Protein Databases for Mass Spectrometric Analysis

Raghothama Chaerkady, Ph.D.

McKusick-Nathans Institute of Genetic Medicine and the Department of Biological Chemistry
http://pandeylab.igm.jhmi.edu/

Data base file

Mass spec file

Search engine that uses mass spectrometry data to identify proteins from primary sequence databases

MASCOT **http://www.matrixscience.com/**

Initiated at Imperial Cancer Research Fund, UK

Mascot utilizes code from MOWSE (Developed by Darryl Pappin and David Perkins) MOWSE → Molecular Weight Search

MOWSE database fully indexed → Rapid search from a non-redundant protein sequence database

MOWSE uses scoring algorithm initially for peptide mass fingerprint on the basis of non-uniform distribution of peptide sizes which result from digestion by an enzyme.

Mass Spectrom. Biol. Sci., 135-150 (1996)

Search engine that uses mass spectrometry data to identify proteins from primary sequence databases

MOWSE II, involve amino acid sequence and composition qualifiers
Uses indexed molecular weight databases

Drawback

Database had to be built for each new enzyme and for each set of amino acid residue masses.

Difficult for searching proteins in which residues had been chemically or post-translationally modified.

MOWSE III New algorithms
Facility to specify selected MS/MS fragment ion masses as an "ions" qualifier to a peptide mass value
Now available as mascot from www.matrixscience.com

SEQUEST

Developed by Jimmy Eng/John Yates

Correlates uninterpreted tandem mass spectra of peptides with amino acid sequences from protein and nucleotide databases.

SEQUEST will determine the amino acid sequence and thus the protein(s) and organism(s) that correspond to the mass spectrum being analyzed.

ProteinProspector
http://prospector.ucsf.edu/

PeptideSearch
http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html

X! Tandem
The Global Proteome Machine, Advanced Search Page (GPM)
X! Tandem open source is software that can match tandem mass spectra with peptide sequences
http://gpm.igm.jhmi.edu/tandem/thegpm_tandem_a.html

OMSSA
The Open Mass Spectrometry Search Algorithm [OMSSA] is a faster search engine for identifying MS/MS peptide spectra by searching libraries of known protein sequences.

Classic probability score using an explicit model for matching experimental spectra to sequences

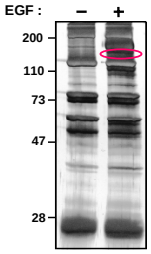
The Myth of Kozak's Consensus Sequence: Translation Initiation Codon

- **CCACCATGG**
- **Most upstream ATG used for translation initiation**
- **Biologists look for this sequence and annotate any ATG near the 5' end of the clone as the initiator methionine**

N-terminal Acetylation

- **Perhaps the most common co-translational modification (60-85% of proteins in yeast)**
- **Usually, aminopeptidases cleave one or two N-terminal amino acids followed by acetylation of the 'mature' protein**
- **So, if you find an N-acetylated peptide, the initiation methionine can be established.**

MS-Based Identification of a 130 kDa Protein in the EGF Receptor Signaling Pathway



Silver stained gel

IP : anti-pTyr

Assignment of the initiator methionine in a cDNA 'fragment' based on an N-terminal peptide

>KIAA0229 (1180 residues) FRAGMENT

SWGKGREGVVSPAGLGGALPGDGKFGSPSRLGCSLGEGVQRVAALGMGKEQ
ELLRAARTGHLPAVEKLLSGKRLSSGFGGGGGGGSGGGGGGSGGGGGGLGS
SSHPLSSLLSMWRGPNVNCVDSTGYTPLHHAALNGHHRRSSSSRSQDSAEGQ
DGQVPEQFSGLLHGSSPVCEVGQDPFQLLCTAGQSHPDGSPQQGACHKASM
QLEETGVHAPGASQPSALDQSKRVGYLTGLPTTNSRSHPETLTHTASPHPGGA
EEGDRSGAR

Assignment of the initiator methionine in a cDNA 'fragment' based on an N-terminal peptide

>KIAA0229 (1180 residues) FRAGMENT

SWGKGREGVVSPAGLGGALPGDGKFGSPSRLGCSLGEGVQRVAALG**MGKEQ**
LLRAARTGHLPAVEKLLSGKRLSSGFGGGGGGGSGGGGGGSGGGGGGLGSS
SHPLSSLLSMWRGPNVNCVDSTGYTPLHHAALNGHHRRSSSSRSQDSAEGQD
GQVPEQFSGLLHGSSPVCEVGQDPFQLLCTAGQSHPDGSPQQGACHKASMQL
EETGVHAPGASQPSALDQSKRVGYLTGLPTTNSRSHPETLTHTASPHPGGAEE
GDRSGAR

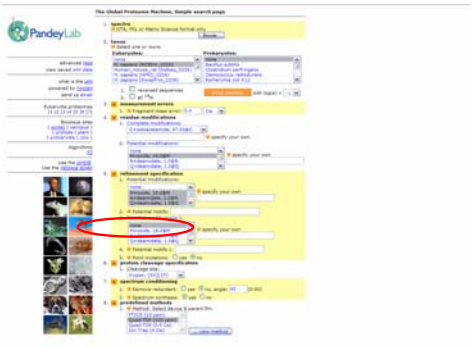
N-terminal Acetylated Peptide – Annotation of Start Codon

Aligment of sequences from 5 species in databases. The sequence at the top (XP_371848) is the human protein predicted by gene prediction programs. Peptides identified by MS/MS are marked in bold red and conserved residues are marked with an asterisk. The open reading frame in the case of zebra fish was the only correctly annotated entry. The acetylated methionine in the case of the peptide provides clear evidence that this methionine residue marks the N-terminus of this family of proteins.

Detecting Single Nucleotide Polymorphisms and Mutations at the Protein Level

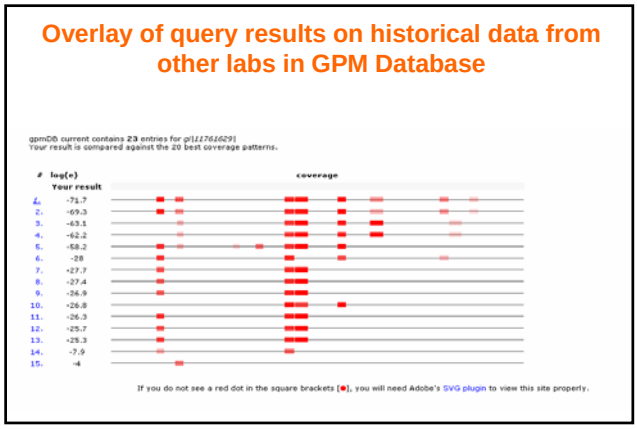
- Hit against the genome/proteome with a single amino acid change
- 'Point mutation' option in X!Tandem (and also available in Spectrum Mill)

X!Tandem

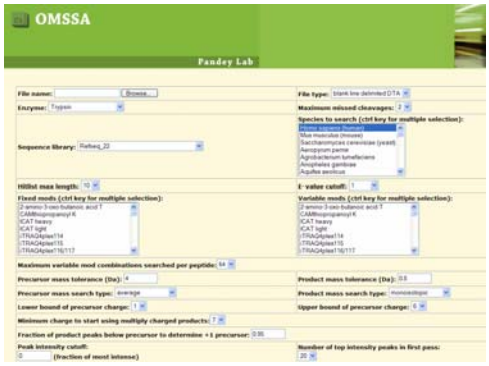


http://human.thegpm.org/tandem/thegpm_tandem.html

Overlay of query results on historical data from other labs in GPM Database



Open Mass Spectrometry Search Algorithm [OMSSA]



Protein Databases

- Swiss-Prot
- nr (non-redundant protein database)
- RefSeq
- IPI (International Protein Index)

Swiss-Prot

http://us.expasy.org/sprot/

- Swiss-prot is part of the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics.
- A highly curated protein sequence database with minimal redundancy
- Swiss-Prot currently contains 172,000 protein sequences representing 8,859 species
- 13,500 Human protein sequences

TrEMBL

http://us.expasy.org/sprot/

- **TrEMBL – A computer annotated supplement of Swiss-Prot containing all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot**
- **TrEMBL can be considered as a preliminary section of Swiss-Prot**
- **TrEMBL is split in two main sections: SPTrEMBL and REMTrEMBL**
- **SPTrEMBL – All TrEMBL entries that should finally be upgraded to the standard Swiss-Prot quality, are assigned Swiss-Prot accessions**
- **REMTrEMBL – Remaining TrEMBL entries**

UniProt

http://www.expasy.uniprot.org/

Swiss-Prot

+

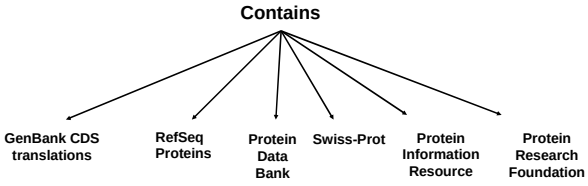
TrEMBL

+

PIR

→ **UniProt**
(Universal Protein Resource)

nr (non-redundant) database



nr (non-redundant) database

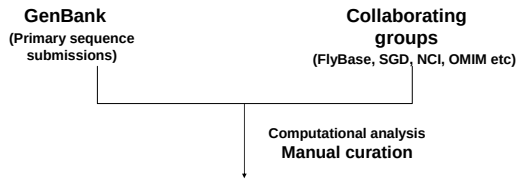
- **All identical sequences from any of the above databases are merged into a single entry**
- **It contains 1,800,000 protein sequences from 33,362 species**
- **Still NOT non-redundant (=VERY Redundant)**

RefSeq (Reference Sequence) database

http://www.ncbi.nlm.nih.gov/RefSeq/

- **RefSeq database is a result of collaborative effort of NCBI and other groups and databases like TIGR, FlyBase, WormBase etc.**
- **A comprehensive, integrated and highly non-redundant curated protein sequence database**
- **29,500 Human protein sequences**
- **Contains protein sequences from all major research organisms**
- **Alternate splice forms listed individually**
- **Also contains predicted proteins translated from predicted transcripts (designated as XP_ entries)**

RefSeq (Reference Sequence) database

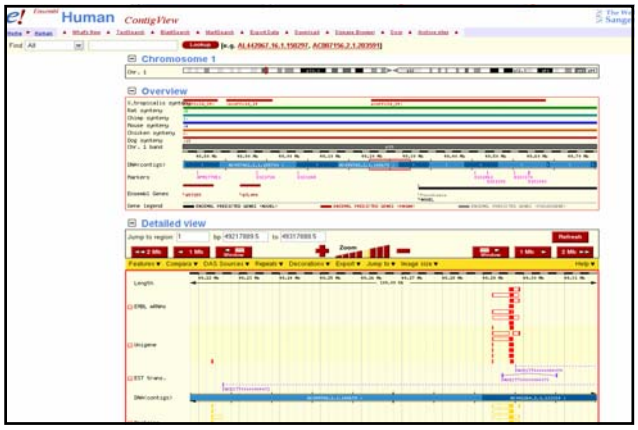


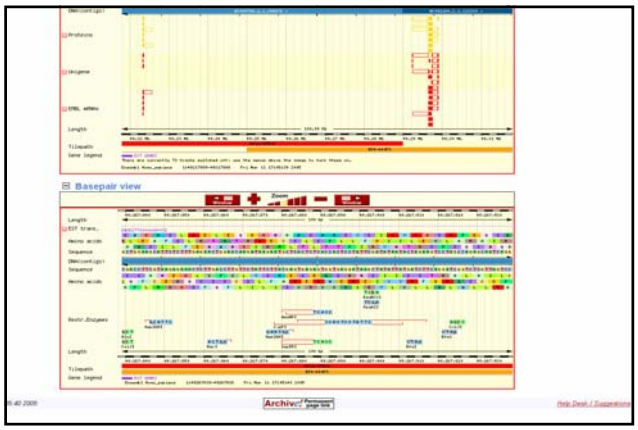
Unique representative, non-redundant RefSeq protein sequence entries

Ensembl database

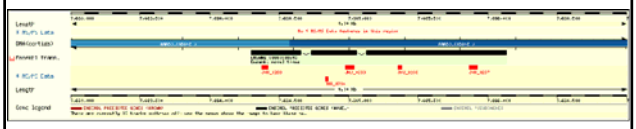
http://www.ebi.ac.uk/ensembl/

- **Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute that aims at developing a system that maintains automatic annotation of large eukaryotic genomes. database is a result of collaborative effort of NCBI and other groups and databases like TIGR, FlyBase, WormBase etc.**
- **It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources.**



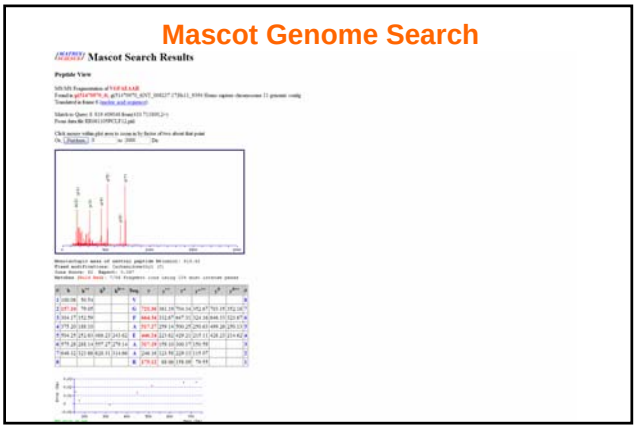


Use of Ensembl Distributed Annotation System to Validate a Predicted Transcript

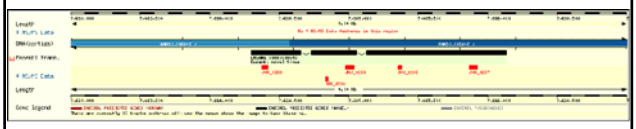


Correction of a Predicted Transcript

Mascot Genome Search



Use of Ensembl Distributed Annotation System to Validate a Predicted Transcript

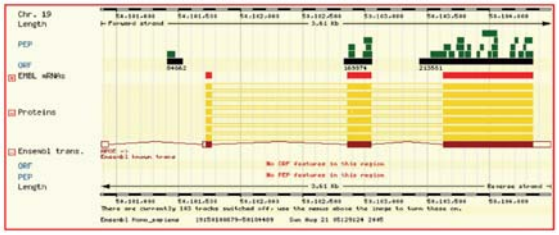


Kalume *et al*
BMC Genomics 2005

Correction of a Predicted Transcript

Kalume *et al*
BMC Genomics 2005

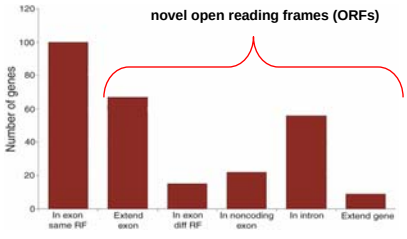
Mapping of Open Reading Frames (ORFs)



Dataset – Human PPP

States *et al*
Nat. Biotechnol. 2006

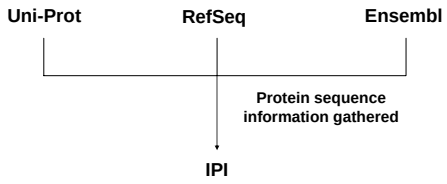
Statistics of ORFs Mapped to Genome



States *et al*
Nat. Biotechnol. 2006

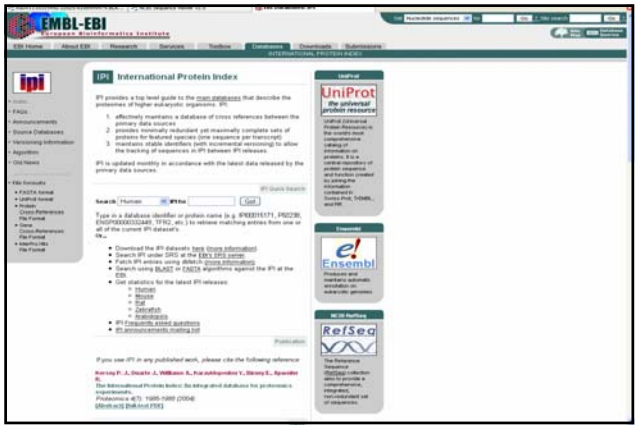
IPI (International Protein Index) database

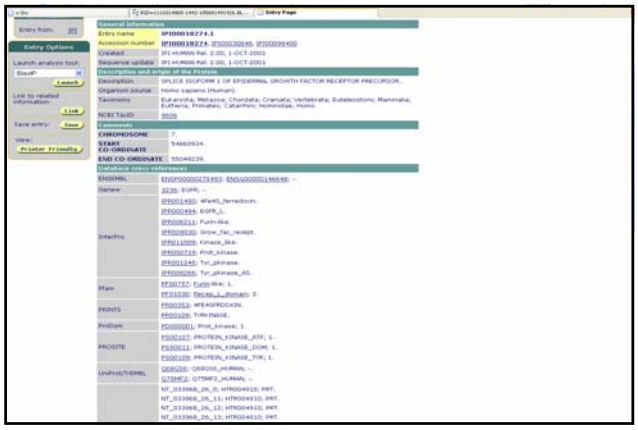
http://www.ebi.ac.uk/IPI/



IPI (International Protein Index) database

- **IPI is a protein database from the European Bioinformatics Institute**
- **Has protein sequence information from Human, Mouse, Rat, Zebra fish and Arabidopsis species only**
- **58,000 Human protein sequences**
- **A redundant database**
- **Has information on protein isoforms**
- **The sequence identifiers and sequence entries are not generally stable**





Databases/Tools for protein information

Protein Information resources

Databases

- **Swiss-Prot (http://us.expasy.org/sprot/)**
- **HPRD (Human Protein Reference Database) (http://www.hprd.org)**

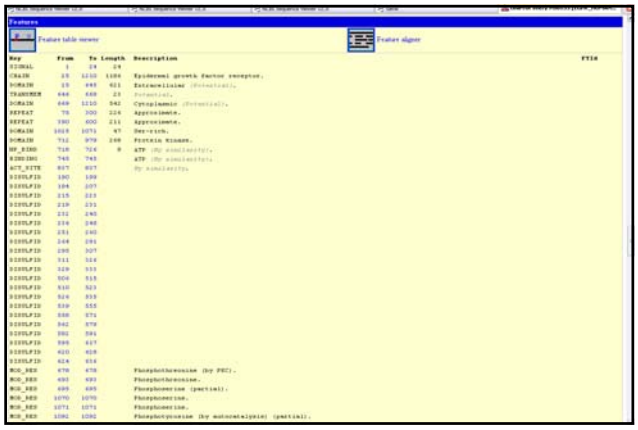
Tools

- **SMART (http://smart.embl-heidelberg.de/)**
- **Pfam (http://www.sanger.ac.uk/Software/Pfam/)**
- **PSORT (http://psort.nibb.ac.jp/)**

Swiss-Prot

Type of information than can be obtained for the protein of interest

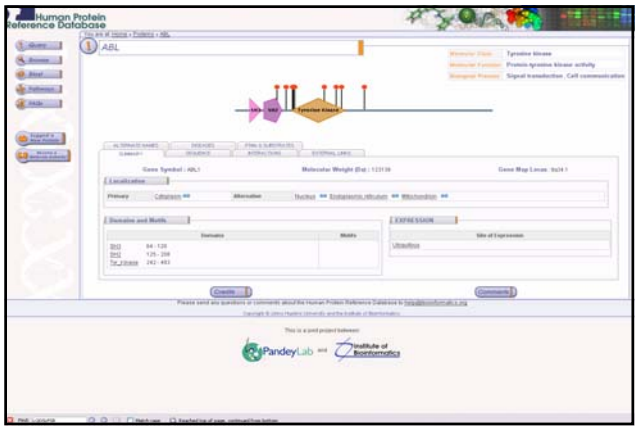
- **Function**
- **Architecture of protein (e.g. Domains, motifs)**
- **Post-translational modifications**
- **Alternate splice forms**
- **Localization**
- **Protein variants**
- **Cross-References to many other databases**



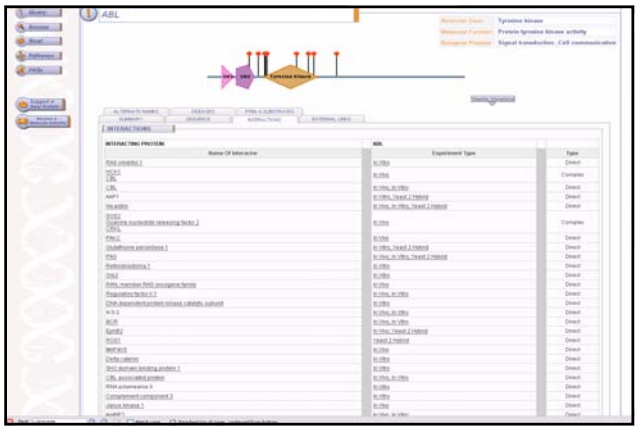
HPRD (Human Protein Reference Database)

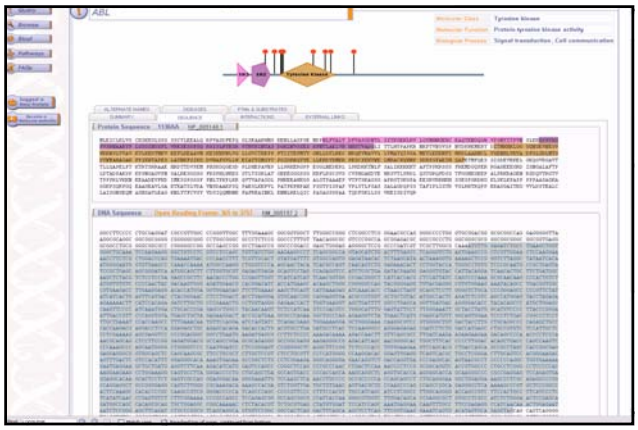
Type of information than can be obtained for the protein of interest

- **Function**
- **Architecture of protein (e.g. Domains, motifs)**
- **Post-translational modifications**
- **Expression**
- **Localization**
- **Disease associations**
- **Protein-protein interactions**



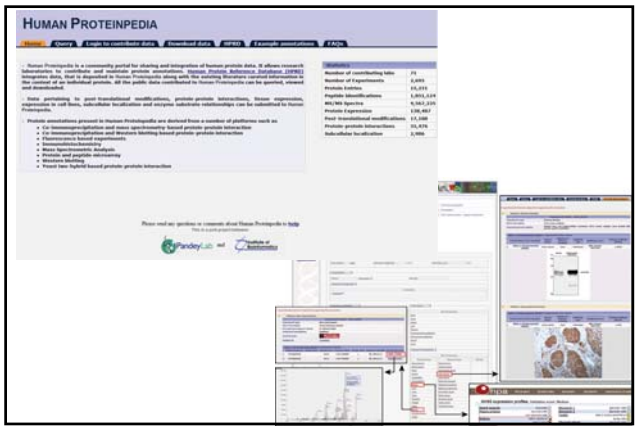

Human Protein Reference Database
ABL


ABL


ABL

Status of HPRD

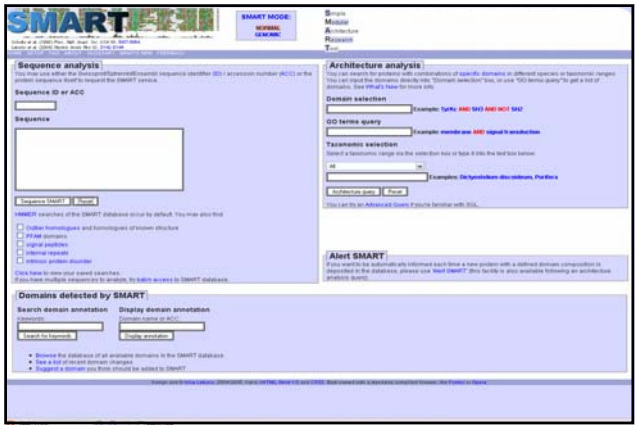
- Over 20,000 proteins annotated
- All 1,864 human disease genes in OMIM annotated
- Over 170,000 PubMed links provided (derived from reading of over 2,000,000 full-text articles)
- 25 types of post-translational modifications (PTM) annotated from literature
- 8,400 PTM sites annotated
- Over ~33,000 binary interactions annotated
- Compatible with Gene Ontology, PSI-MI, Cytoscape



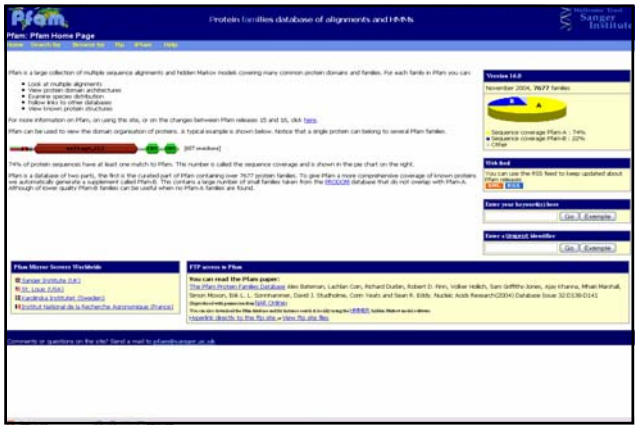
Protein Information resources

Tools

- **SMART (http://smart.embl-heidelberg.de/)**
- **Pfam (http://www.sanger.ac.uk/Software/Pfam/)**
- **PSORT (http://psort.nibb.ac.jp/)**




SMART
Domains within the query sequence gi|51709929|ref|XP_485348.1| of 390 residues


Pfam
Protein families database of alignments and HMMs
Pfam: Pfam Home Page


PSORT:
Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequences
PSORT WWW Server
CONTENTS
PSORT II (Recommended for animal/yeast sequences)
PSORT II Users' Manual
PSORT II Prediction
PSORT (Old version; for bacterial/plant sequences)
PSORT Users' Manual (WWW version)
PSORT Prediction
iPSORT (Detection of N-terminal sorting signals)
iPSORT Prediction
How to Obtain iPSORT (caml-iPSORT)
Other Information

Human Genome Annotation

Genome Annotation by Mass Spectrometry: What Can We Gain?

- Assigning start codons
- Proteins isoforms (alternative splicing, novel exons)
- Novel genes (proteins less than 100 amino acids not predicted by programs)
- cSNPs
- Correction of incorrect gene predictions (50% of the genes in human are predicted)
- Validation of gene predictions

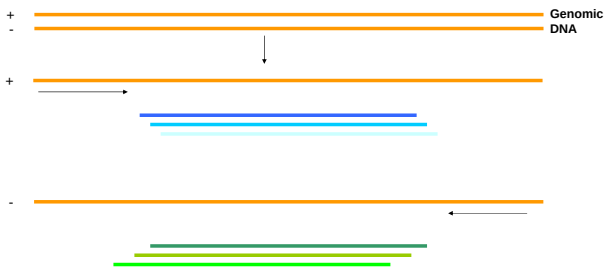
When is a peptide not identified from a database search?

- Protein not described (i.e. novel protein)
- Polymorphisms
- Alternative splice forms
- Novel exon
- Wrong annotation

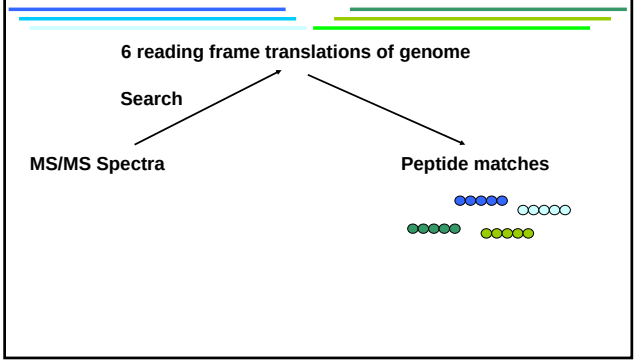
How do you identify such events?

- For novel genes and novel exons use the human genome sequence
- For polymorphisms and alternate splice forms, use a computational strategy

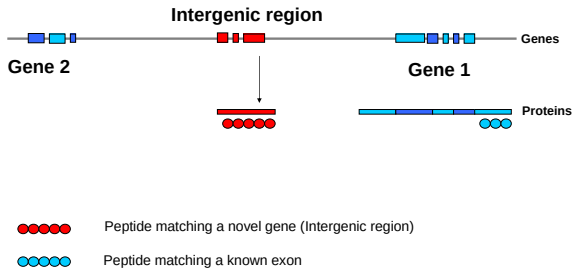
Genome Search



Genome Search



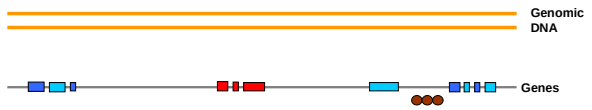
Peptide mapping onto the genome – Identifying a novel gene



Peptide matching a novel gene (Intergenic region)

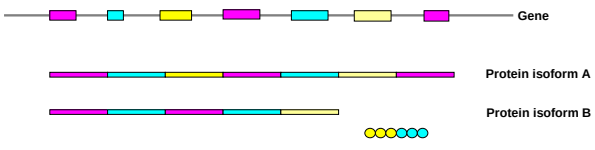
Peptide matching a known exon

Peptide mapping onto the genome – Identifying a novel exon



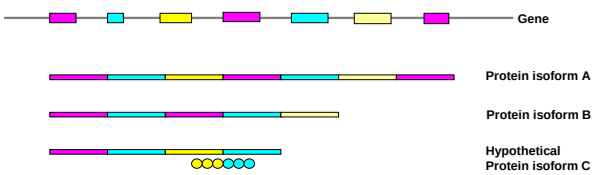
Peptide matching intronic region

Alternate splice forms



No Match!!

Alternate splice forms



Peptide matches!!

Alternate splice forms

Gene Symbol: HSPA8

```
NP_694881.1 MSKGPAVGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQVA 60
NP_006588.1 MSKGPAVGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQVA 60
            ************************************************************

NP_694881.1 MNPTNTVFDAKRLIGRRFDDAVVQSDMKHWPFMVVNDAGRPKVQVEYKGETKSFYPEEVS 120
NP_006588.1 MNPTNTVFDAKRLIGRRFDDAVVQSDMKHWPFMVVNDAGRPKVQVEYKGETKSFYPEEVS 120
            ************************************************************

NP_694881.1 SMVLTKMKEIAEAYLGKTVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAA 180
NP_006588.1 SMVLTKMKEIAEAYLGKTVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAA 180
            ************************************************************

NP_694881.1 IAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDGIFEVKSTAGDTHLGGEDFDNRMVNH 240
NP_006588.1 IAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDGIFEVKSTAGDTHLGGEDFDNRMVNH 240
            ************************************************************

NP_694881.1 FIAEFKRKHKKDISENKRAVRRLRTACERAKRTLSSSTQASIEIDSLYEGIDFYTSITRA 300
NP_006588.1 FIAEFKRKHKKDISENKRAVRRLRTACERAKRTLSSSTQASIEIDSLYEGIDFYTSITRA 300
            ************************************************************

NP_694881.1 RFEELNADLFRGTLDPVEKALRDAKLDKSQIHDIVLVGGSTRIPKIQKLLQDFFNGKELN 360
NP_006588.1 RFEELNADLFRGTLDPVEKALRDAKLDKSQIHDIVLVGGSTRIPKIQKLLQDFFNGKELN 360
            ************************************************************

NP_694881.1 KSINPDEAVAYGAAVQAAILSGDKSENVQDLLLLDVTPLSLGIETAGGVMTVLIKRNTTI 420
NP_006588.1 KSINPDEAVAYGAAVQAAILSGDKSENVQDLLLLDVTPLSLGIETAGGVMTVLIKRNTTI 420
            ************************************************************

NP_694881.1 PTKQTQTFTTYSDNQPGVLIQVYEGERAMTKDNNLLGKFELT---------------------- 462
NP_006588.1 PTKQTQTFTTYSDNQPGVLIQVYEGERAMTKDNNLLGKFELTGIPPAPRGVPQIEVTFDI 480
            ***********************************************

NP_694881.1 ------------------------------------------------------------
NP_006588.1 DANGILNVSAVDKSTGKENKITITNDKGRLSKEDIERMVQEAEKYKAEDEKQRDKVSSKN 540
NP_694881.1 ------------------------------------------------------------
NP_006588.1 SLESYAFNMKATVEDEKLQGKINDEDKQKILDKCNEIINWLDKNQTAEKEEFEHQQKELE 600
NP_694881.1 ---------------GMPGGMPGGFPGGGAPPSGGASSGPTIEEVD 493
NP_006588.1 KVCNPIITKLYQSAGGMPGGMPGGFPGGGAPPSGGASSGPTIEEVD 646
```

Alternate splice forms

Gene Symbol: OGT

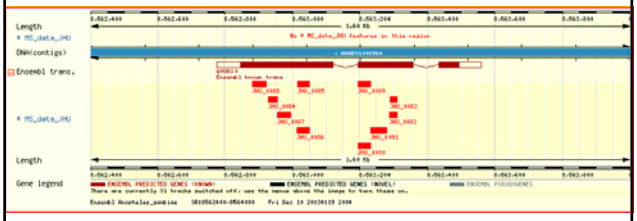
```
NP_858059 MASSVGNVADSTG----------LAELAHREYQAGDFEAAERHCMQLWRQEPDNTGVLLL 50
NP_858058 MASSVGNVADSTEPTKRMLSFQGLAELAHREYQAGDFEAAERHCMQLWRQEPDNTGVLLL 60
          ************           ************************************

NP_858059 LSSIHFQCRRLDRSAHFSTLAIKQNPLLAEAYSNLGNVYKERGQLQEAIEHYRHALRLKP 110
NP_858058 LSSIHFQCRRLDRSAHFSTLAIKQNPLLAEAYSNLGNVYKERGQLQEAIEHYRHALRLKP 120
          ************************************************************

NP_858059 DFIDGYINLAAALVAAGDMEGAVQAYVSALQYNPDLYCVRSDLGNLLKALGRLEEAKACY 170
NP_858058 DFIDGYINLAAALVAAGDMEGAVQAYVSALQYNPDLYCVRSDLGNLLKALGRLEEAKACY 180
          ************************************************************

NP_858059 LKAIETQPNFAVAWSNLGCVFNAQGEIWLAIHHFEKAVTLDPNFLDAYINLGNVLKEARI 230
NP_858058 LKAIETQPNFAVAWSNLGCVFNAQGEIWLAIHHFEKAVTLDPNFLDAYINLGNVLKEARI 240
          ************************************************************

                    ....

NP_858059 TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQRISSPLFNTKQYTMELERLY 1010
NP_858058 TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQRISSPLFNTKQYTMELERLY 1020
          ************************************************************

NP_858059 LQMWEHYAAGNKPDHMIKPVEVTESA 1036
NP_858058 LQMWEHYAAGNKPDHMIKPVEVTESA 1046
          *************************
```

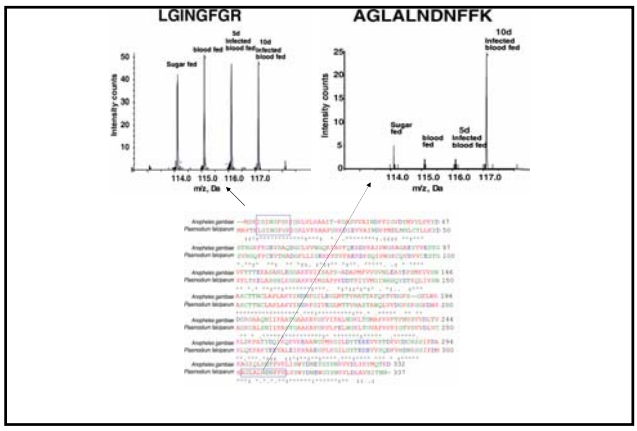
Genome Annotation

Genome Annotation by Mass Spectrometry: What Can We Gain?

- Assigning start codons
- Proteins isoforms (alternative splicing, novel exons)
- Novel genes (proteins less than 100 amino acids not predicted by programs)
- cSNPs
- Correction of incorrect gene predictions (~50% of the genes in human are predicted)
- Validation of gene predictions

Use of Ensembl Distributed Annotation System to Validate a Known Transcript




LGINGFGR
AGLALNDNFFK
Intensity counts
m/z, Da
Sugar fed
5d Infected blood fed
10d Infected blood fed
blood fed