Protein Databases for Mass Spectrometric Analysis

Akhilesh Pandey, M.D., Ph.D.

http://pandeylab.igm.jhmi.edu/

Human Genome Annotation

- A case for proteomics-driven annotation of protein-coding regions
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

6 reading frame translations of genome

MS/MS Spectra

Search

Peptide matches

Peptide matching a novel gene (Intergenic region)

Peptide matching a known exon

Peptide mapping onto the genome – Identifying a novel gene

Intergenic region

Gene 2

Gene 1

Proteins
Peptide mapping onto the genome – Identifying a novel exon

Genomic DNA

Genes

Peptide matching intronic region

Alternate splice forms

Gene

Protein isoform A

Protein isoform B

No Match!!
Gene Symbol: HSPA8

Alternate splice forms

**Gene Symbol:** HSPA8

**Alternate splice forms**

- **Protein isoform A**
- **Protein isoform B**
- **Hypothetical Protein isoform C**

**Peptide matches!!**
Alternate splice forms

Gene Symbol: OGT

```
NP_858059  |  HADVQVQVAGDSTG---------LAEIASSYQAGFPEAEHCHQGARQGSPNTVYLL  50
NP_858058  |  HADVQVQVAGDSTGTPSTMCLTQGALAEIASSYQAGFPEAEHCHQGARQGSPNTVYLL  60
NP_858059  |  LSINHVCQHLDLSAMSFSLAIPQGLALLATAYLMLNYYVQGQQQAEYBYHIALJAP  110
NP_858058  |  LASSIFSILQVGLALLAYLMLNYYVQGQQQAEYBYHIALJAP  120
NP_858059  |  DFIDGYINLAAALVAVGNMEGAVAFVSALQHQKCVDAELISALGKLEAKACH  170
NP_858058  |  DFIDGYINLAAALVAVGNMEGAVAFVSALQHQKCVDAELISALGKLEAKACH  180
NP_858059  |  TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQR  1010
NP_858058  |  TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQR  1020
NP_858059  |  LQMWEHYAAGNKPDHMIKPVEVTESA  1036
NP_858058  |  LQMWEHYAAGNKPDHMIKPVEVTESA  1046

---
```

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](image)
Assignment of the initiator methionine in a cDNA ‘fragment’ based on an N-terminal peptide

> KIAA0229 (1180 residues) FRAGMENT

SWGKGREGVVPAGLGGALPGDGKFSPRSRLGCSLGEVGQRVAALMGKEQ ELLRAARTGHLPAVEKLLSGBKRLSSGFGGGGGGGGGGGGGGGGGGGGGS SSSSHPLSSLLSMWRGPNVCVDSTGYTPLHHAALNGHHRSSSSRSQDSAEGQ DGVPEQFSGLHLHSSPVCEVGQDPFQQLLCTAGQSHPGDPQQGACHKASMQGLEETGVHAPGASQPSALDOQSKRVGYLTGLPTTNSRSHPETLHTASPHPGGAEEGDRSGAR
N-terminal Acetylated Peptide – Annotation of Start Codon

Alignment 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 zebrafish 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.

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

• 12,000 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

Contains

- GenBank CDS translations
- RefSeq Proteins
- Swiss-Prot
- Swiss-Prot Information Resource
- Protein Data Bank
- Protein Research Foundation

- 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


- 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

- 28,000 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**

- **GenBank** (Primary sequence submissions)
- **Collaborating groups** (FlyBase, SGD, NCI, OMIM etc)

  Computational analysis
  Manual curation

  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. The 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
IPI (International Protein Index) database

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

- A redundant database
- The sequence identifiers and sequence entries are not stable

- Has protein sequence information from Human, Mouse, Rat, Zebra fish and Arabidopsis species only
- 49,000 Human protein sequences
- A redundant database
- Has information on protein isoforms
- The sequence identifiers and sequence entries are not 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
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
Targeting of the e-Abl tyrosine kinase to mitochondria is an early cellular response to oxidative stress.

Yoon S, Short A, Bailey SC, Petrie D, Khera M, Simon S, Yaffe D.

Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

This study identified e-Abl tyrosine kinase as an early cellular response to oxidative stress. The presence of e-Abl in mitochondria is dependent on activation of protein kinase C and the e-Abl kinase domain. Targeting of e-Abl to mitochondria is associated with MEL-induced loss of mitochondrial membrane potential. The results also demonstrate that e-Abl is necessary for MEL-induced depletion of ATP and the induction of a pro-apoptotic cell death. These findings indicate that the e-Abl kinase is responsive to oxidative stress and thereby mediates mitochondrial dysfunction and cell death.
Status of HPRD

• Over 18,000 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
• 8,000 PTM sites annotated
• Over 24,000 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/)