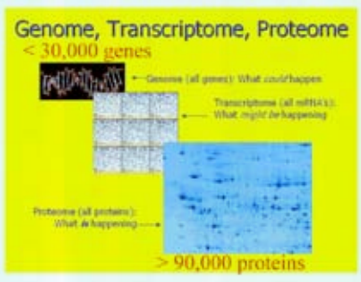
Shotgun Protein Sequencing

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories
Jean-Loup Faulon, Computational Systems Biology Dept.
Sara Gaucher, Zhaoduo Zhang and Anup Singh, Biosystems Research Dept.

Why shotgun protein sequencing?

Problem:



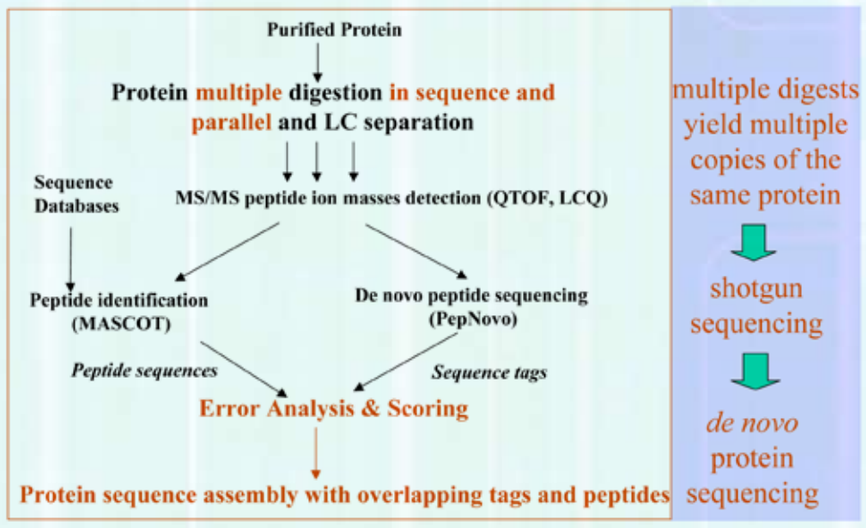
MS techniques relying on genome sequence to identify proteins have a problem…

Solution:

Shotgun protein sequencing for

- Alternative splices forms
- Post-translational modifications
- Proper identification of stop and start codons in genome sequences
- Major histocompatibility complex peptides (adaptive immunity)
- Microbial secretome/peptidome (energy production, water contamination)

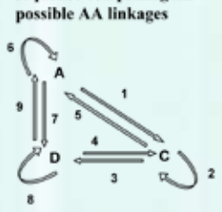
Approach



Developing a scoring metric

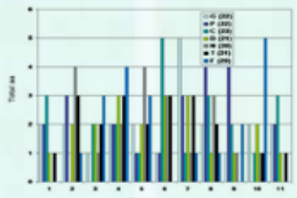
All scoring algorithms (MASCOT, SEQUEST, PepNovo) are biased toward tryptic digests.
What is the rate of false positive (and false negative) with non-tryptic digest?

Designed an *in silico* sequence comprising all possible AA linkages



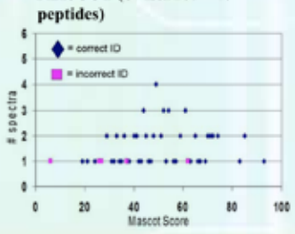
Euler path:
ACCDCAADDA

Synthesized an array of 40 11-mers (Sigma-Genosys)



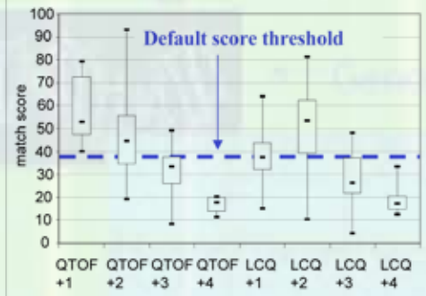
Distribution of AAs is unbiased toward sequence location

Acquired spectra with QTOF and LCQ. Searched Spectra with MASCOT (SwissProt + 40 peptides)



Using assembly to improve scoring accuracy

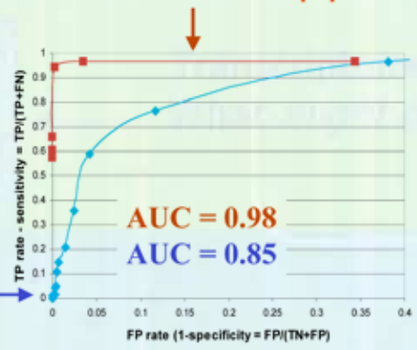
Mascot search: many positives are eliminated (many FNs)



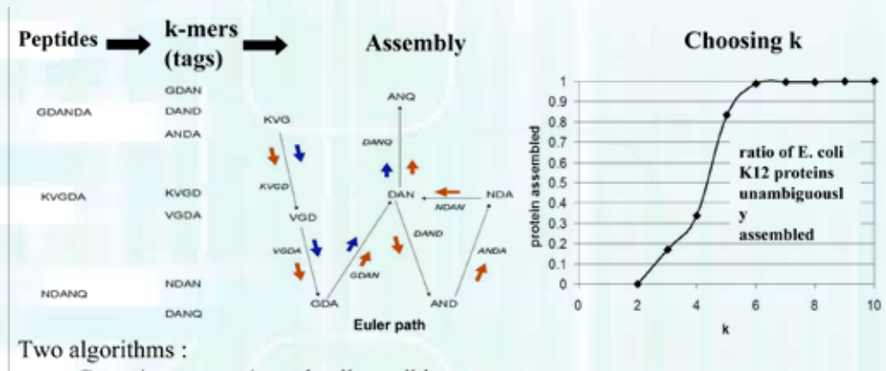
Analytical Chemistry 2007, in press

Selecting all peptides above a given MASCOT score

Selecting all peptides that assemble with another peptide



Assembly algorithms



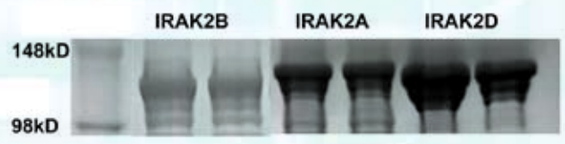
Two algorithms :
- Count/enumerate/sample all possible sequences
- Compute longest unambiguous (unique) subsequence

In silico feasibility study results:

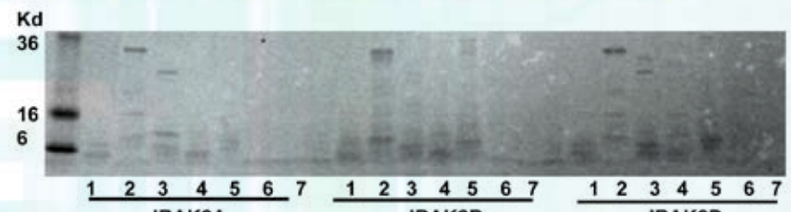
# digest in parallel	15	15	10	10	7	7
% non-identified peptides	0	25	0	25	0	25
average % sequence recovered (tags of length 5)	99±1	95±5	93±7	73±20	90±12	61±20

Sequencing alternative spliced forms

- Purified IRAK2A, IRAK2B, IRAK2D



- Digestion of IRAK2A, IRAK2B, IRAK2D with seven enzymes



1, Trypsin; 2, Lys-C; 3, Glu-C; 4, Chymotrypsin; 5, Elastase; 6, Proteinase K; 7, Pepsin.

Sequencing alternative spliced forms

splice form	# spectra	# peptide sequences	% recovery (using tags of length 5)
IRAK2A	1478	14444	79% (491/622)
IRAK2B	1589	15584	80% (462/574)
IRAK2D	1731	16835	83% (460/551)

- Shotgun sequencing: The splice forms are differentiated (Each splice form is uniquely characterized by a specific recovered (in red) sequence)
- Classical identification method (Trypsin alone): The splice forms cannot be set apart

Significance

- Shotgun sequencing is as fast as protein identification but is more accurate and increases sequence coverage
- Can detect alternative splice forms when identification method cannot
- First product of its kind
- Biodefense applications
  Sequencing spliced variants and complexes in immune signaling pathways (MISL Grand Challenge) Sequencing Major histocompatibility complex peptides
- Bioenergy applications
  Sequencing method to be used for wrongly or incompletely identified proteins and complexes in microbes involved in biofuels production

Acknowledgement: LDRD program (new directions)

Publications:

- S. P. Gaucher, J. A. Morrow, and J.-L. Faulon. *"Use of a Designed Peptide Array to Infer Dissociation Trends for Non-Tryptic Peptides in Quadrupole Ion Trap and Quadrupole Time of Flight Mass Spectrometry"* **Analytical Chemistry, in press**
- J.-L. Faulon. March 8-9, 2007, New Mexico Bioinformatics Symposium, Santa Fe, NM. **Invited Talk** *"Shotgun Protein Sequencing and Analysis"*
- S. P. Gaucher. April 20, 2007, Cal State Chico Department of Chemistry Seminar Series, Chico, CA **Invited Talk:** *" Peptide and Protein Characterization by Mass Spectrometry: Applications to the Desulfovibrio vulgaris Hildenborough Proteome"*

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2007-2007-5985C CreativeArts 98429.LW.8.07

National Nuclear Security Administration

Sandia National Laboratories