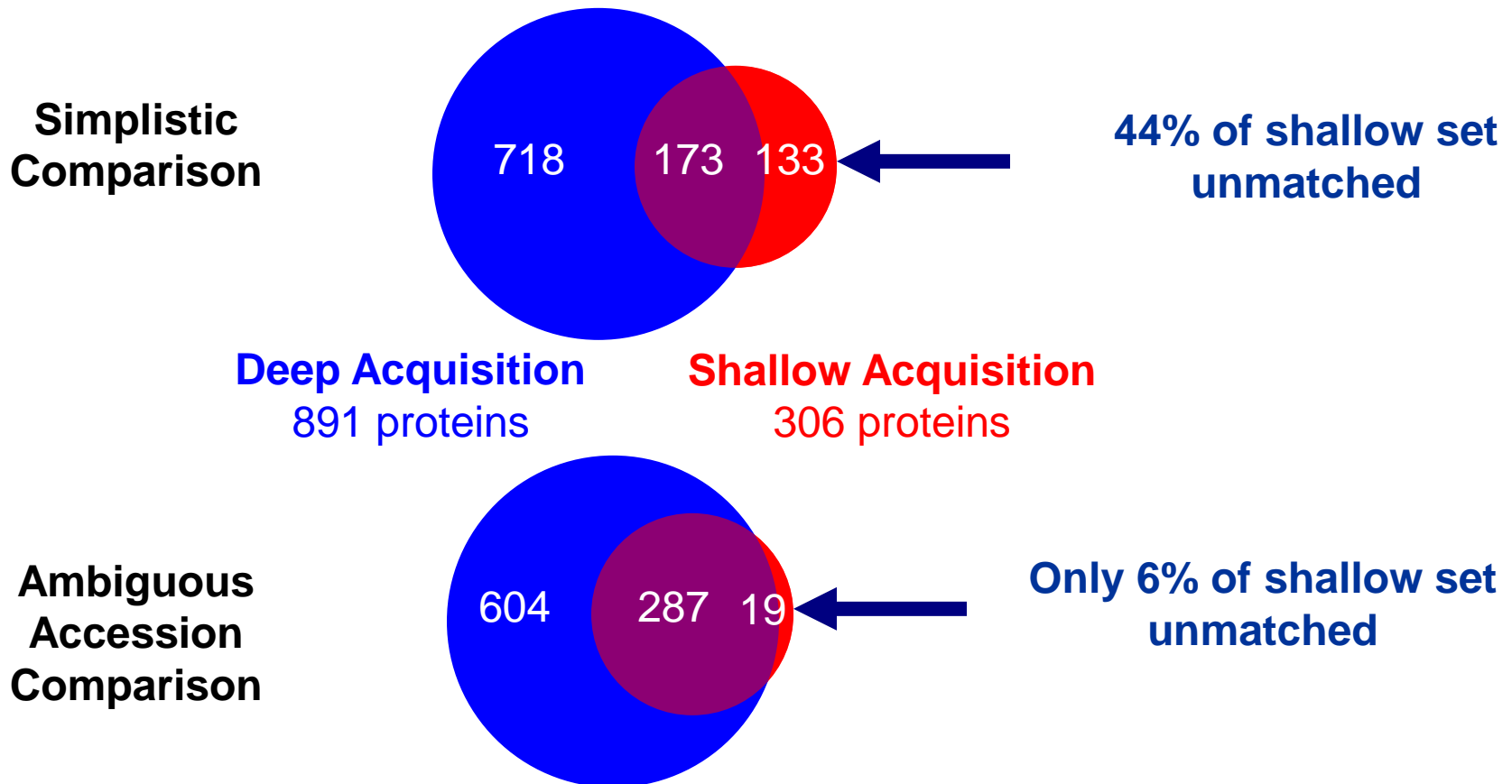
Comparison with Ambiguous Accessions in Protein Groups

Sample 1	=	Sample 2
Protein Group #8: plectin 1 IPI:00398775.3 (isoform 2) IPI:00420096.4 (isoform 3) IPI:00186711.3 (isoform 6)		Protein Group #3: plectin 1 IPI:00186711.3 (isoform 6) IPI:00398775.3 (isoform 2) IPI:00420096.4 (isoform 3)

Correctly conclude we have found the same protein in both samples, although we're unsure of which isoform it is.

Bad Comparison Hurts Reproducibility

A test of comparison methods: deep and shallow acquisitions of the same sample:



Conclusions

- Using a protein inference tool is critical.
- Search results can be dramatically improved by careful selection of the set of proteins you search against.
- Reporting ambiguity among protein accessions (a group, not a single accession) enables better comparison.
- Optimal comparison of protein lists is critical for many kinds of proteomics experiments.

Trademarks/Licensing

For Research Use Only. Not for use in diagnostic procedures.

The trademarks mentioned herein are the property of AB Sciex Pte. Ltd. or their respective owners. AB SCIEX™ is being used under license.

© 2010 AB SCIEX.