Getting Academic Software To Scientists

Brian Searle and Mark Turner

Proteome Software Inc.

Portland, Oregon USA

Brian.Searle@ProteomeSoftware.com

Commercial Search Engines

- SEQUEST
- Mascot
- Spectrum Mill
- Phenyx/OLAV
- ProteinLynx/IdentityE
- Paragon
- Sonar

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Academic Search Engines

- X! Tandem
- OMSSA
- Protein Prospector
- Pep_Probe
- Myrimatch
- SCOPE
- ProbId
- MassMatrix
- DBDigger/MASPIC
- RMET-Luck
- VEMS
- PepSearch
- pepReap
- PepSlice
- PepSOM

- Anal Chem 2003, 75:435-444.
- Bioinformatics, 2004; 20: 1948 - 1954.
- Bioinformatics, 2007; 23: 612 - 618.

PROTEOME SOFTWARE
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Academic Search Engines

- GutenTag
- MultiTag
- InsPecT
- CIDentify
- OpenSea
- Spider
- MS-BLAST
- FASTF/FASTS
- Popitam
- MS-Convolution
- MS-Alignment
- pFind

- Lutefisk
- Sherenga
- Pepnovo
- PILOT
- MSNovo
- Sequit

- DACSIM
- MS2Grouper
- Bonanza
- X! Hunter
- Bibliospec
- SpectraST

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

- J Comput Biol. 2000, 7(6):777-87.
- J Comput Biol. 2001, 8(6): 571-583.
- J Comput Biol. 2001, 8(3): 325-337.
- Anal Chem. 2003, 75(10):2470-7.
- Proteomics 2003, 3(8):1597-610.
- J Comput Biol. 2003, 10(1):1-12.
- Proteomics 2004, 4(4):950-60.
- Bioinformatics 2004, 20: 2296 - 2304.
- J Comput Biol. 2007, 14(8): 1025-1043.

Hits for "Tandem Mass Spectrometry" & "Bioinformatics"

- Analytical Chemistry → 15
- J Proteome Res → 51
- Proteomics → 58
- Mol Cell Proteomics → 45

- Bioinformatics → 66
- J Comp Bio → 12
- BMC Bioinformatics → 29
- J Bioinformatics & Comp Bio → 4

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Academic Software is Awesome

- Huge diversity
- Cutting edge innovation
- Highly adaptable

- Publications – no “black boxes”

Academic Software is Awkward

- Full of bugs
- Difficult to use
- Other people in the lab hate it
- Almost never makes it out of the lab

- It dies after the researcher leaves

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Contents

1. Difficulties Levering Intellectual Property
2. Getting Over the Hurdles of Academic Software
3. Funding Software Development

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Non-Commercial Search Engines

- GutenTag
- MultiTag
- InsPecT
- CIDentify
- OpenSea
- Spider
- MS-BLAST
- FASTF/FASTS
- Popitam
- MS-Convolution
- MS-Alignment
- pFind

- Lutefisk
- Sherenga
- Pepnovo
- PILOT
- MSNovo
- Sequit

- DACSIM
- MS2Grouper
- Bonanza
- X! Hunter
- Bibliospec
- SpectraST

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

- J Comput Biol. 2000, 7(6):777-87.
- J Comput Biol. 2001, 8(6): 571-583.
- J Comput Biol. 2001, 8(3): 325-337.
- Anal Chem. 2003, 75(10):2470-7.
- Proteomics 2003, 3(8):1597-610.
- J Comput Biol. 2003, 10(1):1-12.
- Proteomics 2004, 4(4):950-60.
- Bioinformatics 2004, 20: 2296 - 2304.
- J Comput Biol. 2007, 14(8): 1025-1043.

Academic Software is Awkward

- Full of bugs
- Difficult to use
- Other people in the lab hate it
- Almost never makes it out of the lab

- It dies after the researcher leaves

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Academic Software is Awkward

- Full of bugs ✔
- Difficult to use
- Other people in the lab hate it
- Almost never makes it out of the lab

- It dies after the researcher leaves

Academic Software is Awkward

- Full of bugs ✔
- Difficult to use ✔
- Other people in the lab hate it
- Almost never makes it out of the lab

- It dies after the researcher leaves

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Academic Software is Awkward

- Full of bugs ✔
- Difficult to use ✔
- Other people in the lab hate it ✔
- Almost never makes it out of the lab

- It dies after the researcher leaves

Academic Software is Awkward

- Full of bugs ✓
- Difficult to use ✓
- Other people in the lab hate it ✓
- Almost never makes it out of the lab ✓

- It dies after the researcher leaves

Academic Software is Awkward

- Full of bugs ✓
- Difficult to use ✓
- Other people in the lab hate it ✓
- Almost never makes it out of the lab ✓

- It dies after the researcher leaves ?

Lesson #1

…Don’t patent your software

Lesson #2

…If you do,
don't write it yourself

(it won't hold up)

PROTEOME SOFTWARE

\\Snapserver\proteomesoft\users\brian\example_control_sprot.xml - OpenSea(TM) Browser Client

File Edit View Print Help

Identification	Prob	Score
040203_seq_1_peaks_alkylated		
control_sprot.fasta		
Serum Albumin Precursor	100....	87.8
Ovotransferrin	100....	53.8
IGG_CONTR (C00010) NI	100....	29.4
IGGL_CONTR (AAK84156	100....	23.0
Myoglobin	100....	17.7
LACB_CONTR (C00006) I	100....	17.6
Peroxidase C1a	100....	10.5
Superoxide Dismutase	100....	10.1
Cytochrome C	100....	9.1
Ubiquitin	100....	8.5
IGGH_CONTR (AAB3738	97.0...	2.6
Tryp_pig Trypsin	95.0...	1.7
Insulin	95.0...	1.6
Unmatched Spectra		

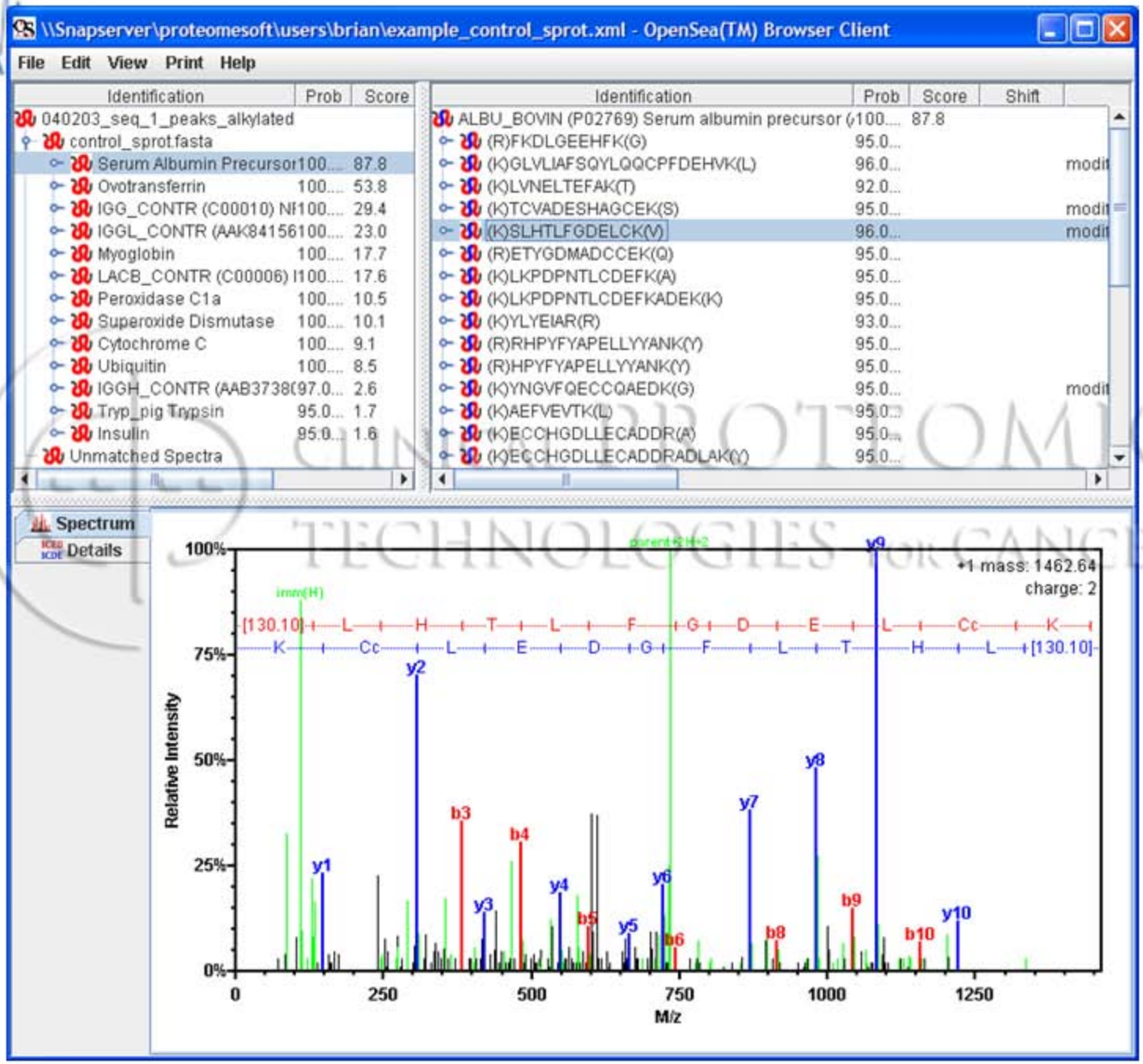
Identification	Prob	Score	Shift
ALBU_BOVIN (P02769) Serum albumin precursor (/	100....	87.8	
(R)FKDLGEEHFK(G)	95.0...		
(K)GLVLIAFSQYLQQCPFDEHVK(L)	96.0...		modif
(K)LVNELTEFAK(T)	92.0...		
(K)TCVADESHAGCEK(S)	95.0...		modif
(K)SLHTLFGDELCK(V)	96.0...		modif
(R)ETYGDMADCCEK(Q)	95.0...		
(K)LKPDPNTLCDEFK(A)	95.0...		
(K)LKPDPNTLCDEFKADEK(K)	95.0...		
(K)YLYEIAR(R)	93.0...		
(R)RHPYFYAPELLYYANK(Y)	95.0...		
(R)HPYFYAPELLYYANK(Y)	95.0...		
(K)YNGVFQECCQAEDK(G)	95.0...		modif
(K)AEFVEVTK(L)	95.0...		
(K)ECCHGDLLECADDR(A)	95.0...		
(K)ECCHGDLLECADDRADLAK(Y)	95.0...		

ALBU_BOVIN (P02769) Serum albumin precursor (Allergen Bos d 6)
length: 607 amino acids, mass: 71272.19 AMU

```
M K W V T F I S L L   L L F S S A Y S R G   V F R R D T H K S E   I A H R F K D L G E   40
E H F K G L V L I A   F S Q Y L Q Q C P F   D E H V K L V N E L   T E F A K T C V A D   80
E S H A G C E K S L   H T L F G D E L C K   V A S L R E T Y G D   M A D C C E K Q E P   120
E R N E C F L S H K   D D S P D L P K L K   P D P N T L C D E F   K A D E K K F W G K   160
Y L Y E I A R R H P   Y F Y A P E L L Y Y   A N K Y N G V F Q E   C C Q A E D K G A C   200
L L P K I E T M R E   K V L A S S A R Q R   L R C A S I Q K F G   E R A L K A W S V A   240
R L S Q K F P K A E   F V E V T K L V T D   L T K V H K E C C H   G D L L E C A D D R   280
A D L A K Y I C D N   Q D T I S S K L K E   C C D K P L L E K S   H C I A E V E K D A   320
I P E N L P P L T A   D F A E D K D V C K   N Y Q E A K D A F L   G S F L Y E Y S R R   360
H P E Y A V S V L L   R L A K E Y E A T L   E E C C A K D D P H   A C Y S T V F D K L   400
K H L V D E P Q N L   I K Q N C D Q F E K   L G E Y G F Q N A L   I V R Y T R K V P Q   440
V S T P T L V E V S   R S L G K V G T R C   C T K P E S E R M P   C T E D Y L S L I L   480
N R L C V L H E K T   P V S E K V T K C C   T E S L V N R R P C   F S A L T P D E T Y   520
V P K A F D E K L F   T F H A D I C T L P   D T E K Q I K K Q T   A L V E L L K H K P   560
K A T E E Q L K T V   M E N F V A F V D K   C C A A D D K E A C   F A V E G P K L V V   600
S T Q T A L A                                                                           607
```

percent coverage: 55.8%, identified region mass: 66252.31 AMU

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER


\\Snapserver\proteomesoft\users\brian\example_control_sprot.xml - OpenSea(TM) Browser Client
File Edit View Print Help
Identification Prob Score
040203_seq_1_peaks_alkylated
control_sprot.fasta
Serum Albumin Precursor 100.... 87.8
Ovotransferrin 100.... 53.8
IGG_CONTR (C00010) NI 100.... 29.4
IGGL_CONTR (AAK84156 100.... 23.0
Myoglobin 100.... 17.7
LACB_CONTR (C00006) I 100.... 17.6
Peroxidase C1a 100.... 10.5
Superoxide Dismutase 100.... 10.1
Cytochrome C 100.... 9.1
Ubiquitin 100.... 8.5
IGGH_CONTR (AAB3738(97.0... 2.6
Tryp_pig Trypsin 95.0... 1.7
Insulin 95.0... 1.6
Unmatched Spectra
Identification Prob Score Shift
ALBU_BOVIN (P02769) Serum albumin precursor (100.... 87.8
(R)FKDLGEEHFK(G) 95.0...
(K)GLVLIAFSQYLQQCPFDEHVK(L) 96.0... modif
(K)LVNELTEFAK(T) 92.0...
(K)TCVADESHAGCEK(S) 95.0... modif
(K)SLHTLFGDELCK(V) 96.0... modif
(R)ETYGDMADCCEK(Q) 95.0...
(K)LKPDPNTLCDEFK(A) 95.0...
(K)LKPDPNTLCDEFKADEK(K) 95.0...
(K)YLYEIAR(R) 93.0...
(R)RHPYFYAPELLYYANK(Y) 95.0...
(R)HPYFYAPELLYYANK(Y) 95.0...
(K)YNGVFQECCQAEDK(G) 95.0... modif
(K)AEFVEVTK(L) 95.0...
(K)ECCHGDLLECADDR(A) 95.0...
(K)ECCHGDLLECADDRADLAK(Y) 95.0...
Spectrum
Details
+1 mass: 1462.64
charge: 2
Relative Intensity
M/z
100%
75%
50%
25%
0%
0
250
500
750
1000
1250
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
b3
b4
b5
b6
b8
b9
b10

Lesson #3

…Don’t forget
“How To Load Data”

Lesson #4

… Your boss doesn't own your IP

PROTEOME SOFTWARE
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

PROTEOME SOFTWARE

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

OOPS

Lesson #5

…Research science will never be a "Big Market"

Lesson #6

…Sometimes your first book never makes it

Seriously,

- Patenting software takes a long time and doesn't necessarily get you anywhere

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

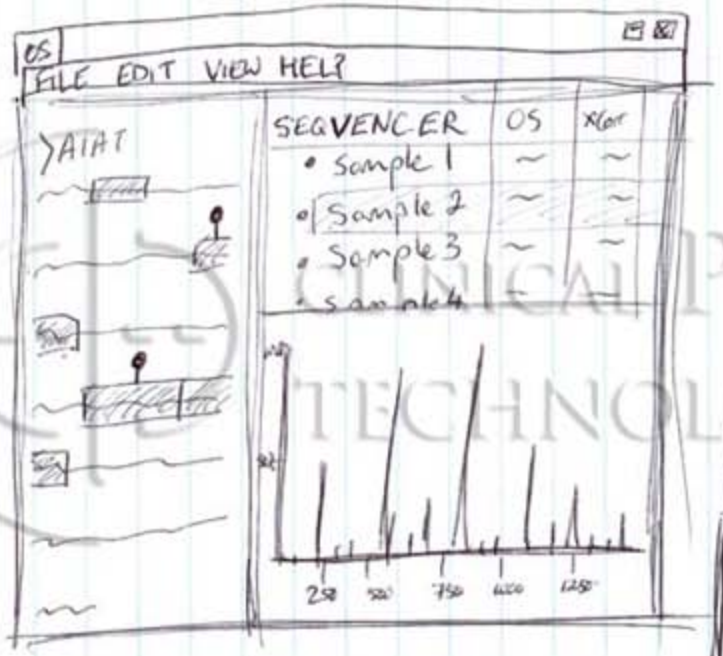
Seriously,

- Patenting software takes a long time and doesn't necessarily get you anywhere
- Your university owns your software, and they might not give it up

Seriously,

- Patenting software takes a long time and doesn't necessarily get you anywhere
- Your university owns your software, and they might not give it up
- Open sourcing your software will guarantee that you can use it later

Results Viewer



What do you want to know?

- peptides / modifications identified across samples?

- out of these samples, what proteins are in all of them?

- of this protein, which samples is it found in?

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Scaffold

Scaffold - trap_control_1.9.30_analyzed

File Edit View Experiment Export Window Help

Min Protein: 95.0% Min # Peptides: 1 Min Peptide: 95%

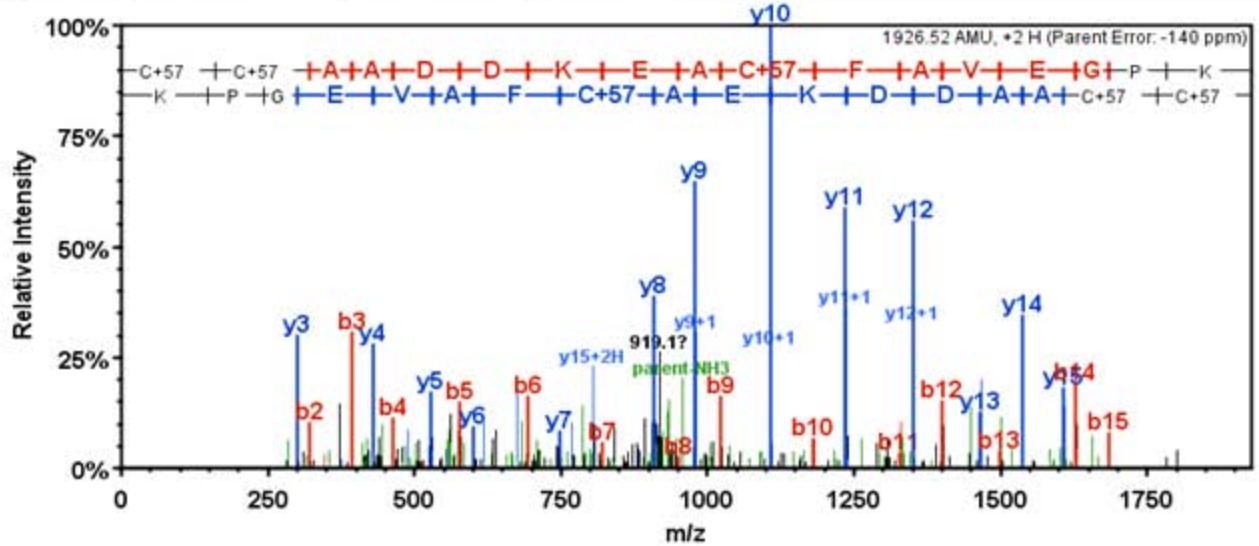
Load Data
Samples
Proteins
Quantify
Publish
Statistics
Peptides

NF00160898 Serum al... All MS/MS Samples

Sequence Coverage	Protein	Category
	NF0016089...	Uncategori

G...	...	Sequence	Prob	S...	SEQ...	Mas...	Mas...	XI T...
✓	✓	(K)LFTFHADICTLPDTEK(Q)	95%	5.84	0.48	52.3	44.9	8.96
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.81	0.59	81.1	43.6	13.17
✓	✓	(K)LFTFHADICTLPDTEK(Q)	95%	5.79	0.42	52.1	44.8	6.24
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.70	0.59	90.6	43.8	9.55
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.66	0.56	80.6	43.2	10.54
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.60	0.51	65.6	43.8	10.23
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.60	0.60	63.4	44.9	9.36
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.57	0.53	76.9	44.2	8.96
✓	✓	(K)LFTFHADICTLPDTEK(Q)	95%	5.52	0.44	53.4	44.6	3.96
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.52	0.49	58.5	43.8	8.77
✓	✓	(K)CCAADDKEACFAVEGPK(L)	95%	5.48	0.51	71.8	44.0	7.54
✓	✓	(K)GLVLIAFSQYLQQCPFDEHVK(L)	95%	5.47	0.49	42.3	42.7	4.70
✓	✓	(K)LFTFHADICTLPDTEK(Q)	95%	5.46	0.42	59.9	44.7	5.34

Protein Sequence | Similar Proteins | Spectrum | Spectrum/Model Error | Fragmentation Table



Lesson #7

…GUIs and "Ease of Use" are critical

Lesson #8

…GUIs take time and "Ease of Use" means doing it wrong first

Over 1000 recorded bugs since June 2005

- 13% in the Algorithms

Over 1000 recorded bugs since June 2005

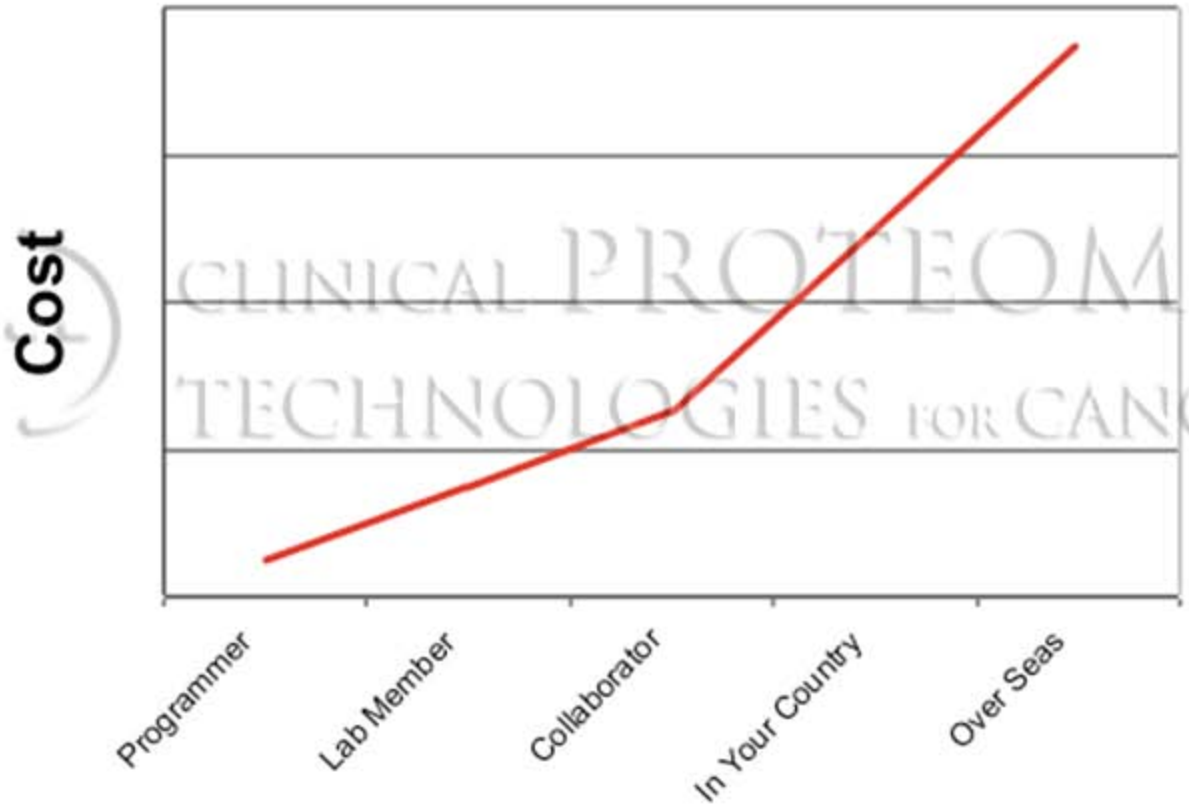
- 13% in the Algorithms
- 23% in File Loading

Over 1000 recorded bugs since June 2005

- 13% in the Algorithms
- 23% in File Loading
- 64% in the Graphical Interface

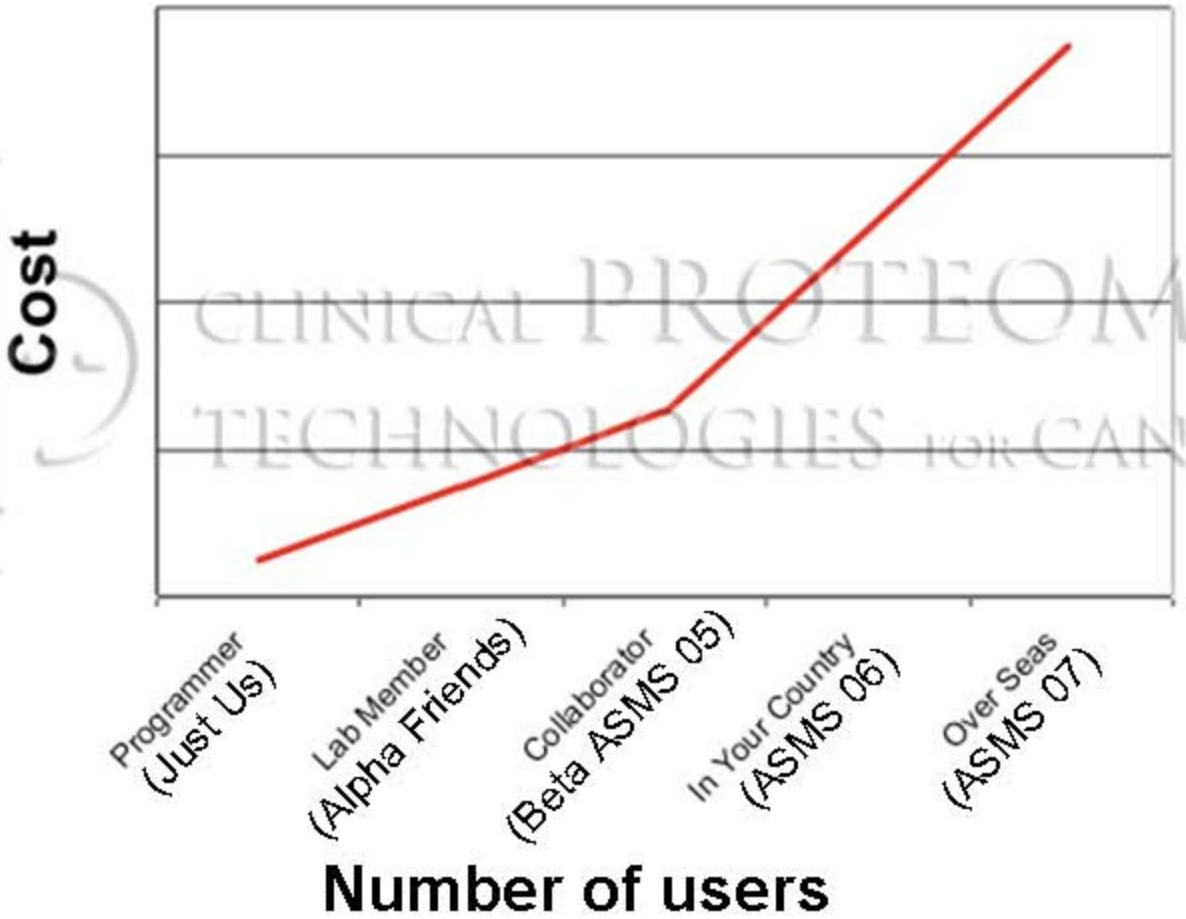
Lesson #9

…Bugs get worse when you distribute widely



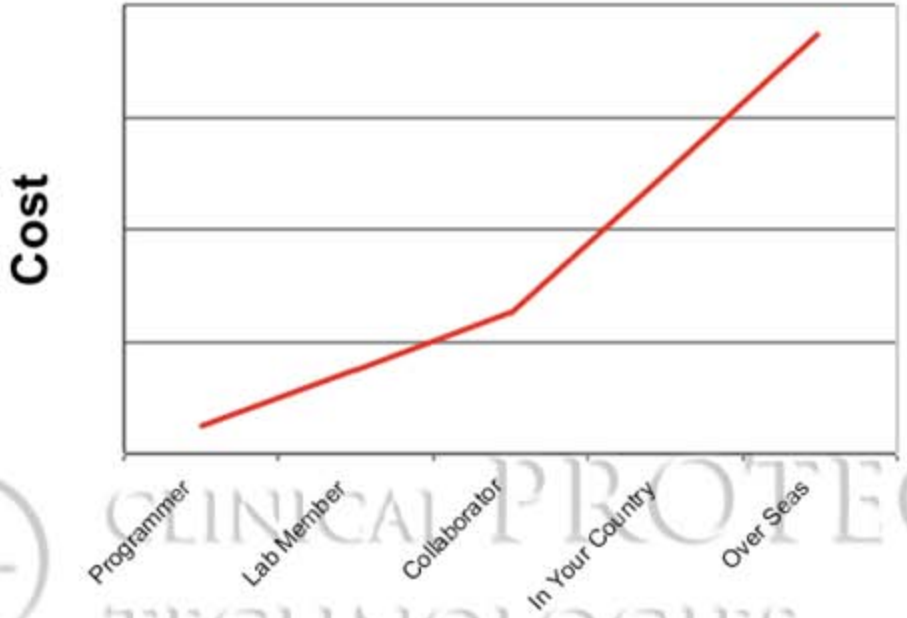
Number of users

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

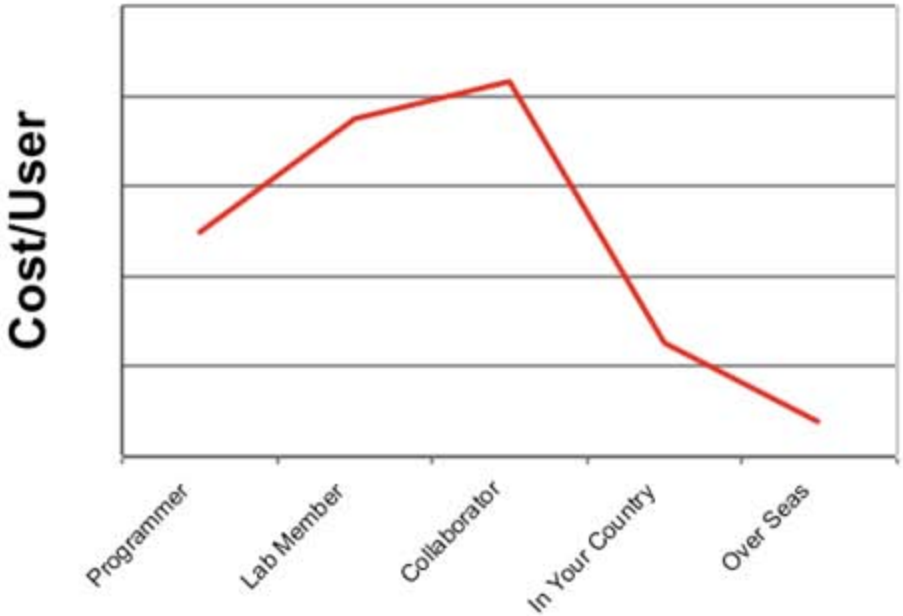


CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Brian, why would we want to do this anyways?



Number of users



PROTEOME SOFTWARE
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Seriously,

- Continued software support is primarily user interface work

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Seriously,

- Continued software support is primarily user interface work
- “Ease of use” means fixing bugs

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Seriously,

- Continued software support is primarily user interface work
- “Ease of use” means fixing bugs
- None of this produces publications

PROTEOME SOFTWARE
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

No publications means:

- PIs interest in software must go beyond the short term

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

No publications means:

- PIs interest in software must go beyond the short term
- We have to hire real software engineers

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

No publications means:

- PIs interest in software must go beyond the short term
- We have to hire real software engineers
- Almost zero funding for maintenance

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

No publications means:

- PIs interest in software must go beyond the short term
- We have to hire real software engineers
- Almost zero funding for maintenance

Funding Opportunity Number: PAR-07-235

How Should Funding Work?

- We don't need R01 caliber money

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How Should Funding Work?

- We don't need R01 caliber money
- Software need SBIR caliber money

PROTEOME SOFTWARE
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How Should Funding Work?

- We don't need R01 caliber money
- Software need SBIR caliber money
- SBIRs (and R01s) aren't good for software with iterative design phases

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How Should Funding Work?

- We don't need R01 caliber money
- Software need SBIR caliber money
- SBIRs (and R01s) aren't good for software with iterative design phases
- We need smaller grants with iterative design in mind

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER