Improvements in the Selection of Peptide Biomarkers for Environmental Proteomics

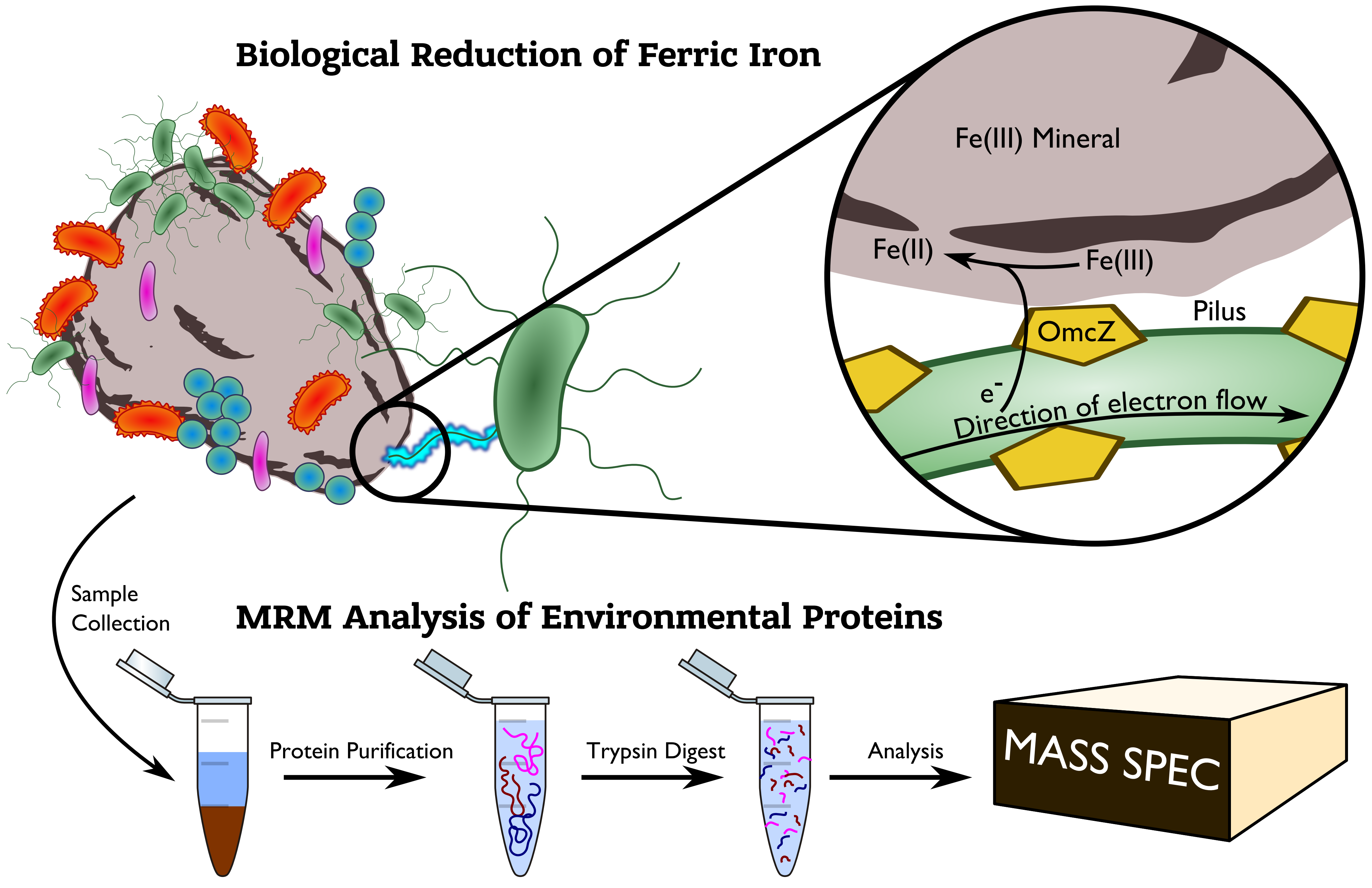
Alex Hernsdorf
awh58@cornell.edu
ENVE '14

Background and Objectives:

Targeted proteomics experiments allow for accurate estimation of protein abundance from the analysis of peptides present in biological samples.[1] These peptides serve as **biomarkers** that give investigators an estimate of biological activity in environments like the human body or bioremediation sites. Focusing on a finite number of peptides in targeted proteomics experiments gives high reproducibility and a better signal-to-noise ratio than discovery-based experiments. **Multiple reaction monitoring (MRM)** is an ideal mass-spec (MS) workflow for biomarker assays due to its high selectivity, sensitivity, and robustness. Successful implementation of MRM requires the selection of **proteotypic peptides (PTPs)**, meaning that they are likely to be observed with current MS-based proteomics methods (not all peptides produce strong signals).[2] However, physical and biochemical factors complicates the prediction of which peptides are PTPs. As such, the best resource for selecting PTPs is existing MS data from prior proteomics experiments. Regrettably, this data is often unavailable, and so predictions based on sequence are necessary.

PTP selection is convoluted in **environmental proteomics**, when biological activity may be attributed to many species. In this instance, an investigator may want to select PTPs from functionally similar proteins (**orthologs**) expressed by different organisms. MRM analysis can be simplified by selecting peptides with **conserved sequence** across a list of orthologs. To simplify the task of selecting PTPs for MRM, I have developed analytical tools that integrate existing MS data with sequence information from bioinformatics databases and present results in graphical and tabular output.

Biological Reduction of Ferric Iron



MRM Analysis of Environmental Proteins

1. Generation of Candidate Peptides:

First, a tabular list of **candidate peptides** is generated from a list of target proteins grouped by orthologous cluster. The amino acid sequences of each protein are retrieved and subjected to an *in silico* **tryptic digest** to find all peptides predicted from sequence. Peptides are filtered according to length and (optionally, see right) amino acid content. Lists of peptides are output along with MS data (if available) and grouped by ortholog.

Peptide Discovery Tools:

To reduce the number of unique peptides, PTPs that indicate the presence of multiple orthologs can be selected. When the peptides predicted from the sequence of target proteins are insufficient, new peptides can be discovered using sequence analysis via NCBI BLAST. The protein sequences of orthologs are first downloaded and digested. Two options were developed for discovering new peptides from the protein sequences of these orthologs:

Most common peptides are identified among the orthologs predicted by BLAST. These peptides are then searched for uniqueness using the Peptide Match tool. This is the preferred PTP selection technique, and it works particularly well when a large number (>10) of orthologs are found.

Protein alignments are used when the above technique fails, showing where the predicted peptides lie on the protein sequences from which they derive.. Proteins are first aligned using MAFFT, and then peptides are overlaid onto the sequence alignment and visualized in a graphical output.[4,5] This technique allows selection of PTPs with roughly equivalent length and location in the protein sequence, but is limited in the number of proteins (<10) that can be assessed simultaneously.

2. Peptide Filtering:

Peptides with residues undergoing easy modification excluded if possible:

Met (M), Trp (W) - easily oxidized
Gln (Q), Asn (N) - converted to acidic forms
N-terminal Gln (Q) and Glu (E) - converted to pyro-glutamate
Cys (C) - covalent linkages
His (H) - extra charge on peptide

Peptides with two **neighboring basic AAs** at cleavage sites excluded

Length must be between **7 and ~15 AAs**

3. Checking Peptide Uniqueness:

Given the enormous diversity of bacteria and proteins in environmental samples, non-unique peptides pose a risk for **interference** in MRM analysis. To identify non-unique peptides, the Protein Information Resource developed **Peptide Match**, a web tool that retrieves "all occurrences for a given query peptide from the UniProtKB protein sequences".[3] The Peptide Match API is queried with the list of candidate peptides and the output is visually inspected to determine whether the uniqueness, or lack thereof, of the peptide is satisfactory. As mentioned previously, it may be valuable to select peptides that are conserved among orthologs of phenotypically similar organisms.

Commonality-Based Peptide Discovery

Top Orthologs (NADH:flavin oxidoreductase)		Peptides	BLAST Hits		Matching Peptides
Desulfotomaculum reducens	*In silico* digest →	TDEYGGSLEN**R**	9	Peptide Matching →	Non-specific
Desulfosporosinus orientis		DLGSST**R**	8		Non-specific
Desulfosporosinus meridiei		EIPHELTVEEI**K**	7		2 *Desulfitobacterium* (specific)
Desulfitobacterium dichloroeliminans		IAIIGGGLPGCELA**K**	7		2 *Desulfitobacterium* (specific)
Desulfurispora thermophila		FFYTPDLA**R**	5		4 *Desulfitobacterium* (conserved)
		IVPQLYQAG**R**	4		4 *Desulfosporosinus* (conserved)

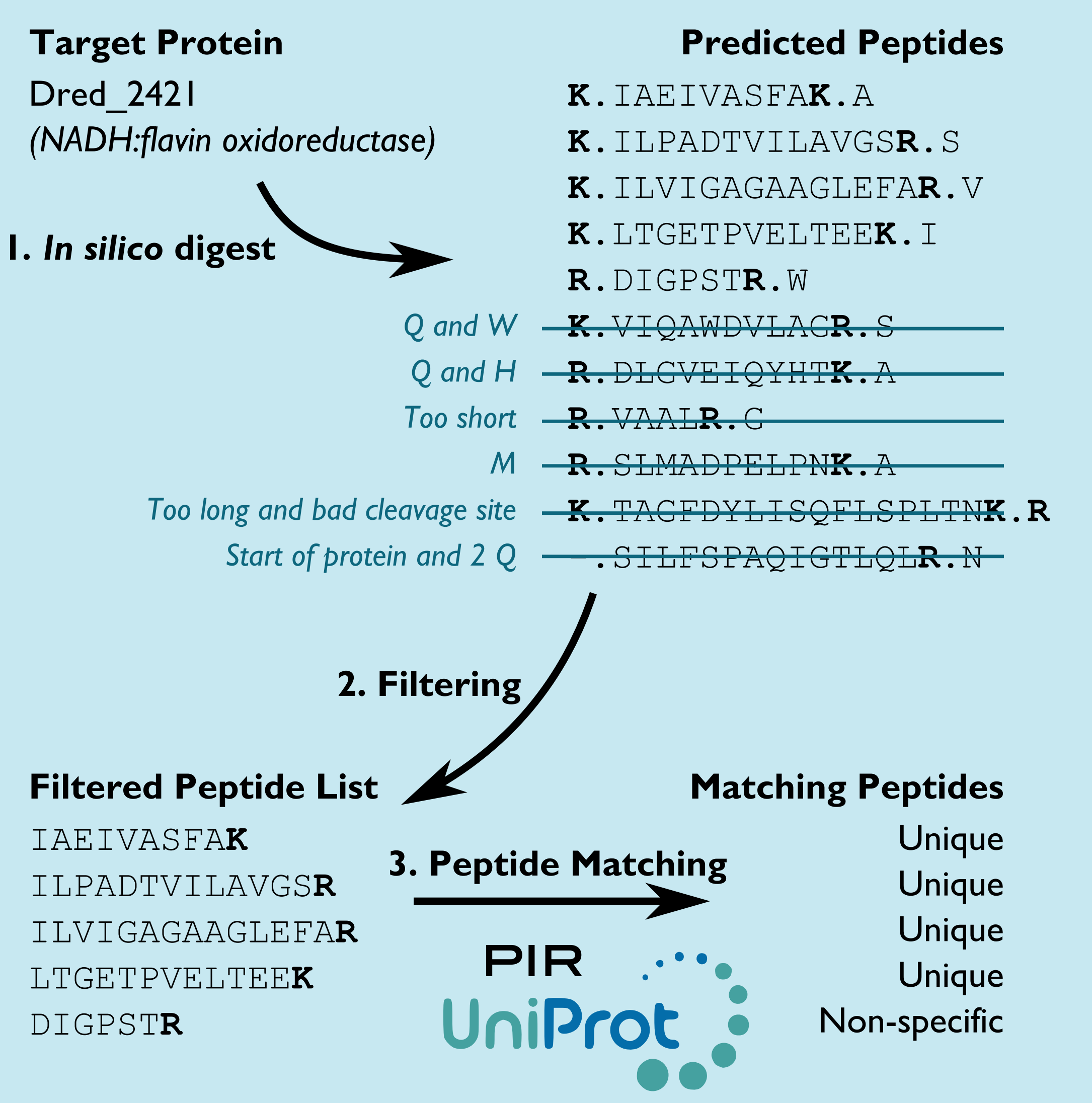
Alignment-Based Peptide Discovery

Orthologs	Sequence		Matching Peptides
GSU_0466	...**R**EDPGPVV**R**PVDDTG**R**Y**K**VTSTAAD**K**YVF**R**SPSL**R**NVAI...	Peptide Matching →	Unique
Gura_1316	...**K**ESPAADV**R**PTEDLG**R**F**K**VTNTAAD**K**YVF**K**SPSL**R**NIEL...		Gmet and DaAHT2 orthologs
Gbem_0020	...**R**EIPTAEI**R**PESDTG**R**F**K**VTNTASD**K**YVF**R**APSL**R**NVAL...		GM21 and Desaf orthologs

Acknowledgements:

The author would like to thank **Professor Ruth Richardson** for her continual guidance and support, microbiology graduate student **Annie Otwell** for inspiring me to study how microbes do what they do, and **Cresten Mansfeldt** for his thoughts on programming and design and for always making sure I'm focused in lab. Funding was generously provided through ELI by **James Moore** '62 EE. Supplies and materials were provided by DOE Genomic Sciences Program Grant #DE-SC0006644.

Filtering of Candidate Peptides and Uniqueness Check



Target Protein
Dred_2421
(NADH:flavin oxidoreductase)

1. *In silico* digest

Predicted Peptides
K.IAEIVASFA**K**.A
K.ILPADTVILAVGS**R**.S
K.ILVIGAGAAGLEFA**R**.V
K.LTGETPVELTEE**K**.I
R.DIGPST**R**.W
Q and W ~~**K**.VIQAWDVLAG**R**.S~~
Q and H ~~**R**.DLGVEIQYHT**K**.A~~
Too short ~~**R**.VAAL**R**.G~~
M ~~**R**.SLMADPELPN**K**.A~~
Too long and bad cleavage site ~~**K**.TAGFDYLISQFLSPLTN~~**K**.**R**
Start of protein and 2 Q ~~-.SILFSPAQIGTLQL**R**.N~~

2. Filtering

Filtered Peptide List		Matching Peptides
IAEIVASFA**K**	3. Peptide Matching → PIR UniProt	Unique
ILPADTVILAVGS**R**		Unique
ILVIGAGAAGLEFA**R**		Unique
LTGETPVELTEE**K**		Unique
DIGPST**R**		Non-specific

Conclusions and Future Work:

This research focuses on the optimization of a procedure that incorporates empirical data with *in silico* predictions to select biomarkers that are both proteotypic and representative of proteins from important biodegradation pathways. The software developed provides researchers with a fast protocol for selecting PTPs, a necessary procedure for protein abundance experiments.

Currently, the process outlined in the Methods sections requires familiarity with MATLAB and Python. However, it is predicted that some users may not have coding experience, and so a GUI or web tool will be developed to integrate the various components. Furthermore, programs initially developed in MATLAB will be rewritten in Python and put online so that they may be freely accessible for other researchers.

References:

1. James A, Jorgensen C. Basic Design of MRM Assays for Peptide Quantification. In: Cutillas PR, Timms JF, eds. LC-MS/MS in Proteomics.Vol 658. Methods in Molecular Biology. Totowa, NJ: Humana Press; 2010:167–185.
2. Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. Rapid Commun mass Spectrom. 2005;19(13):1844–50.
3. Chen C, Li Z, Huang H, Suzek BE, Wu CH. A fast Peptide Match service for UniProt Knowledgebase. Bioinformatics. 2013;29(21):2808–9.
4. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
5. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics. 2012;28(23):3144–6.