Faster, More Sensitive Peptide ID by Sequence DB Compression

Nathan Edwards

Center for Bioinformatics and Computational Biology

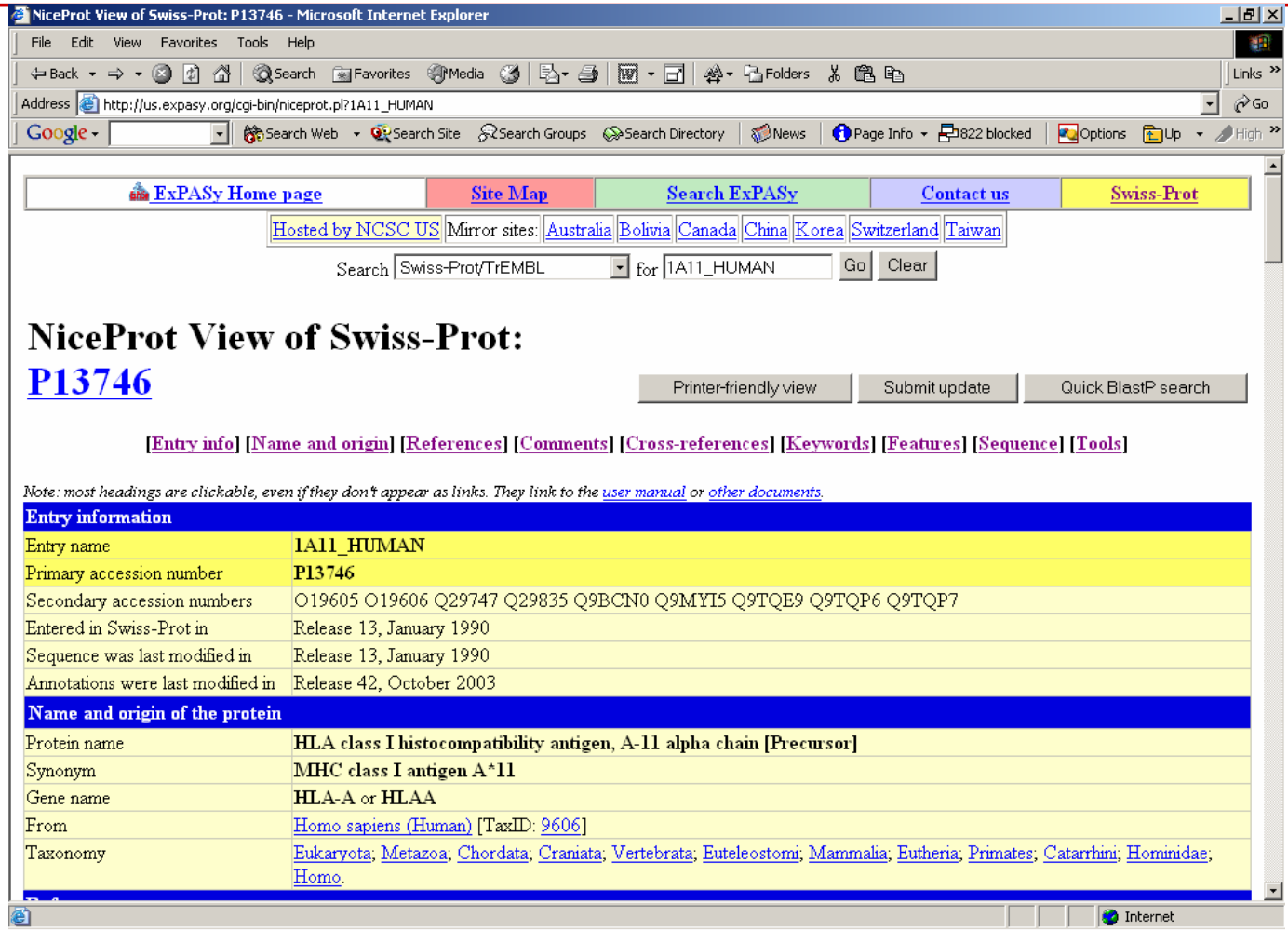
UMIACS
University of Maryland Institute for Advanced Computer Studies

UNIVERSITY OF MARYLAND 18 56

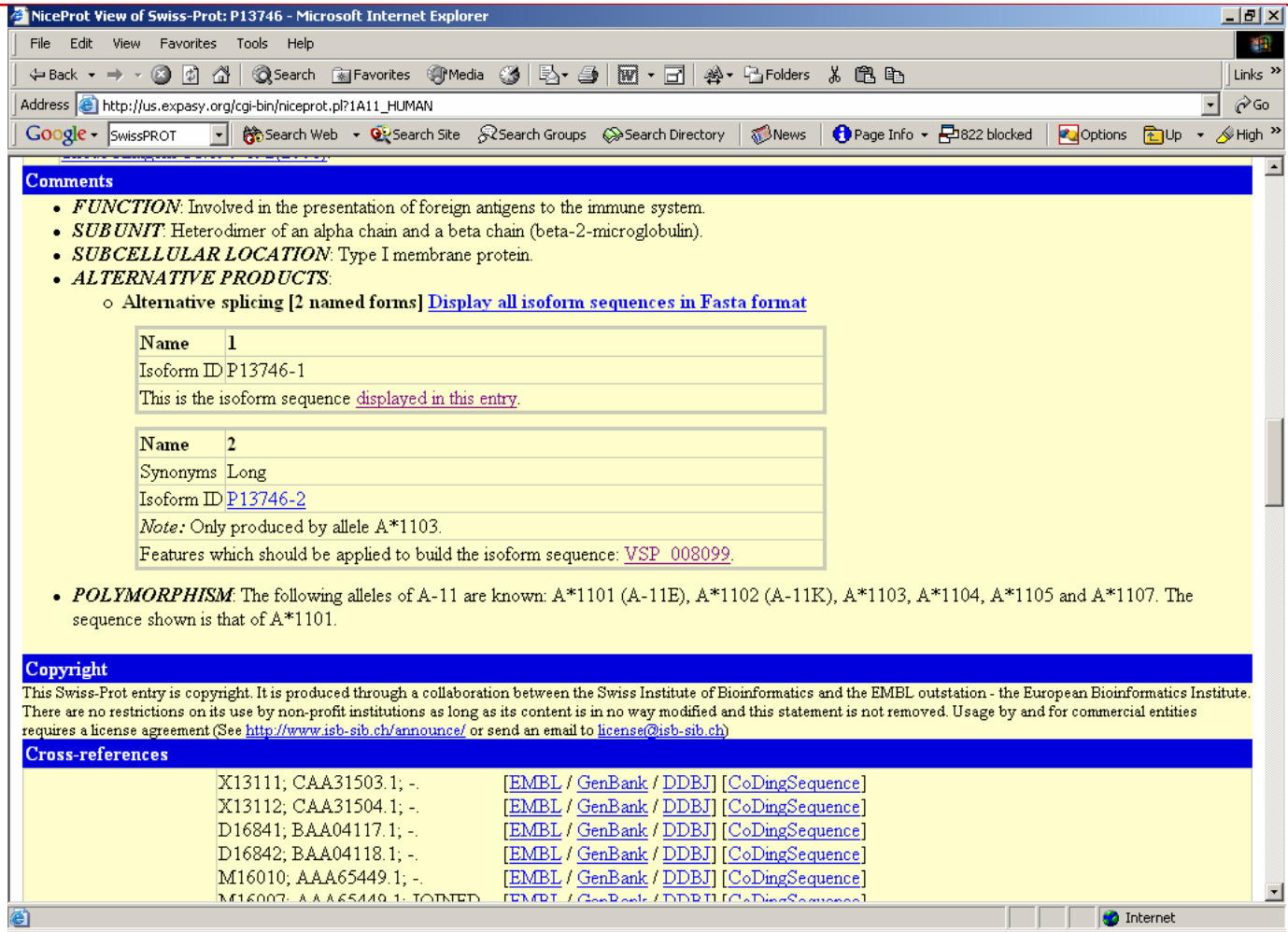
MS/MS Search Engines

- Fail when peptides are missing from sequence database
- Protein sequence databases serve many masters
  - Full length protein sequences not needed for MS/MS
  - Explicit variant enumeration is needed for MS/MS
- Much peptide sequence information is lost, inaccessible, or not integrated
  - Protein isoforms, Sequence variants, SNPs, alternate splice forms, ESTs
- Some peptides are more interesting than others
  - Protein identification is only part of the story

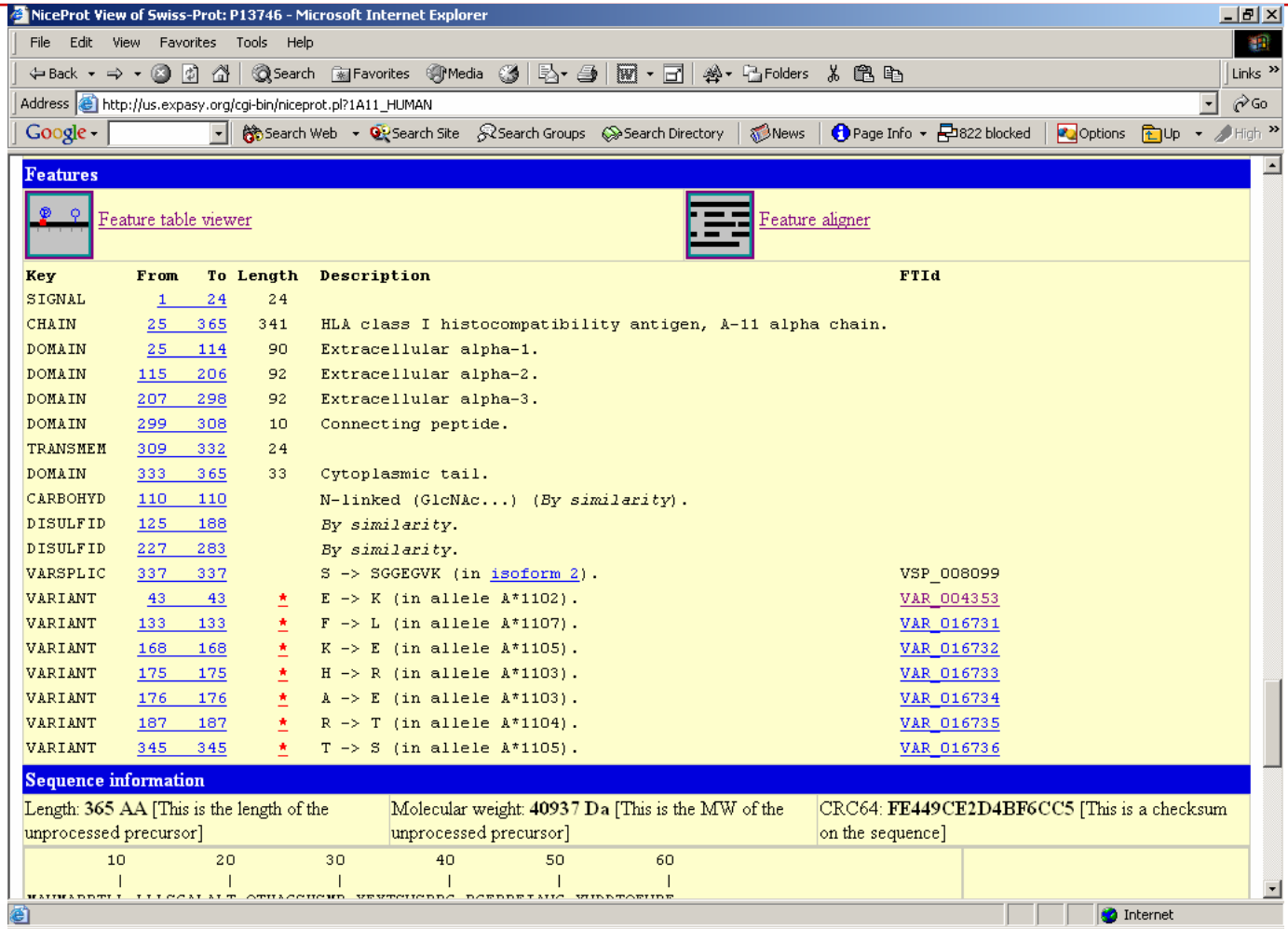
Swiss-Prot


NiceProt View of Swiss-Prot: P13746 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://us.expasy.org/cgi-bin/niceprot.pl?1A11_HUMAN
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Hosted by NCSC US
Mirror sites: Australia Bolivia Canada China Korea Switzerland Taiwan
Search Swiss-Prot/TrEMBL for 1A11_HUMAN Go Clear
NiceProt View of Swiss-Prot:
P13746
Printer-friendly view
Submit update
Quick BlastP search
[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]
Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.
Entry information
Entry name 1A11_HUMAN
Primary accession number P13746
Secondary accession numbers O19605 O19606 Q29747 Q29835 Q9BCN0 Q9MYI5 Q9TQE9 Q9TQP6 Q9TQP7
Entered in Swiss-Prot in Release 13, January 1990
Sequence was last modified in Release 13, January 1990
Annotations were last modified in Release 42, October 2003
Name and origin of the protein
Protein name HLA class I histocompatibility antigen, A-11 alpha chain [Precursor]
Synonym MHC class I antigen A*11
Gene name HLA-A or HLAA
From Homo sapiens (Human) [TaxID: 9606]
Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
Internet

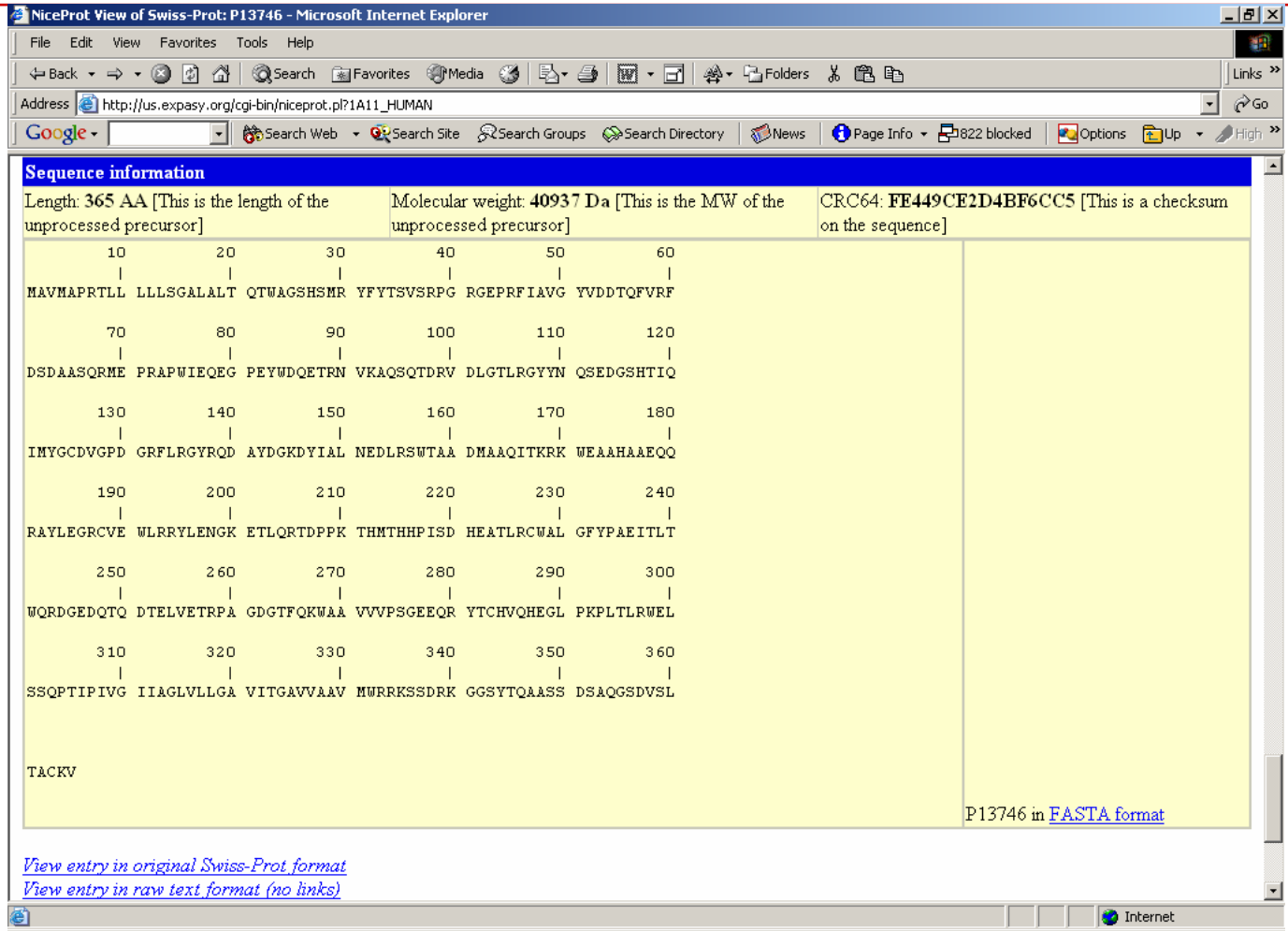
Swiss-Prot Variant Annotations



Swiss-Prot Variant Annotations



Swiss-Prot Sequence

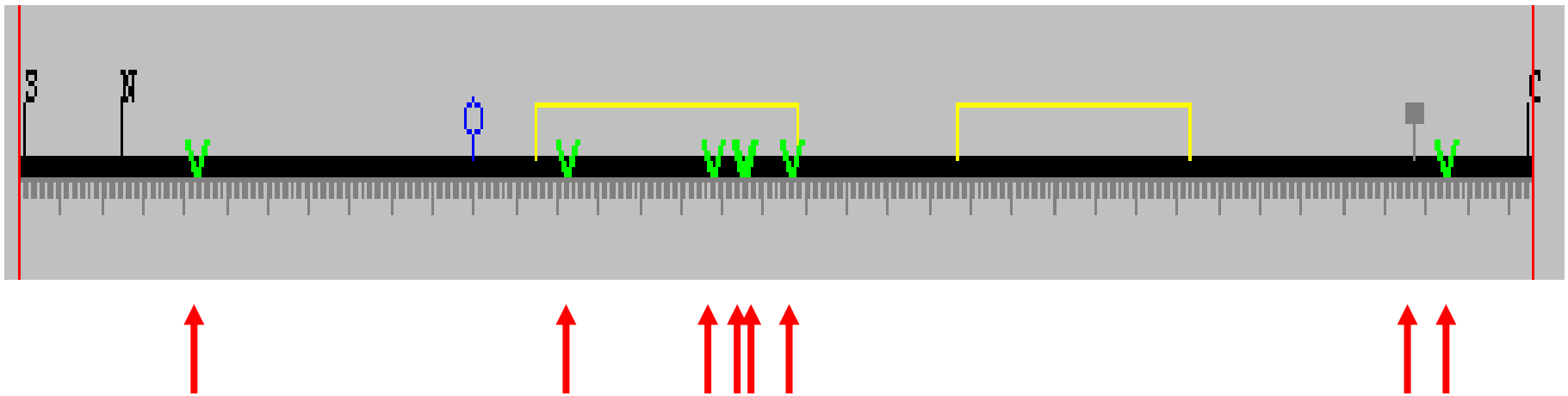


Swiss-Prot

- VarSplic enumerates all variants, conflicts, isoforms
- Swiss-Prot sequence size:
  - 60 Mb
- VarSplic sequence size:
  - 95 Mb
- How many more peptide candidates?

Swiss-Prot
Variant Annotations

Feature viewer



Variants

Swiss-Prot VarSplic Output

```
P13746-00-01-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-01-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-00-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-03-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-03-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-04-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGKPRFIAVGYVDDTQFVRF
P13746-01-04-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGKPRFIAVGYVDDTQFVRF
P13746-00-05-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-05-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-00-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-02-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-02-00     MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
                    ****************************************:*****************
```

Swiss-Prot VarSplic Output

```
P13746-00-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-00-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-01-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYSQAASSDSAQ
P13746-01-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYSQAASSDSAQ
                     ************************************       *******:*********
```

Peptide Candidates

- Parent ion
  - Typically < 3000 Da
- Tryptic Peptides
  - Cut at K or R
- Search engines
  - Don’t handle > 4+ well
  - Long peptides don’t fragment well
- # of distinct 30-mers upper bounds total peptide content

Peptide Candidates

- At most 1% additional peptides in ~ 1.6 times as much sequence

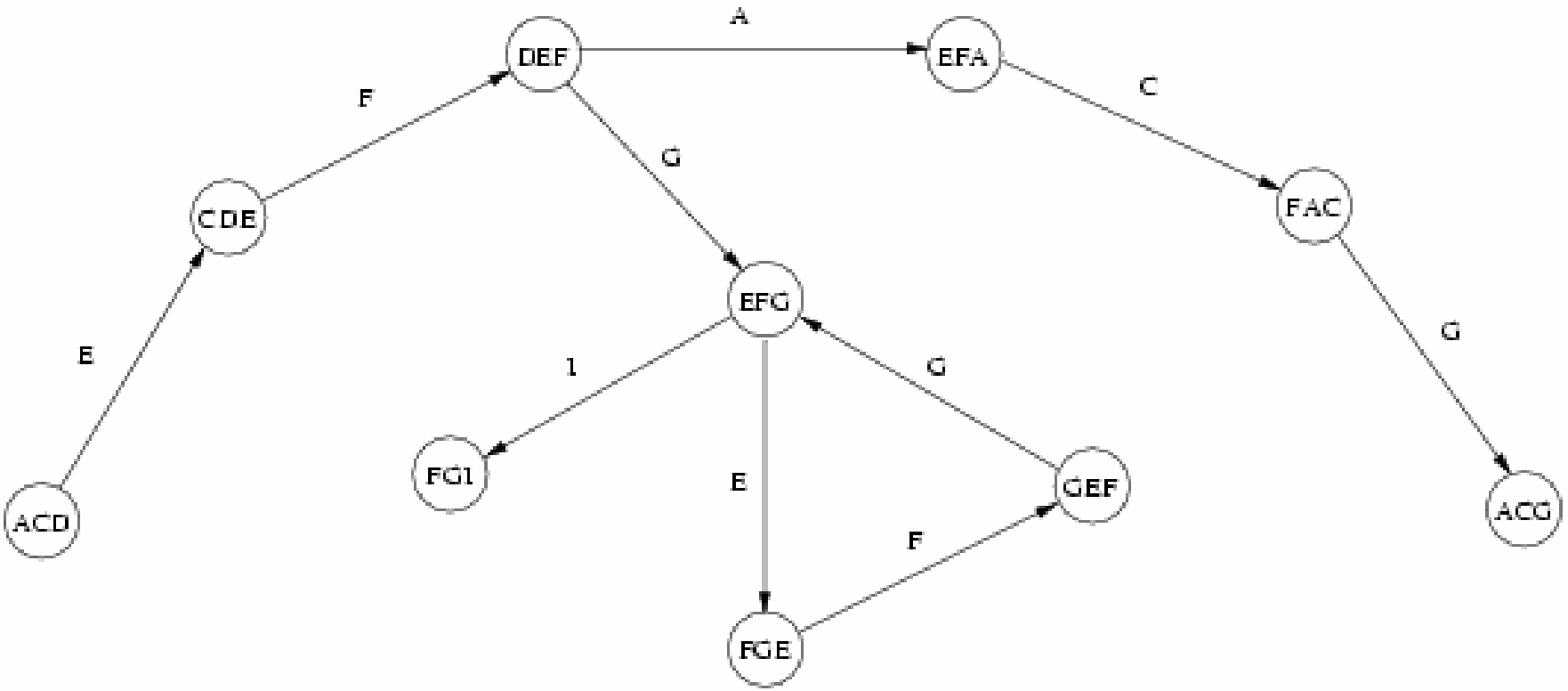
Sequence Database	Swiss-Prot	VarSplic
Size	60 Mb	95 Mb
30-mers (N_{30})	46 Mb	47 Mb
Overhead	29%	101%

Sequence Database Compression

Construct sequence database that is

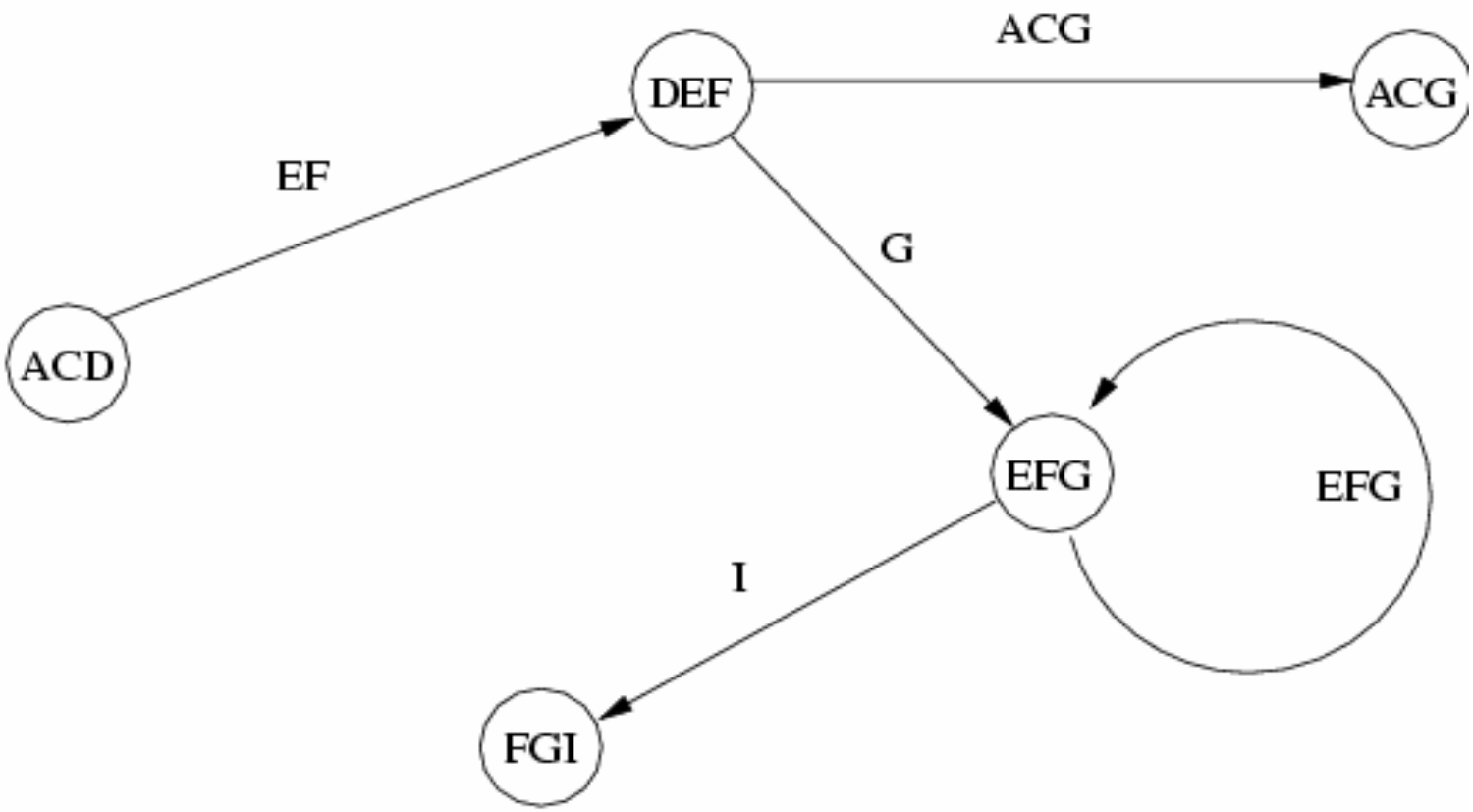
- Complete
  - All 30-mers are present
- Correct
  - No other 30-mers are present
- Compact
  - No 30-mer is present more than once

SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

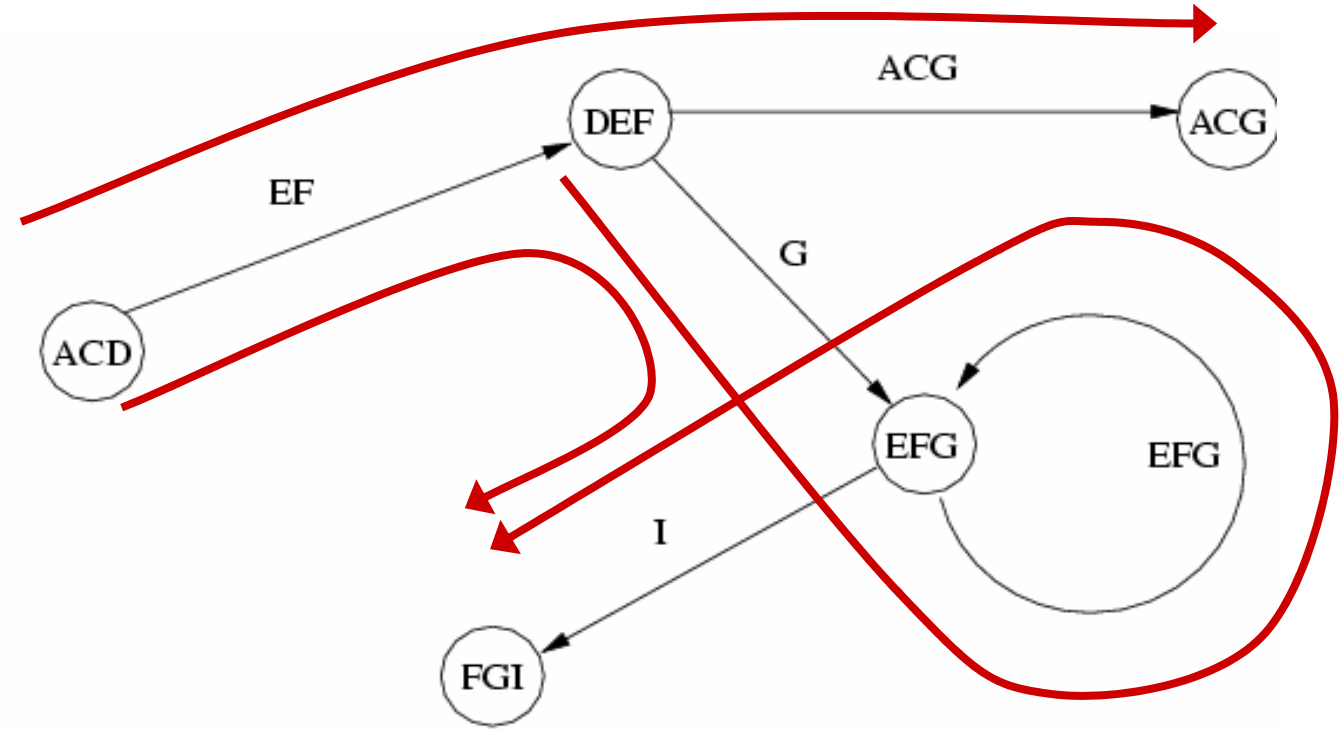
Compressed SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

Sequence Databases & CSBH-graphs

- Sequences correspond to paths



ACDEFGI, ACDEFACG, DEFGEFGI

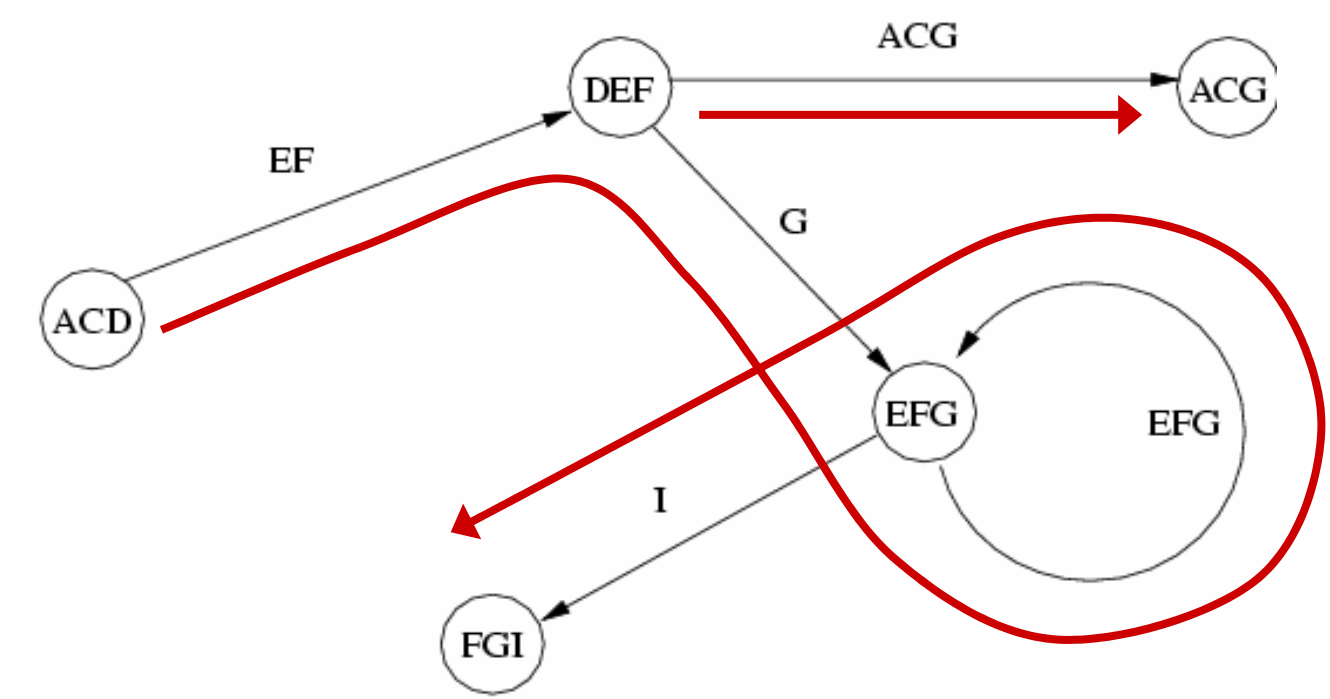
Sequence Databases & CSBH-graphs

- Complete
  - All edges are on some path
- Correct
  - Output path sequence only
- Compact
  - No edge is used more than once

- C^3 Path Set uses all edges exactly once.

Sequence Databases & CSBH-graphs

- Use each edge exactly once



ACDEFGEFGI, DEFACG

AA Sequence Databases

Sequence Database	Sequence Length	Distinct 30-mers	Overhead
IPI-HUMAN	20358846	12115520	68%
IPI	54145883	29769766	81%
Swiss-Prot	56454588	44374286	27%
Swiss-Prot-VS	89541275	45307827	97%
UniProt	472581860	274510105	72%
UniProt-VS	506796094	275391669	84%
MSDB	481919777	276523755	74%
NRP	495502241	283160529	75%
NCBI-nr	619132252	378721915	63%
UnionNR	674700840	385369671	75%
Union	2157353500	385369671	460%

Minimum Size C^3 Sequence Database

Sequence Database	C^3 30-mer Enumeration	Overhead	Compression	Compression Bound
IPI-HUMAN	13854679	14.35%	68.05%	59.51%
IPI	37961385	27.52%	70.11%	54.98%
Swiss-Prot	52662145	18.68%	93.28%	78.60%
Swiss-Prot-VS	54534356	20.36%	60.90%	50.60%
UniProt	337119564	22.81%	71.34%	58.09%
UniProt-VS	338890778	23.06%	66.87%	54.34%
MSDB	342924164	24.01%	71.16%	57.38%
NRP	351600578	24.17%	70.96%	57.15%
NCBI-nr	463517034	22.39%	74.87%	61.17%
UnionNR	473665310	22.91%	70.20%	57.12%
Union	473665310	22.91%	21.96%	17.86%

Implementation

- Suitable for use by Mascot, SEQUEST, X-Tandem, etc…
  - FASTA format
  - Use special “restart symbol”!
- All connection to protein context is lost
  - Must do exact string search to find peptides in original database

Implementation

>1
ACDEFGI
>2
ACDEFACG
>3
DEFGEFGI

>A
ACDEFGEFGI
>B
DEFACG

- Can't guarantee that "restarts" won't create "false" tryptic peptides

Implementation

- Mark original entries' start and end
- "Restart" with unusable symbol
  - Weight(J) = 10kDa
- Forbid tryptic digestion at J
  - Mascot cofiguration:

```
Title:Trypsin
Cleavage:KR
Restrict:PJ
Cterm
```

- What about "short" proteins?

- C^3 guarantee for tryptic peptides of length ≤ (k-2)

Faster Peptide ID

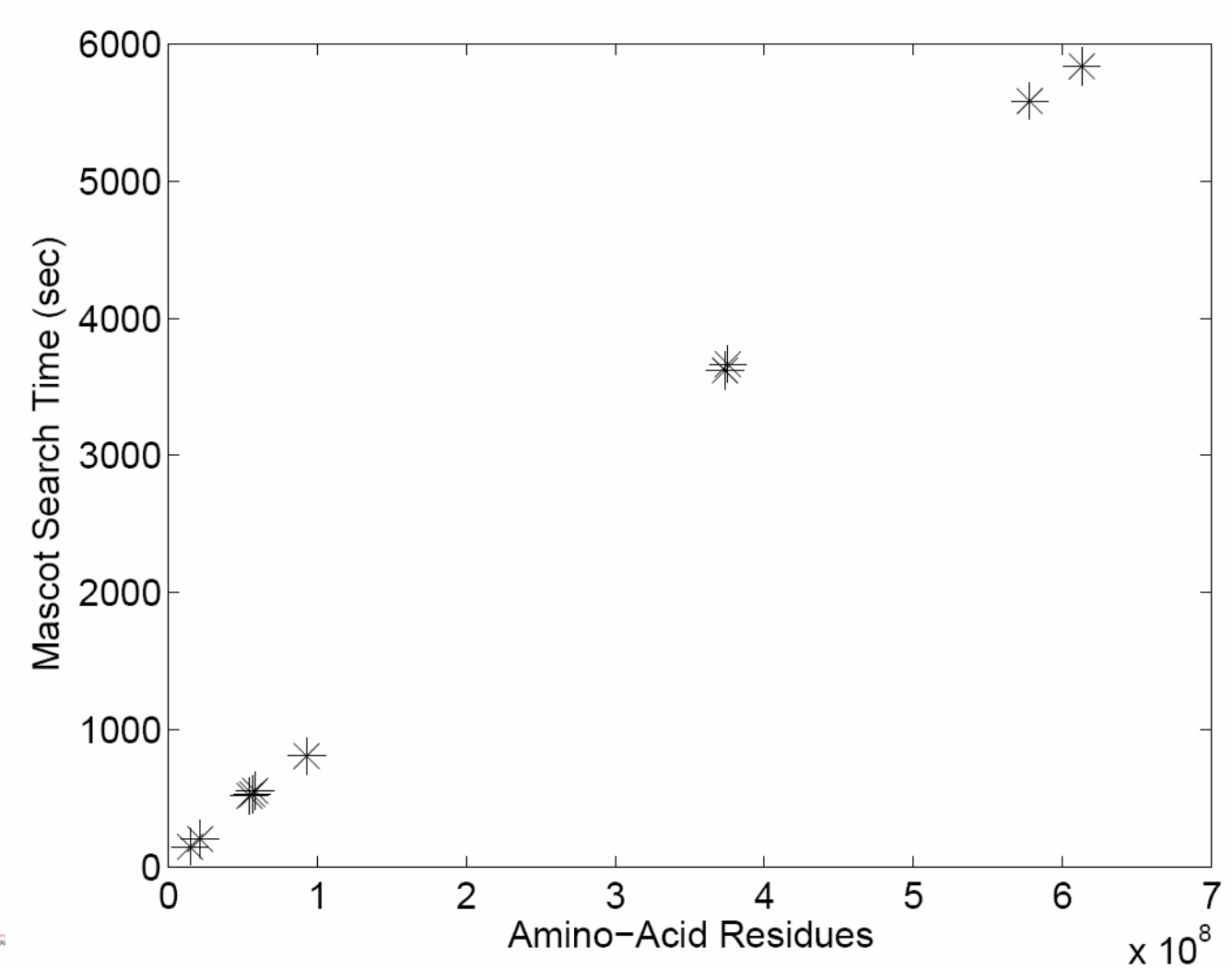
Peptide identification workflow

- Search spectra against C^3 sequence
- Insert protein context using exact string search
  - Parse search output
  - Search original sequence using exact peptide sequence
  - Re-construct search output

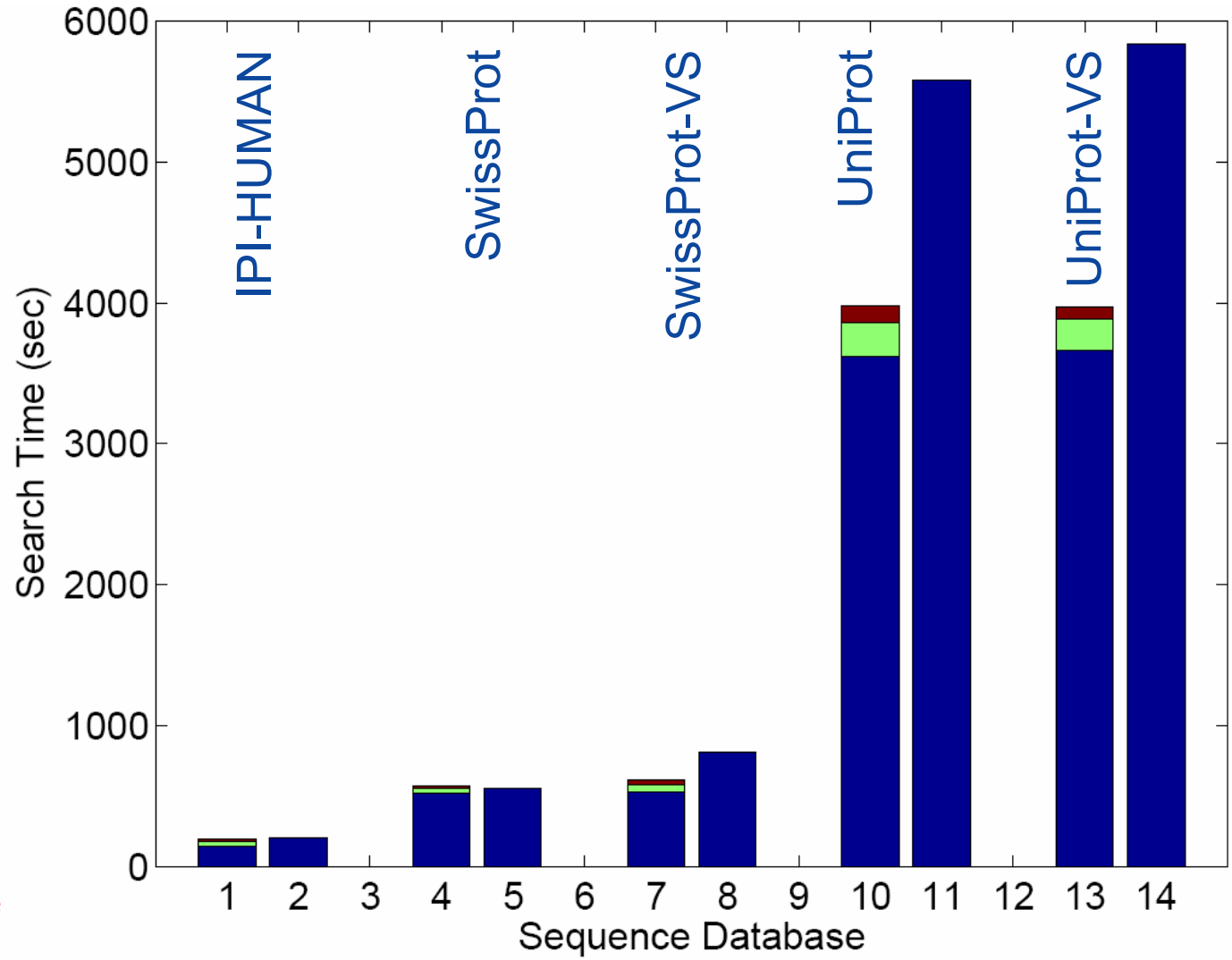
Faster Peptide ID

- ISB 17 Protein Mix
  - LC/MS/MS, 2023 MS/MS spectra
- Mascot v2.0
  - Dell PC w/ 512 Mb RAM
  - Parent Tolerance 2Da
  - Fragment Tolerance 0.15Da
  - Up to 2 Missed Cleavages
- IPI-HUMAN, SwissProt, UniProt
  - plus SwissProt-VS UniProt-VS versions

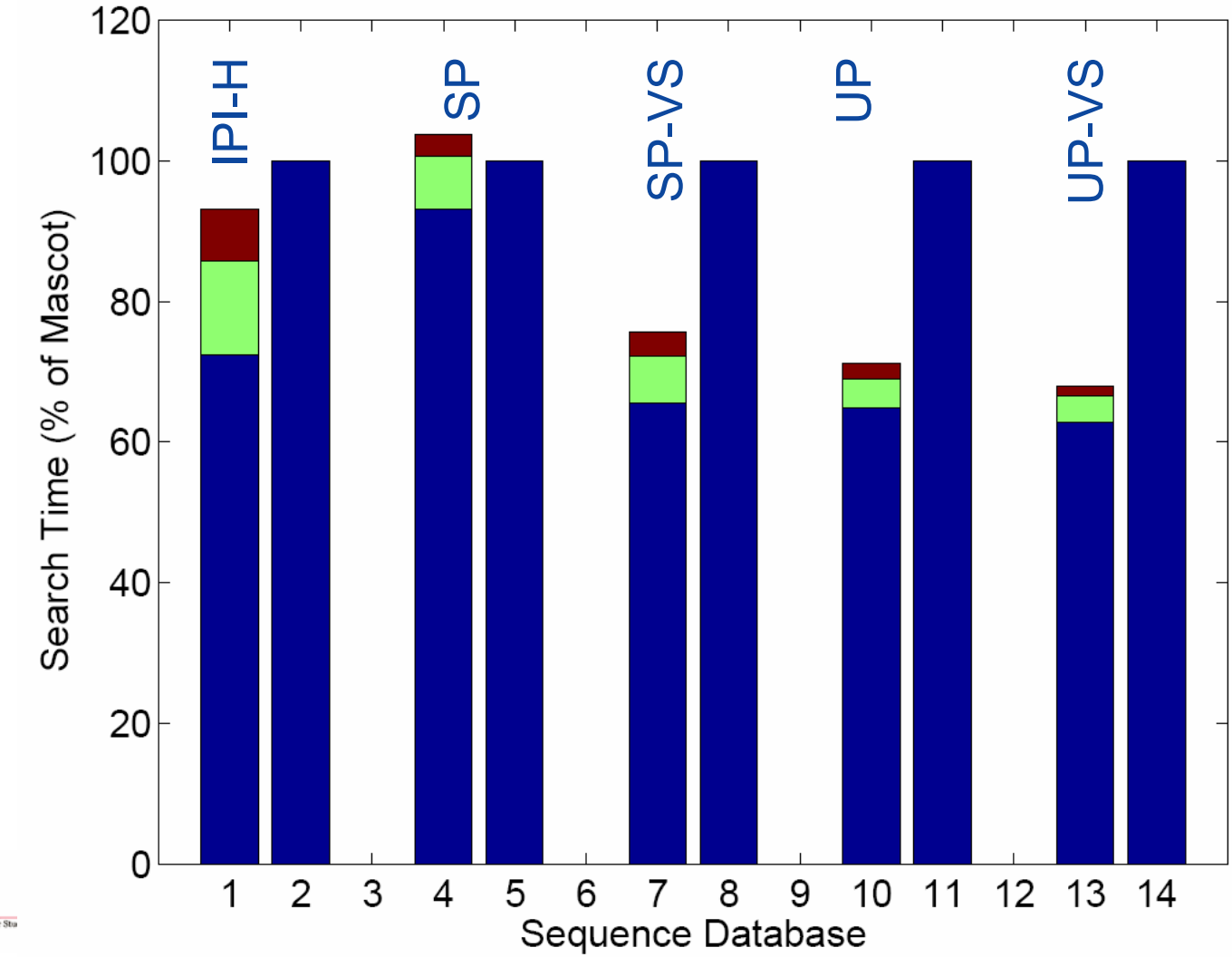
Mascot Running Time



Total Search Time



Relative Search Time



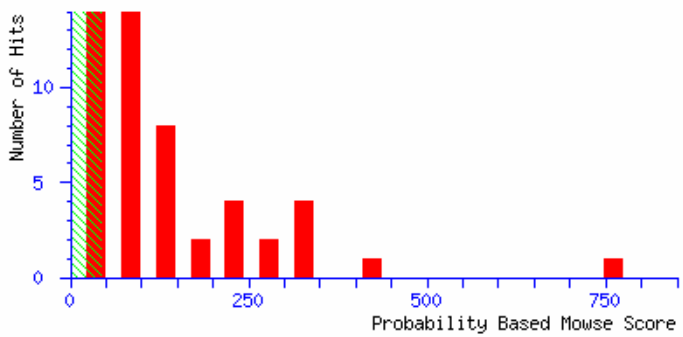
Mascot Results: C³ DB

{MATRIX SCIENCE} Mascot Search Results

```
User              : Seach for Compression Paper
Email             :
Search title      :  Conversion of 17mix_test2.xml to mascot generic by msxml2other
MS data file      : C:\Documents and Settings\edwardnj\Desktop\17prot_mix.mgf
Database          : VarsplicC3  (183075 sequences; 56651639 residues)
Timestamp         : 5 Oct 2004 at 14:09:37 GMT
Significant hits: AAAAAAHJAX  AAAAAAHJAX
                  AAAAAAEIYP  AAAAAAEIYP
                  AAAAAAJSWI  AAAAAAJSWI
                  AAAAAABOBV  AAAAAABOBV
                  AAAAAABITK  AAAAAABITK
                  AAAAAAAKFD  AAAAAAAKFD
                  AAAAAABITN  AAAAAABITN
                  AAAAAAANIU  AAAAAAANIU
                  AAAAAABOCD  AAAAAABOCD
                  AAAAAABOBE  AAAAAABOBE
                  AAAAAAENJK  AAAAAAJBOS
                  AAAAAABOCE  AAAAAAAKEZ
                  AAAAAAJBOS  AAAAAABITO
                  AAAAAAAKEZ  AAAAAAAEBT
                  AAAAAABITO  AAAAAAAKEY
                  AAAAAAAEBT  AAAAAABTLR
                  AAAAAAAKEY  AAAAAAAECF
                  AAAAAABTLR  AAAAAAAAPF
                  AAAAAAAECF
                  AAAAAAAAPF
```

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 43 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.



UMIACS
University of Maryland Institute for Advanced Computer Studies

Mascot Results: C³ DB

Peptide Summary Report

Switch to Protein Summary Report

To create a bookmark for this report, right click this link: Peptide Summary Report (Conversion of 17mix_test2.xml to mascot generic by msxml2other)

Select All | Select None | Search Selected | ☐ Error tolerant | Archive Report

1. AAAAAAHJAX **Mass:** 116278 **Score:** 762 **Peptides matched:** 38

AAAAAAHJAX

☐ Check to include this hit in error tolerant search or archive report

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑	587	435.70	869.39	869.42	-0.03	0	(24)	4.3	1	YWQAFR
☑	588	435.70	869.39	869.42	-0.03	0	25	3.6	1	YWQAFR
☑	589	435.70	869.39	869.42	-0.03	0	(22)	6.2	1	YWQAFR
☑	641	450.67	899.32	899.38	-0.06	0	39	0.19	1	FNDDFSR
☑	764	495.22	988.42	988.48	-0.06	0	39	0.2	1	WLPAMSER
☑	765	495.22	988.42	988.48	-0.06	0	(36)	0.36	1	WLPAMSER
☑	767	495.22	988.42	988.48	-0.06	0	(32)	0.94	1	WLPAMSER
☑	769	495.22	988.42	988.48	-0.06	0	(38)	0.25	1	WLPAMSER
☑	845	534.23	1066.44	1066.48	-0.05	0	31	0.72	1	WVGYGQDSR
☑	873	550.23	1098.45	1098.55	-0.10	0	(72)	6.5e-005	1	IDPNAWVER
☑	874	550.23	1098.45	1098.55	-0.10	0	73	5.9e-005	1	IDPNAWVER
☑	1020	626.80	1251.58	1251.65	-0.07	0	36	0.24	1	LAAHPPFASWR
☑	1021	626.80	1251.58	1251.65	-0.07	0	(33)	0.45	1	LAAHPPFASWR
☑	1037	633.26	1264.51	1264.61	-0.10	0	(24)	4.5	1	HQQQFFQFR
☑	1038	633.26	1264.51	1264.61	-0.10	0	37	0.23	1	HQQQFFQFR
☑	1072	650.28	1298.55	1298.62	-0.07	0	47	0.023	1	ELNYGPHQWR
☑	1132	671.31	1340.60	1340.66	-0.06	0	(63)	0.00049	1	VDEDQPFPAVPK
☑	1133	671.31	1340.60	1340.66	-0.06	0	(57)	0.0023	1	VDEDQPFPAVPK
☑	1134	671.31	1340.60	1340.66	-0.06	0	70	0.00011	1	VDEDQPFPAVPK
☑	1135	671.31	1340.60	1340.66	-0.06	0	(38)	0.15	1	VDEDQPFPAVPK
☑	1172	681.33	1360.65	1360.71	-0.06	0	38	0.16	1	LWSAEIPNLYR
☑	1173	681.33	1360.65	1360.71	-0.06	0	(29)	1.5	1	LWSAEIPNLYR

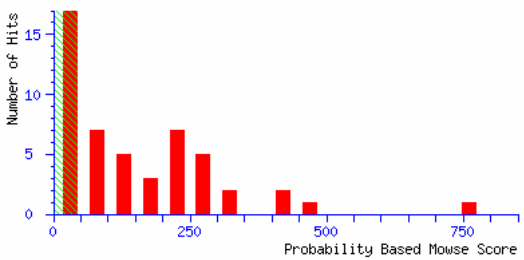
Mascot Results: C³ DB ++

MATRIX SCIENCE Mascot Search Results

```
User            : Seach for Compression Paper
Email           :
Search title    :  Conversion of 17mix_test2.xml to mascot generic by msxml2other
MS data file    : C:\Documents and Settings\edwardnj\Desktop\17prot_mix.mgf
Database        : Varsplic  (203425 sequences; 92944773 residues)
Timestamp       : 5 Oct 2004 at 14:09:37 GMT
Significant hits: P00722          (BGAL_ECOLI) Beta-galactosidase (EC 3.2.1.23) (Lactase)
                  P00634-00-00-00 (PPB_ECOLI) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P00634 Alkaline phosphatase precursor (
                  P00489          (PHS2_RABIT) Glycogen phosphorylase, muscle form (EC 2.4.1.1) (Myophosphorylase)
                  P00921-00-01-00 (CAH2_BOVIN) Splice isoform Displayed; Variant one of the major forms; Conflict Displayed; from P00921 Carbonic anhydrase
                  Q29443          (TRFE_BOVIN) Serotransferrin precursor (Transferrin) (Siderophilin) (Beta-1-metal binding globulin)
                  P02769-00-00-00 (ALBU_BOVIN) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P02769 Serum albumin precursor (Allerg
                  P21948          (PPB_ESCFE) Alkaline phosphatase precursor (EC 3.1.3.1) (APase)
                  Q28641          (MYH4_RABIT) Myosin heavy chain, skeletal muscle, juvenile
                  P02608          (MLRS_RABIT) Myosin regulatory light chain 2, skeletal muscle isoform type 2 (G2) (DTNB) (MLC-2)
                  P11217-00-05-00 (PHS2_HUMAN) Splice isoform Displayed; Variant GSD-V-VAR_014005; Conflict Displayed; from P11217 Glycogen phosphorylase,
                  P11217-00-12-00 (PHS2_HUMAN) Splice isoform Displayed; Variant GSD-V-VAR_014009; Conflict Displayed; from P11217 Glycogen phosphorylase,
                  P49064          (ALBU_FELCA) Serum albumin precursor (Allergen Fel d 2)
                  P12882          (MYH1_HUMAN) Myosin heavy chain, skeletal muscle, adult 1 (Myosin heavy chain IIx/d) (MyHC-IIx/d)
                  P07724          (ALBU_MOUSE) Serum albumin precursor
                  Q9Y623          (MYH4_HUMAN) Myosin heavy chain, skeletal muscle, fetal (Myosin heavy chain IIb) (MyHC-IIb)
                  P13538          (MYSS_CHICK) Myosin heavy chain, skeletal muscle, adult
                  P11055          (MYH3_HUMAN) Myosin heavy chain, fast skeletal muscle, embryonic (Muscle embryonic myosin heavy chain) (SMHCE)
                  P00432          (CATA_BOVIN) Catalase (EC 1.11.1.6)
                  P06278          (AMY_BACLI) Alpha-amylase precursor (EC 3.2.1.1) (1,4-alpha-D-glucan glucanohydrolase) (BLA)
                  P09812          (PHS2_RAT) Glycogen phosphorylase, muscle form (EC 2.4.1.1) (Myophosphorylase)
```

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 43 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.



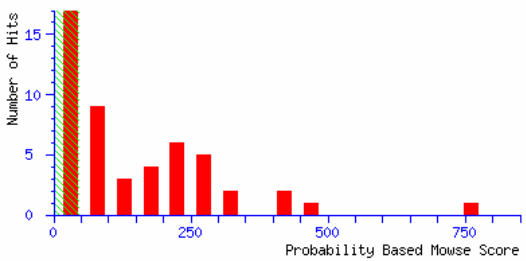
Mascot Results: Original DB

{MATRIX SCIENCE} Mascot Search Results

```
User            : Seach for Compression Paper
Email           :
Search title    :  Conversion of 17mix_test2.xml to mascot generic by msxml2other
MS data file    : C:\Documents and Settings\edwardnj\Desktop\17prot_mix.mgf
Database        : Varsplic  (203425 sequences; 92944773 residues)
Timestamp       : 5 Oct 2004 at 13:55:52 GMT
Significant hits: P00722           (BGAL_ECOLI) Beta-galactosidase (EC 3.2.1.23) (Lactase)
                  P00634-00-00-00  (PPB_ECOLI) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P00634 Alkaline p
                  P00921-00-01-00  (CAH2_BOVIN) Splice isoform Displayed; Variant one of the major forms; Conflict Displayed; from P00
                  P00489           (PHS2_RABIT) Glycogen phosphorylase, muscle form (EC 2.4.1.1) (Myophosphorylase)
                  Q29443           (TRFE_BOVIN) Serotransferrin precursor (Transferrin) (Siderophilin) (Beta-1-metal binding globulin)
                  P02769-00-00-00  (ALBU_BOVIN) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P02769 Serum alb
                  P21948           (PPB_ESCFE) Alkaline phosphatase precursor (EC 3.1.3.1) (APase)
                  P02608           (MLRS_RABIT) Myosin regulatory light chain 2, skeletal muscle isoform type 2 (G2) (DTNB) (MLC-2)
                  Q28641           (MYH4_RABIT) Myosin heavy chain, skeletal muscle, juvenile
                  P11217-00-05-00  (PHS2_HUMAN) Splice isoform Displayed; Variant GSD-V-VAR_014005; Conflict Displayed; from P11217 Gl
                  P11217-00-12-00  (PHS2_HUMAN) Splice isoform Displayed; Variant GSD-V-VAR_014009; Conflict Displayed; from P11217 Gl
                  P49064           (ALBU_FELCA) Serum albumin precursor (Allergen Fel d 2)
                  P07724           (MYH4_HUMAN) Myosin heavy chain, skeletal muscle, fetal (Myosin heavy chain IIb) (MyHC-IIb)
                  Q9Y623           (MYH1_HUMAN) Myosin heavy chain, skeletal muscle, adult 1 (Myosin heavy chain IIx/d) (MyHC-IIx/d)
                  P12882           (MYSS_CHICK) Myosin heavy chain, skeletal muscle, adult
                  P13538           (CATA_BOVIN) Catalase (EC 1.11.1.6)
                  P00432           (MYH3_HUMAN) Myosin heavy chain, fast skeletal muscle, embryonic (Muscle embryonic myosin heavy cha
                  P11055           (AMY_BACLI) Alpha-amylase precursor (EC 3.2.1.1) (1,4-alpha-D-glucan glucanohydrolase) (BLA)
                  P06278           (PHS2_RAT) Glycogen phosphorylase, muscle form (EC 2.4.1.1) (Myophosphorylase)
                  P09812
```

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 45 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.



Mascot Results: C³ DB ++

Peptide Summary Report

Switch to Protein Summary Report

To create a bookmark for this report, right click this link: Peptide Summary Report (Conversion of 17mix_test2.xml to mascot generic by msxml2other)

Select All | Select None | Search Selected | ☐ Error tolerant | Archive Report

1. P00722 **Mass:** 116278 **Score:** 762 **Peptides matched:** 38
(BGAL_ECOLI) Beta-galactosidase (EC 3.2.1.23) (Lactase)
☐ Check to include this hit in error tolerant search or archive report

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑	587	435.70	869.39	869.42	-0.03	0	(24)	4.3	1	YWQAFR
☑	588	435.70	869.39	869.42	-0.03	0	25	3.6	1	YWQAFR
☑	589	435.70	869.39	869.42	-0.03	0	(22)	6.2	1	YWQAFR
☑	641	450.67	899.32	899.38	-0.06	0	39	0.19	1	FNDDFSR
☑	764	495.22	988.42	988.48	-0.06	0	39	0.2	1	WLPAMSER
☑	765	495.22	988.42	988.48	-0.06	0	(36)	0.36	1	WLPAMSER
☑	767	495.22	988.42	988.48	-0.06	0	(32)	0.94	1	WLPAMSER
☑	769	495.22	988.42	988.48	-0.06	0	(38)	0.25	1	WLPAMSER
☑	845	534.23	1066.44	1066.48	-0.05	0	31	0.72	1	WVGYGQDSR
☑	873	550.23	1098.45	1098.55	-0.10	0	(72)	6.5e-005	1	IDPNAWVER
☑	874	550.23	1098.45	1098.55	-0.10	0	73	5.9e-005	1	IDPNAWVER
☑	1020	626.80	1251.58	1251.65	-0.07	0	36	0.24	1	LAAHPPFASWR
☑	1021	626.80	1251.58	1251.65	-0.07	0	(33)	0.45	1	LAAHPPFASWR
☑	1037	633.26	1264.51	1264.61	-0.10	0	(24)	4.5	1	HQQQFFQFR
☑	1038	633.26	1264.51	1264.61	-0.10	0	37	0.23	1	HQQQFFQFR
☑	1072	650.28	1298.55	1298.62	-0.07	0	47	0.023	1	ELNYGPHQWR
☑	1132	671.31	1340.60	1340.66	-0.06	0	(63)	0.00049	1	VDEDQPFPAVPK
☑	1133	671.31	1340.60	1340.66	-0.06	0	(57)	0.0023	1	VDEDQPFPAVPK
☑	1134	671.31	1340.60	1340.66	-0.06	0	70	0.00011	1	VDEDQPFPAVPK
☑	1135	671.31	1340.60	1340.66	-0.06	0	(38)	0.15	1	VDEDQPFPAVPK
☑	1172	681.33	1360.65	1360.71	-0.06	0	38	0.16	1	LWSAEIPNLYR

Mascot Results: Original DB

Peptide Summary Report

Switch to Protein Summary Report

To create a bookmark for this report, right click this link: Peptide Summary Report (Conversion of 17mix_test2.xml to mascot generic by msxml2other)

Select All | Select None | Search Selected | ☐ Error tolerant | Archive Report

1. P00722 **Mass:** 116278 **Score:** 760 **Peptides matched:** 38
(BGAL_ECOLI) Beta-galactosidase (EC 3.2.1.23) (Lactase)
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑ 587	435.70	869.39	869.42	-0.03	0	(24)	6.7	1	YWQAFR
☑ 588	435.70	869.39	869.42	-0.03	0	25	5.7	1	YWQAFR
☑ 589	435.70	869.39	869.42	-0.03	0	(22)	9.6	1	YWQAFR
☑ 641	450.67	899.32	899.38	-0.06	0	39	0.29	1	FNDDFSR
☑ 764	495.22	988.42	988.48	-0.06	0	39	0.35	1	WLPAMSER
☑ 765	495.22	988.42	988.48	-0.06	0	(36)	0.61	1	WLPAMSER
☑ 767	495.22	988.42	988.48	-0.06	0	(32)	1.6	1	WLPAMSER
☑ 769	495.22	988.42	988.48	-0.06	0	(38)	0.43	1	WLPAMSER
☑ 845	534.23	1066.44	1066.48	-0.05	0	31	1.2	1	WVGYGQDSR
☑ 873	550.23	1098.45	1098.55	-0.10	0	(72)	0.00011	1	IDPNAWVER
☑ 874	550.23	1098.45	1098.55	-0.10	0	73	9.8e-005	1	IDPNAWVER
☑ 1020	626.80	1251.58	1251.65	-0.07	0	36	0.42	1	LAAHPPFASWR
☑ 1021	626.80	1251.58	1251.65	-0.07	0	(33)	0.78	1	LAAHPPFASWR
☑ 1037	633.26	1264.51	1264.61	-0.10	0	(24)	8	1	HQQQFFQFR
☑ 1038	633.26	1264.51	1264.61	-0.10	0	37	0.4	1	HQQQFFQFR
☑ 1072	650.28	1298.55	1298.62	-0.07	0	47	0.039	1	ELNYGPHQWR
☑ 1132	671.31	1340.60	1340.66	-0.06	0	(63)	0.00082	1	VDEDQPFPAVPK
☑ 1133	671.31	1340.60	1340.66	-0.06	0	(57)	0.0038	1	VDEDQPFPAVPK
☑ 1134	671.31	1340.60	1340.66	-0.06	0	70	0.00018	1	VDEDQPFPAVPK
☑ 1135	671.31	1340.60	1340.66	-0.06	0	(38)	0.24	1	VDEDQPFPAVPK
☑ 1172	681.33	1360.65	1360.71	-0.06	0	38	0.26	1	LWSAEIPNLYR

More Sensitive Peptide ID

- Significances, *p*-values, Expect values
  - Normalize for number of trials
  - Blast:
    - Size of sequence database
  - Mascot etc.:
    - Peptides with appropriate parent mass
  - Redundant peptide sequences increase the number of trials, artificially.
    - Trials are not independent!
  - Less redundancy results in a better significance estimate

Available for Download

Available for download:

- C^3 Sequence Databases
- Exact peptide search code
- Mascot parse + rewrite scripts

Other search engines and sequence databases on request

Genomic Peptide Sequences

- Many putative peptide sequences never become “protein” sequences
  - Genomic DNA,
  - Refseq mRNA, ESTs
  - SNP/Polymorphism databases
  - Variant records in SwissProt
- Genomic annotation seeks “full length” genes and proteins

Genomic Peptide Sequences

- Genomic DNA
  - Exons & introns, 6 frames, large (6 * 3Gb)
- Refseq mRNA
  - No introns, 3 frames, small (36Mb)
  - Most protein sequences already represented in sequence databases
- ESTs
  - No introns, 3 frames, large (3 * 3Gb)
  - Used by gene & alternative splicing pipelines
  - Highly redundant, nucleotide error rate ~ 1%

Compressed EST Peptides

- 3 frame translation
- Break at non-amino-acids
  - stop codons + X
- Discard AA sequence < 50 AA long
- Result: 1.1Gb
- C3 Compress
  - Discard 30-mers that occur only once
- Result: 83Mb (< 2 * SwissProt)

Compressed EST Peptides

- Genomic Peptides
  - Refseq mRNA & Proteins, ESTs
  - Size: 96Mb
- Genomic + IPI-HUMAN Peptides
  - Size: 99Mb
  - 30-mers:
    - 11.2Mb 30-mers in common
    - 1.6Mb in IPI-HUMAN alone
    - 37.2Mb in putative genomic peptides alone

Extensions and Futher Work

- Better compression
  - Relax compactness constraint
  - Enumerate tryptic peptides only
  - Relax correctness constraint
- Decouple peptide ID from protein ID
- Genomic peptides from GenScan exons

Thanks

- Informatics Research @ ABI & Celera
  - Ross Lippert, Clark Mobarry, Bjarni Halldorsson
- UMIACS @ University of Maryland, CP
  - V.S. Subrahmanian, Fritz McCall, Doan Pham
- Fenselau Lab @ UM, CP