



# Protein Identification using Mass Spectrometry data and Genome Fingerprint Scanning

Abdoulaie F. Lowe Nicolas\*, Jainab Khatun and Morgan C. Giddings

Department of Microbiology and Immunology, University of North Carolina at Chapel Hill  
Chapel Hill, NC 27599

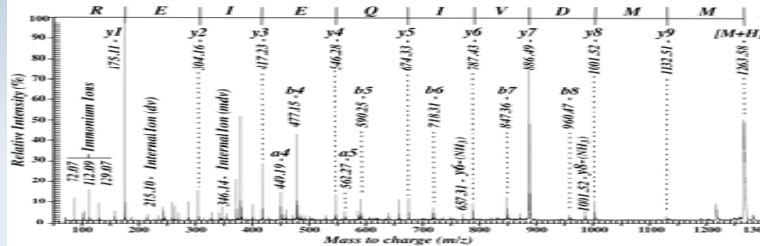
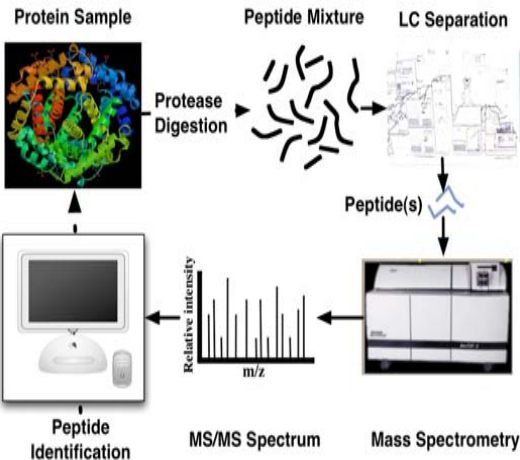


## BACKGROUND

Peptide mass fingerprinting is a principal protein identification technique that was introduced in 1993 by several groups. Our method identifies proteins by locating the genomic origins of sample proteins by scanning their peptide-mass fingerprint against the theoretical translation and proteolytic digest of an entire genome. Also, it identifies proteins by identifying each individual peptides which was originated from the complex mixture of proteins.

## PURPOSE AND HYPOTHESIS

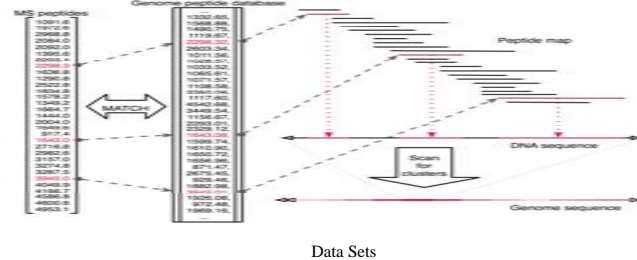
Use proteomics to compare protein expression profiles of different diseases to healthy samples. This can potentially identify cause and the origin (gene and its location) of diseases.



A MALDI TOF/TOF MS/MS spectrum for the peptide MMDVIQEIER. The precursor mass for this spectrum is 1263.54. Some of the common ion types are labeled, and the constructed peptide sequence is shown using the observed y-ions, which are generally the most intense. Note that peak intensities for all ion types are quite variable, and many of the expected ions are not observed. Furthermore, there are many peaks that remain unassigned/unidentifiable, possibly the product of secondary reactions in collision induced dissociation.

## MATERIALS AND METHODS

Our software (GFS) generates a database of all possible peptides that might be produced by the organism under study and match the mass spectrometry measured mass list against the database. We then map back the positions on the chromosome sequence hitting the matching peptides and find cluster of fragments in a confined region which may be the gene encoding the protein present.



Origin	Description	URL
Open Proteomics Data	SqCC human cell line used to model head and neck cancer. Grown under normal conditions. No perturbation.	<a href="http://apropos.icmb.utexas.edu/OPD">http://apropos.icmb.utexas.edu/OPD</a>
Open Proteomics Data	SiHa human cell line used to model cervical cancer. Grown under normal conditions. No perturbation.	<a href="http://apropos.icmb.utexas.edu/OPD">http://apropos.icmb.utexas.edu/OPD</a>
Peptide Atlas	Human prostate cancer cell lines: LNCaP, CL-1, microsomal. Labeled with old ICAT.	<a href="http://peptideatlas.org/repository">http://peptideatlas.org/repository</a>
Peptide Atlas	Human prostate cancer cell lines: LNCaP, CL-1, nuclear. Labeled with old ICAT.	

General overview of protein identification using tandem mass spectrometry. Proteins are digested into peptides using a specific enzyme like trypsin. Peptides are separated by one or more stages of separation usually using liquid chromatography, ionized, and have their mass/charge (m/z) ratios measured producing MS spectra showing the precursor masses of peptide ions. In tandem mass spectrometry individual peptide ions are chosen for fragmentation usually with collision-induced dissociation, that cleaves the peptides primarily along the peptide backbone. Peptide fragment m/z ratios are then measured, producing MS/MS spectra, used to identify a peptide.

## RESULTS

We ran GFS using some of our data and found the following locations as possible hits for those spectra. However, since we ran the program with only 2 chromosomes, the obtained hits may not be the actual hits for those spectra. The actual hits may lie in some other chromosomes, which we did not have time to run against all chromosomes.

GFS takes about 15 hours to run a single spectrum to run against whole Human Genome. Therefore, it was not possible to run Whole Human Genome.

Sequence	Precursor neutral mass	Theoretical peptide mass	Peptide sequence	Peptide start	Peptide stop	Tag sequence	Score	HMM Score
chr2 rev-comp	1363.3322	1364.818	NIYSLVFLKRR	160372647	160372680	LV	9.3201	28.5919
chr2 rev-comp	1363.3322	1363.7975	KHKPALVLAP	72343810	72343852	VL	7.3229	31.1341
chr2 rev-comp	1363.3322	1364.3838	SFFFEVIRCLG	57888940	57888904	VI	9.4577	34.9124
chr11	1363.3322	1364.7087	SVYQIVPEPQ	32331705	32331738	IV	10.317	5 29.2852
chr11	1363.3322	1362.7256	NPHLVVVVAT	74285083	74285119	LV	8.5434	25.9985

## CONCLUSIONS + FUTURE DIRECTIONS

We identified possible locations in the genome but need to run against all chromosomes. Download more data for diseased tissue. Run GFS for normal and diseased tissue proteomics data against whole human genome. Compare Results for different proteomics data.

## ACKNOWLEDGEMENTS

\*AFLN was supported by the Shaw University-University of North Carolina at Chapel Hill Undergraduate Program in Prostate Cancer Research and Training (SUUPPRT), funded by the Department of Defense Prostate Cancer Research Program CDMRP (PC061634).

## REFERENCES

1. Giddings, M.C.; Shah, A.A.; Gesteland, R.F.; Moore, M., "Genome-based peptide fingerprint scanning", *Proc. Natl. Acad. Sci. U.S.A.*, 100, 20-25, 2003.