BIOINFORMATICS
2002

Proteomics - what genes give rise to and does anyone have a spare terabyte?

Stephen Barnes
sbarnes@uab.edu

Outline of class

- **What are the best websites for proteomics?**
- **What can I do after I've identified the protein?**
- **Help! Too many proteins**
- **Computing/ bioinformatics/computing issues**

Post-1995 revolution in protein sequencing

- **Due to two major factors**
  - ***The cataloging of many genomes***
  - ***The development of protein/peptide mass spectrometry***
  - ***Speed and sensitivity***

Why is genomic information valuable to protein research?

- **Computer analysis of genomic sequences allows for the detection of individual genes and the regions of their ORFs**
- **From these, the amino acid sequence of individual proteins can be deduced even if a protein has never been isolated or identified**
- **The peptides resulting from protease-induced cleavage of a protein can be deduced**

From Proteins to Sequence Tags

- If each protein (average 500 residues) had a cleavage site every 10 residues, then about 1.5 million peptides describe the expressed products of the human genome
- Each peptide has a molecular weight value that is its individual sequence tag
- Any modification will increase the peptide's molecular weight

Peptide information needed for protein identification

- Peptide-mass fingerprinting and the ideal covering set for protein characterization. M. Wise et al. *Electrophoresis* 18:1399-1409, 1997
- Purpose: To determine the efficiency and nature of protein identification by the use of endoproteinases and mass spectrometry to create and identify the resulting peptides

Setup

Database of 128,719 non-redundant protein entries

Assumptions:

1. **Digestion is always perfect (value of being *in silico*)**
2. **Cleavage always occurs on the carboxy terminal of each amino acid**
3. **Fragment masses were accurate to the nearest dalton, i.e., ± 0.5 Da**

Theoretical proteolysis of derived protein database

***In silico* endoproteinases**

- **All possible single amino acid sites**
- **Biochemical endoproteinases, chymotrypsin, trypsin and Glu-C**

Results for chymotrypsin

Database entries:	128,719
# of peptide fragments:	3,086,608
# of distinct fragments	14,778
Size of largest fragment:	243,718 Da
Max # of entries for a particular fragment	20,926 (260 Da)
Average # entries for a given fragment:	209
Average number of fragments for an entry:	24
# of uncut entries:	3,059
Average size of uncut entries:	3,194 Da
Max size of uncut entry:	65,243 Da

# of entries defined by X fragments

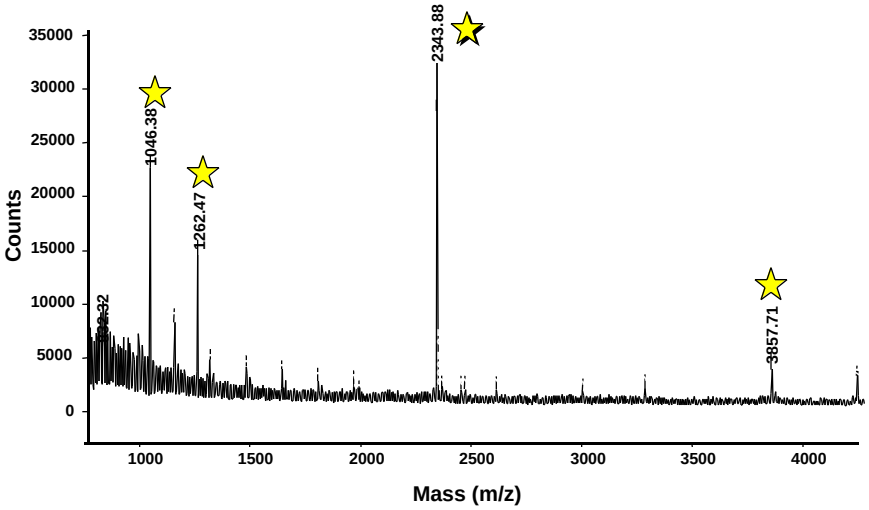
X=1:	2,900
X=2:	88,118
X=3:	26,369
X=4:	952
X=5:	48
X=6:	13
X=7:	2
X=8:	1
X=9:	1

Average # of fragments to define a protein:2.216

Summary of digestion data

Amino acid	Distinct Fragments	Avg # fragments	#Uncut	Avg ident
A alanine	15,372	21.45	3,468	2.13
C cysteine	38,661	6.40	21,525	1.91
D aspartate	17,163	16.15	6,936	2.05
E glutamate	16,960	18.43	6,555	2.08
F phenylalanine	21,642	12.92	7,788	2.00
G glycine	16,490	20.42	3,531	2.13
H histidine	28,695	7.72	18,104	1.96
I isoleucine	18,227	17.36	6,735	2.08
K lysine	19,821	17.50	6,673	2.07
L leucine	12,490	26.19	3,598	2.23
M methionine	29,873	7.88	14,409	1.95
N asparagine	19,765	14.41	8,077	2.03
P proline	19,437	15.34	6,590	2.04
N glutamine	20,182	12.84	8,062	2.01
R arginine	18,754	16.07	6,633	2.07
S serine	13,829	21.51	3,446	2.15
T threonine	15,455	18.21	4,451	2.11
V valine	15,089	19.61	5,084	2.11
W tryptophan	39,643	5.09	26,214	1.91
Y tyrosine	24,343	10.79	9,738	1.98
Glu-C	11,291	30.88	2,808	2.28
Chymotrypsin	14,780	25.42	2,822	2.22
Trypsin	10,846	30.37	2,418	2.34

MALDI-TOF mass spectrum of tryptic digest of p22 band purified by 6xHis-tag



Searching databases with peptide masses to identify proteins

p22: 1046.38 1262.47 2343.88 3857.71

Best site is at www.matrixscience.com

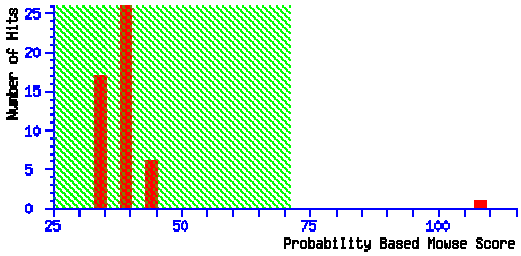
The program (MASCOT) can search the OWL or NCBI databases using a set of tryptic peptide masses, or the fragment ions (specified or unspecified) of peptides

Presents the expected set of tryptic peptides for each matched protein

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.

Protein scores greater than 71 are significant (p<0.05).



Accession	Mass	Score	Description
1. gi|548939	20840	108	FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE SLYD (PPIASE) (ROTAMA
2. gi|13384624	46931	45	myocyte enhancer factor 2C [Mus musculus]
3. gi|5257384	43424	44	(AF137308) phytochrome B [Lolium perenne]
4. gi|4505147	50305	44	MADS box transcription enhancer factor 2, polypeptide C (myocyte enhan
5. gi|1515365	44552	43	(U52596) nucleocapsid protein [Avian infectious bronchitis virus]
6. gi|6093850	49443	42	PRESENILIN 2 (PS-2)
7. gi|15225198	47999	42	hypothetical protein [Arabidopsis thaliana]
8. gi|113854	58376	41	NITROGENASE IRON-IRON PROTEIN ALPHA CHAIN (NITROGENASE COMPONENT I)
9. gi|13928425	13831	40	(AB040419) envelope protein [Bovine immunodeficiency virus]
10. gi|4389228	56064	40	Chain Z, Crystal Structure Of The Complex Between Escherichia Coli Glycerol

MASCOT SEARCH SUMMARY

1. gi|548939 Mass: 20840 Score: 108
FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE SLYD (PPIASE) (ROTAMA

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.59	-0.22	132 -	140	0	FNVEVVAIR
1262.47	1261.46	1261.70	-0.24	6 -	16	0	DLVVSLAYQVR
2343.88	2342.87	2343.08	-0.20	58 -	78	0	FDVAVGANDAYGQYDENLVQR
3857.71	3856.70	3856.89	-0.19	96 -	131	0	FLAETDQGPVPVEITAVEDDHVVVDGNHMLAGQNLK

2. gi|13384624 Mass: 46931 Score: 45
myocyte enhancer factor 2C [Mus musculus]

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.50	-0.13	263 -	271	0	NTMPSVNQR
3857.71	3856.70	3856.76	-0.06	178 -	218	0	NSMSPGVTHRPPSAGNTGGLMGGDLTSGAGTSAGNGYGNPR

No match to: 1262.47, 2343.88

3. gi|5257384 Mass: 43424 Score: 44
(AF137308) phytochrome B [Lolium perenne]

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.54	-0.17	380 -	389	0	GIDELSSVAR
3857.71	3856.70	3856.72	-0.02	86 -	122	0	SPHGCHAQYMANMGSIASLVMAVIISSGGEDEHNMGR

No match to: 1262.47, 2343.88

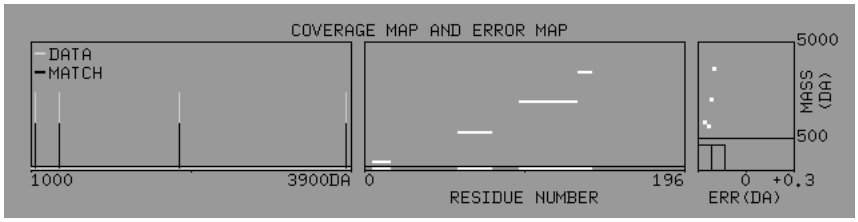
4. gi|4505147 Mass: 50305 Score: 44
MADS box transcription enhancer factor 2, polypeptide C (myocyte enhan

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.50	-0.13	265 -	273	0	NTMPSVNQR
3857.71	3856.70	3856.76	-0.06	180 -	220	0	NSMSPGVTHRPPSAGNTGGLMGGDLTSGAGTSAGNGYGNPR

No match to: 1262.47, 2343.88

Other web sites for peptide analysis

- **http://prowl.rockefeller.edu/**
  - **Choose ProFound**
- **http://prospector.ucsf.edu/**
  - **Choose MS-fit**



Details for rank 1 candidate in search C0B50591-03B8-23FB31C2
gi|548939|sp|P30856|**SLYD_ECOLI FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE SLYD (PPIASE) (ROTAMASE) (HISTIDINE RICH PROTEIN) (WHP)**
gi|1073559|pir||A49987 probable fkbP-type peptidyl-prolyl cis-trans isomerase slyD - Escherichia coli
gi|394720|emb|CAA79705.1| (Z21496) histidine rich protein [Escherichia coli]
gi|475995|gb|AAA18574.1| (L13261) sensitivity to lysis gene [Escherichia coli]
gi|606283|gb|AAA58146.1| (U18997) histidine rich protein [Escherichia coli]
gi|862299|gb|AAC41458.1| (L28082) slyD gene product [Escherichia coli]
gi|1789748|gb|AAC76374.1| (AE000411) FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase) [Escherichia coli K12]
gi|12517970|gb|AAG58456.1|AE005558_9 (AE005558) FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase) [E coli
gi|13363673|dbj|BAB37623.1| (AP002564) FKBP-type peptidyl-prolyl cis-trans isomerase [Escherichia coli O157:H7]
Sample ID : [Pass:0]
Measured peptides : 4
Matched peptides : 4
Min. sequence coverage: 39%

Measured Mass(M)	Avg/ Mono	Computed Mass	Error (Da)	Residues Start	To	Peptide sequence
1045.372	M	1045.592	-0.219	132	140	FNVEVVAIR
1261.462	M	1261.703	-0.240	6	16	DLVVSLAYQVR
2342.872	M	2343.076	-0.204	58	78	FDVAVGANDAYGQYDENLVQR
3856.702	M	3856.892	-0.190	96	131	FLAETDQGPVPVEITAVEDDHVVVDGNHMLAGQNLK

E. coli: FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE

Nominal mass of protein (Mr): 20840

```
  1 MKVAKDLVVS LAYQVRTEDG VLVDESPVSA PLDYLHGHGS
 41 LISGLETALE GHEVGDKFDV AVGANDAYGQ YDENLVQRVP
 81 KDVFMGVDEL QVGMRFLAET DQGPVPVEIT AVEDDHVVVD
121 GNHMLAGQNL KFNVEVVAIR EATEEELAHG HVHGAHDHHH
161 DHDHDGCCGG HGHDHGHEHG GEGCCGGKGN GGCGCH
```

Tryptic fragments detected by MALDI-TOF-MS

```
132-140 FNVEVVAIR
  6- 16 DLVVSLAYQVR
 58- 78 FDVAVGANDAYGQYDENLVQR
 96-131 FLAETDQGPVPVEITAVEDDHVVVDGNHMLAGQNLK
```

Linking protein data to other databases

- **Mascot provides:**
  - ***Sequences of the individual peptides for a given protein***
  - ***Molecular weights (observed and expected)***
  - ***The full list of peptides for a given protein***
  - ***The opportunity to carry out BLAST or psi-BLAST searches of the genome databases to find homologous proteins***

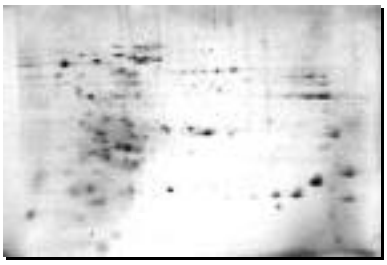
What is BLAST?

- **Part of a set of programs available through http://www.ncbi.nlm.nih.gov**
- **It allows discovery of proteins with sequence alignment similarities to the protein or peptide sequence recovered from MS experiments**
- **Available as BLASTp (single comparisons) or psi-BLAST (iterative)**

Alternative searching method for protein similarities at sequence and structural levels

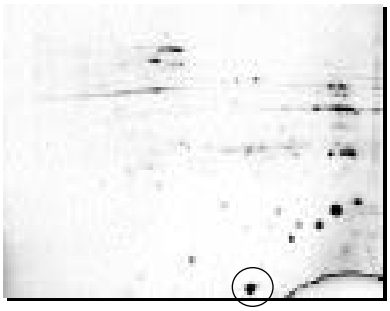
- **Once protein has been identified, go to Entrez at ncbi.nlm.nih.gov and enter the protein name**
- **Once the correct protein is highlighted, click on *blink* - this does a BLAST-style similarity search and identifies related proteins whose structure is known**

Human leukemic cell line (HL-60)



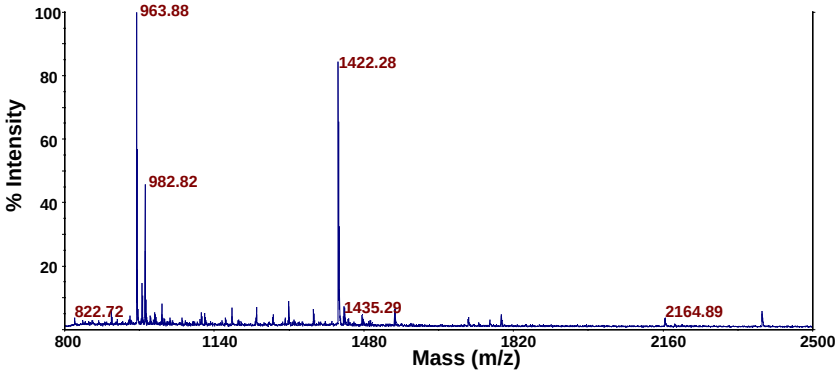
DMSO-induced differentiation

Differentiated neutrophil-like cells



New spot

MALDI-TOF analysis of trypsinized spot from 2D-IEF/SDS-PAGE analysis of DMSO-differentiated HL-60 cells



BLAST analysis

p12: 963.88 982.82 1422.28 1435.29

- **DMSO-induced spot was shown to be S-100 (or calgranulin A) - a calcium binding protein**
- **BLAST showed that the sequence of S-100 is shared by migratory inhibitory factor related protein 8**
- **Two of these entries have a pdb entry (4-letter alphanumeric descriptor) - this means there is a molecular structure available**

Visualization of protein structure

- **A useful website**
  - **http://scop.mrc-lmb.cam.ac.uk/scop/ (structural classification of proteins)**
  - **Necessary to first download the plug-in Chime**
  - **Enter the 4-letter alphanumeric**

Where do we go now?

- **Use of Clustal analysis to localize which are the crucial residues**
  - **Load sequence data of related proteins at *http://www.cmbi.kun.nl/cgi-bin/clustalw.pl***
- **To determine the members of protein networks**
  - **A place to start is at BIND (http://www.bind.ca/)**

Clustal analysis of BATs and thioesterase

```
Kan-1           -MAKLTAVPLS-ALVDEPVHIRVTGLTPFQVVCLQASLKDDKGNLFNSQAFYRASEVGEV
mBAT            -MAKLTAVPLS-ALVDEPVHIQVTGLAPFQVVCLQASLKDER-KPVSSQAFYRASEVGEV
hBAT            -MIQLTATPVS-ALVDEPVHIRATGLIPFQMVSFQASLEDENGDMFYSQAHYRANEFGEV
PTE-2           MAATLILEPAGRCCWDEPVRIAVRGLAPEQPVTLRASLRDEKGALFQAHARYRADTLGEL
                    *   * . .  ****:* . ** * * * ::***.*:.   . ::* ***. .**:

Kan-1           DLERDSSLGGDYMGVHPMGLFWSMKPEKLLTRLVKRDVMNRPHKVHIKLCHPYFPVEGKV
mBAT            DLEHDPSLGGDYMGVHPMGLFWSLKPEKLLGRLIKRDVINSPYQIHIKACHPYFPLQDLV
hBAT            DLNHASSLGGDYMGVHPMGLFWSLKPEKLLTRLLKRDVMNRPFQVQVKLYDLELIVNNKV
PTE-2           DLERAPALGGSFAGLEPMGLLWALEPEKPLVRLVKRDVR-TPLAVELEVLDGHDPDPGRL
                **:: .:***.: *:.****:*:::*** * **:****   *  :.::  .      . :

Kan-1           ISSSLDSLILERWYMAPGVTRIHVKEGRIRGALFLPPGEGPFPGVIDLFGGAGGLFEFRA
mBAT            VSPPLDSLTLERWYVAPGVKRIQVKESRIRGALFLPPGEGPFPGVIDLFGGAGGLMEFRA
hBAT            ASAPKASLTLERWYVAPGVTRIKVREGRLRGALFLPPGEGLFPGVIDLFGGLGGLLEFRA
PTE-2           LCQTRH----ERYFLPPGVRREPVRVGRVRGTLFLPPEPGPFPGIVDMFGTGGGLLEYRA
                 . .      **:::.*** *  *: .*:**:*****  * ***::*:**  ***:*:**

Kan-1           SLLASHGFATLALAYWGYDDLPSRLEKVDLEYFEEGVEFLLRHPKVLGPGVGILSVCIGA
mBAT            SLLASRGFATLALAYWNYDDLPSRLEKVDLEYFEEGVEFLLRHPKVLGPGVGILSVCIGA
hBAT            SLLASRGFASLALAYHNYEDLPRKPEVTDLEYFEEAANFLLRHPKVFGSGVGVVSVCQGV
PTE-2           SLLAGKGFAVMALAYYNYEDLPKTMETLHLEYFEEAMNYLLSHPEVKGPGVGLLGISKGG
                ****.:*** .**** .*:***   *  .******. ::** **:* *.***::.:. *

Kan-1           EIGLSMAINLKQITATVLINGPNFVSSNPHVYRGKVFQPTPCSEEFVTTNALGLVEFYRT
mBAT            EIGLSMAINLKQIRATVLINGPNFVSQSPHVYHGQVYPPVPSNEEFVVTNALGLVEFYRT
hBAT            QIGLSMAIYLKQVTATVLINGTNFPFGIPQVYHGQIHQPLPHSAQLISTNALGLLELYRT
PTE-2           ELCLSMASFLKGITAAVVINGSVANVGGTLRYKGETLPPVGVNRNRIKVTKDGYADIVDV
                :: ****  ** : *:*:***.      .  *:*:   *   . : : ..  *  ::  .
```

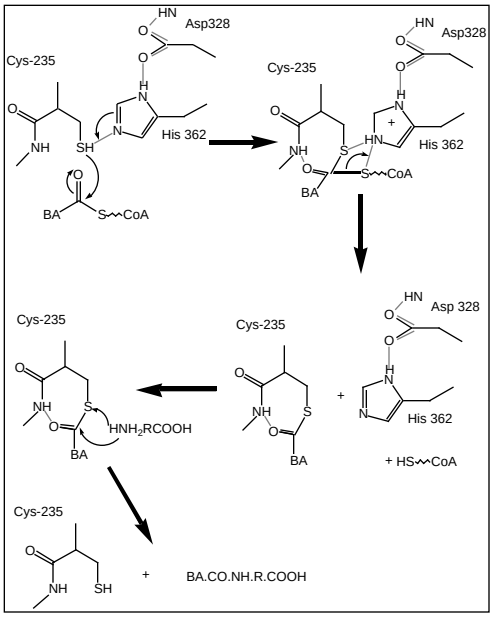
Clustal analysis of BATs and thioesterase

```
Kan-1           FEETADK-DSKYCFPIEKAHGHFLFVVGED DKNLNSKVHAKQAIAQLMKSGKKNWTLLSY
mBAT            FQETADK-DSKYCFPIEKAHGHFLFVVGED DKNLNSKVHANQAIAQLMKNGKKNWTLLSY
hBAT            FETTQVG-ASQYLFPIEEAQGQFLFIVGEG DKTINSKAHAEQAIGQLKRHGKNNWTLLSY
PTE-2           LNSPLEGPDQKSFIPVERAESTFLFLVGQD DHNWKSEFYANEACKRLQAHGRRKPQIICY
                :: .     .:  :*:*.*.. ***:**:.*:. :*: :*::*  :*   *:.:  ::.*

Kan-1           PGAGHLIEPPYSPLCSASRMPFVIPSINWGGEVIPH-AAAQEHSWKEIQKFLKQHLNP--
mBAT            PGAGHLIEPPYTPLCQASRMPILIPSLSWGGEVIPHSQAAQEHSWKEIQKFLKQHLLP--
hBAT            PGAGHLIEPPYSPLCCASTTHDLR--LHWGGEVIPH-AAAQEHAWKEIQRFLRKHLIP--
PTE-2           PETGHYIEPPYFPLCRASLHALVGSPIIWGGEPRAH-AMAQVDAWKQLQTFFHKHLGGRE
                * :** ***** *** **    :   : ****  .*   ** .:**::* *:::**

Kan-1           -GFNSQL
mBAT            -DLSSQL
hBAT            -DVTSQL
PTE-2           GTIPSKV
                  . *::
```

***Current interpretation*: the critical residues that govern the reaction of bile acid CoA with a conjugating enzyme or a thioesterase are Cys 235, Asp 328 and His 362**



Mechanism of action of hBAT derived from sequence and Clustal analysis

In thioesterases, Cys 235 is replaced by a Ser residue - this produces a more unstable intermediate that decomposes before attack by the amino acid second substrate

How can diversity be accomplished with a limited number of genes?

- One can consider that interactions between gene products (proteins) are not linear
  - K *(biological complexity)* = *f*(N)
  - Proportional - K = αN (unlikely)
  - Polynomial - K = αN^u
  - Exponential - K = α^N
  - Factorial - K = N!
- If proteins have two states (ON/OFF), then there are $2^{30,000}$ possible combinations. A human in this model has $2^{30,000}/2^{20,000}$ more combinations than a neomatode (approx 10^{300}).

Geneticists forget about protein sequence and structure

- Each gene product (the protein) has regions that are critical to the function of the protein (*intrinsic activity* – its enzyme catalysis potential), its modifications, and its ability to form protein complexes.
- Although the number of combinations that are theoretically possible for forming protein complexes seems utterly enormous, this set is very much less than that since the sequence and the folding of the protein is crucial for proper protein-protein interaction, thereby limiting the degrees of freedom.

Statistical issues in proteomics

- ***Measurement quality*** **– is it reproducible?**
- ***Experimental design*** **– how can variance be measured?**
- ***Classification*** **– are proteins acting in concert?**
- ***Inference*** **– what is the likelihood that the expression of a protein is different from a "control" situation? What is the likelihood of a false positive or negative?**
- ***Estimation*** **– what is the best estimate of an effect and what are the 95% confidence limits?**

What information is contained within a proteomics data set?

- **The total data can be represented by the following equation:**

$$X = t_1 p_1^T + t_2 p_2^T + \ldots t_k p_k^T + E$$

- **The sum $\Sigma t_n p_n^T$ is the amount of variation that can be accounted for by combination of factors (t_n) and their magnitudes (p_n).**
- **To calculate a vector containing the weights one must invert the observation matrix. This is a major computational task and only works if the number of sample sets is > than the number of variables**

What are the computing challenges in proteomics and mass spectrometry?

- To effectively store large data sets
- To automatically analyze a large data set and distribute the information to users via the web
- To build real time data acquisition and analysis so that informed decisions can be made on-the-fly

Peptide analysis by mass spec and protein identification in real time

- The complement of tryptic peptides for a given protein are correlated
- Once we know the masses and partial sequence of two peptides, the remaining peptides from that protein that elute from the HPLC column can be predicted and thereby excluded
- Greater effort can be placed on examining the non-predicted peptides which may include those with posttranslational modification