Protein Functional Annotation

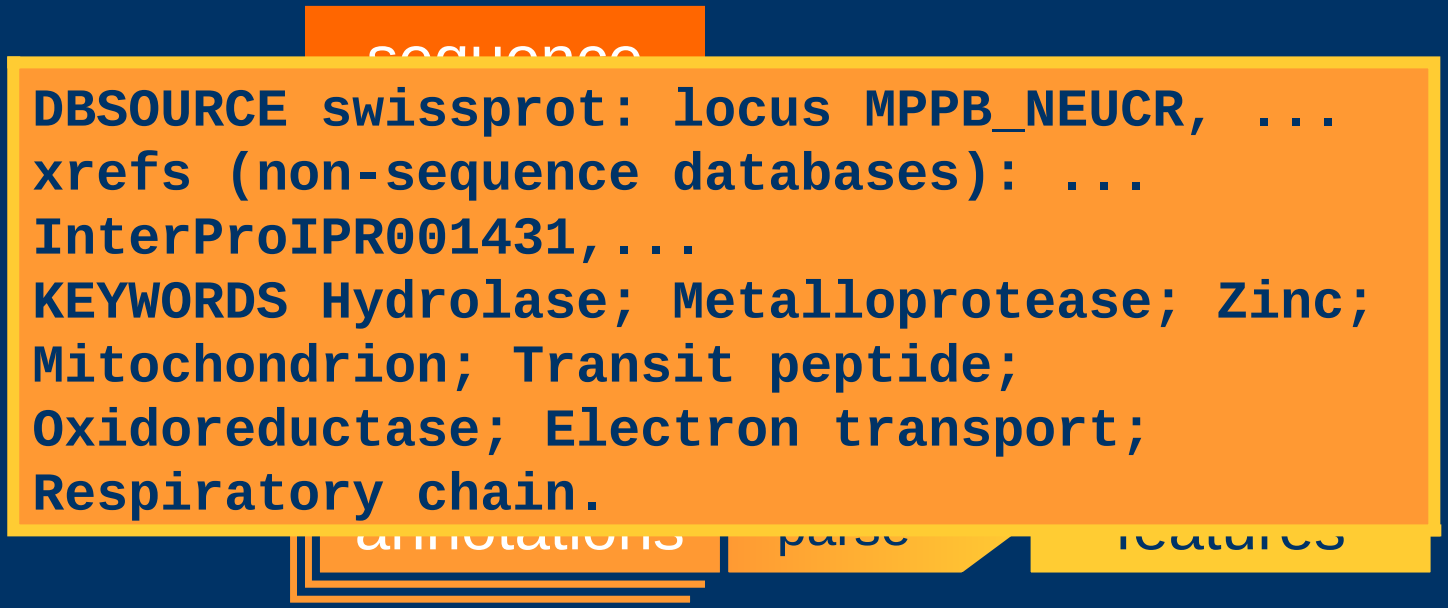
Dr G.P.S. Raghava

Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools
  - Annotation by Subcellular localization
  - requires computational tools for better subcellular localization prediction.

Annotation by Homology

- Statistically significant sequence matches identified by BLAST searches against GenBank (nr), SWISS-PROT, PIR, ProDom, BLOCKS, KEGG, WIT, Brenda, BIND
- Properties or annotation inferred by name, keywords, features, comments



Databases Are Key

Different Levels of Database Annotation

- GenBank (minimal annotation)
- PIR (slightly better annotation)
- SwissProt (even better annotation)
- Organsim-specific DB (best annotation)

Structure Databases

- **RCSB-PDB**
  - **http://www.rcsb.org/pdb/**
- **MSD**
  - **http://www.ebi.ac.uk/msd/index.html**
- **CATH**
  - **http://www.biochem.ucl.ac.uk/bsm/cath/**
- **SCOP**
  - **http://scop.mrc-lmb.cam.ac.uk/scop/**

Expression Databases

- Swiss 2D Page
  - http://ca.expasy.org/ch2d/
- SMD
  - http://genome-www5.stanford.edu/MicroArray/SMD/

Metabolism Databases

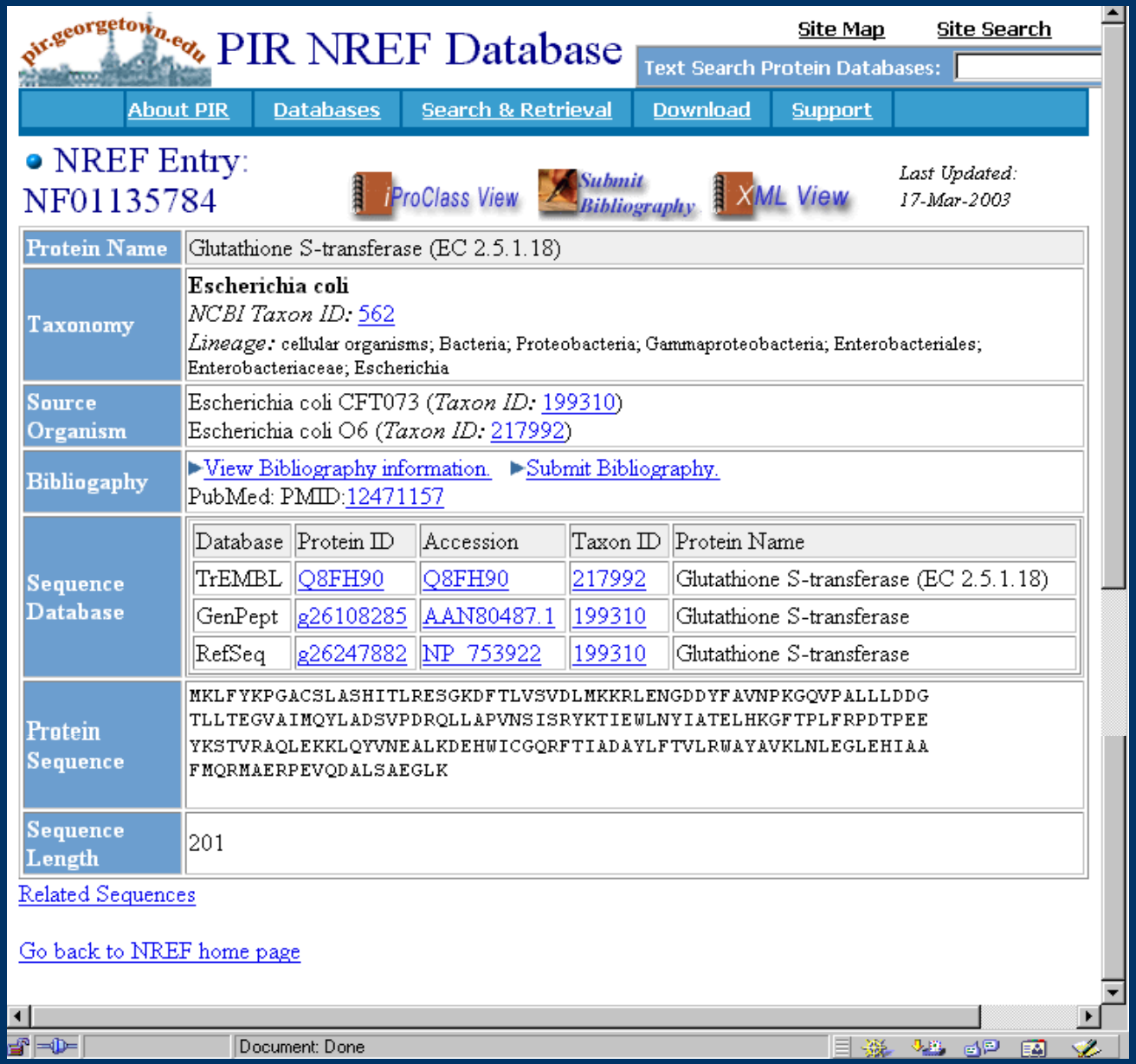
- KEGG
  - http://www.genome.ad.jp/kegg/metabolism.html
- EcoCyc
  - www.ecocyc.org/

Interaction Databases

- BIND
  - http://www.blueprint.org/bind/bind.php

GenBank Annotation

PIR Annotation



pir.georgetown.edu PIR NREF Database

Site Map Site Search

Text Search Protein Databases:

About PIR | Databases | Search & Retrieval | Download | Support

• NREF Entry: NF01135784

iProClass View | Submit Bibliography | XML View

Last Updated: 17-Mar-2003

Protein Name	Glutathione S-transferase (EC 2.5.1.18)
Taxonomy	**Escherichia coli** *NCBI Taxon ID:* 562 *Lineage:* cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia
Source Organism	Escherichia coli CFT073 (*Taxon ID:* 199310) Escherichia coli O6 (*Taxon ID:* 217992)
Bibliography	▶View Bibliography information. ▶Submit Bibliography. PubMed: PMID:12471157

Sequence Database

Database	Protein ID	Accession	Taxon ID	Protein Name
TrEMBL	Q8FH90	Q8FH90	217992	Glutathione S-transferase (EC 2.5.1.18)
GenPept	g26108285	AAN80487.1	199310	Glutathione S-transferase
RefSeq	g26247882	NP_753922	199310	Glutathione S-transferase

Protein Sequence

```
MKLFYKPGACSLASHITLRESGKDFTLVSVDLMKKRLENGDDYFAVNPKGQVPALLLDDG
TLLTEGVAIMQYLADSVPDRQLLAPVNSISRYKTIEWLNYIATELHKGFTPLFRPDTPEE
YKSTVRAQLEKKLQYVNEALKDEHWICGQRFTIADAYLFTVLRWAYAVKLNLEGLEHIAA
FMQRMAERPEVQDALSAEGLK
```

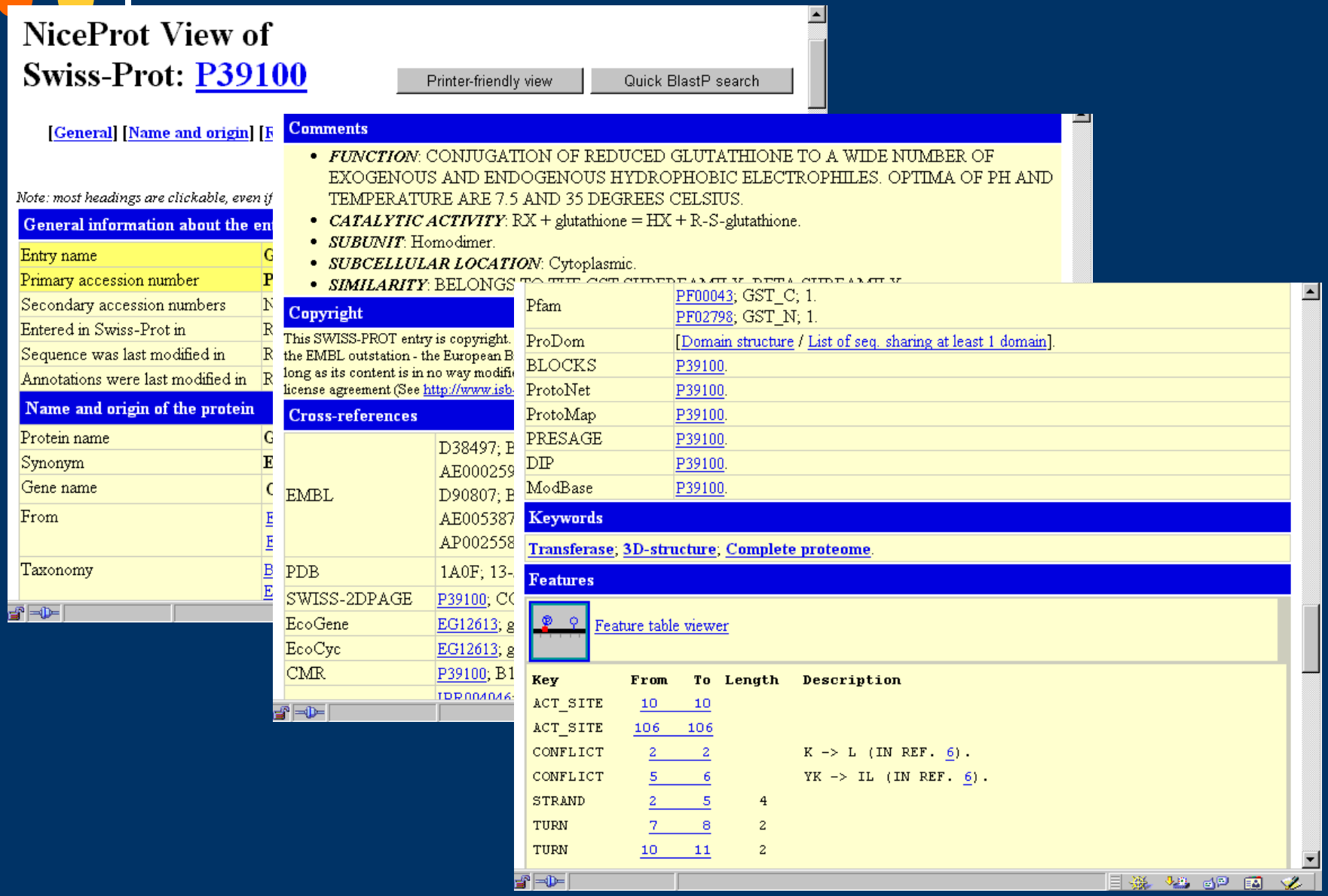
Sequence Length 201

Related Sequences

Go back to NREF home page

Document: Done

Swiss-Prot Annotation



NiceProt View of
Swiss-Prot: P39100

Printer-friendly view | Quick BlastP search

[General] [Name and origin] [F

Note: most headings are clickable, even if

General information about the en

Entry name	G
Primary accession number	P
Secondary accession numbers	N
Entered in Swiss-Prot in	R
Sequence was last modified in	R
Annotations were last modified in	R

Name and origin of the protein

Protein name	G
Synonym	E
Gene name	C
From	E E
Taxonomy	B E

Comments

- ***FUNCTION***: CONJUGATION OF REDUCED GLUTATHIONE TO A WIDE NUMBER OF EXOGENOUS AND ENDOGENOUS HYDROPHOBIC ELECTROPHILES. OPTIMA OF PH AND TEMPERATURE ARE 7.5 AND 35 DEGREES CELSIUS.
- ***CATALYTIC ACTIVITY***: RX + glutathione = HX + R-S-glutathione.
- ***SUBUNIT***: Homodimer.
- ***SUBCELLULAR LOCATION***: Cytoplasmic.
- ***SIMILARITY***: BELONGS TO THE GST SUPERFAMILY, BETA SUBFAMILY.

Copyright

This SWISS-PROT entry is copyright. the EMBL outstation - the European B long as its content is in no way modifi license agreement (See http://www.isb

Cross-references

EMBL	D38497; E AE000259 D90807; E AE005387 AP002558
PDB	1A0F; 13-
SWISS-2DPAGE	P39100; C
EcoGene	EG12613; g
EcoCyc	EG12613; g
CMR	P39100; B1

Pfam	PF00043; GST_C; 1. PF02798; GST_N; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
BLOCKS	P39100.
ProtoNet	P39100.
ProtoMap	P39100.
PRESAGE	P39100.
DIP	P39100.
ModBase	P39100.

Keywords

Transferase; 3D-structure; Complete proteome.

Features

Feature table viewer

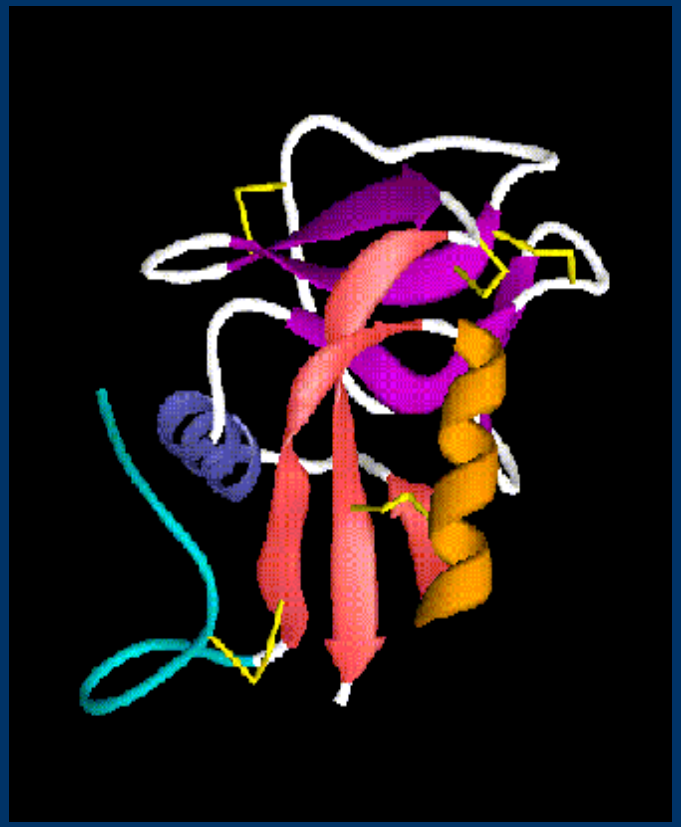
Key	From	To	Length	Description
ACT_SITE	10	10		
ACT_SITE	106	106		
CONFLICT	2	2		K -> L (IN REF. 6).
CONFLICT	5	6		YK -> IL (IN REF. 6).
STRAND	2	5	4	
TURN	7	8	2	
TURN	10	11	2	

Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools
- Annotation by Subcellular localization
  - requires computational tools for better subcellular localization prediction.

Annotation by Composition

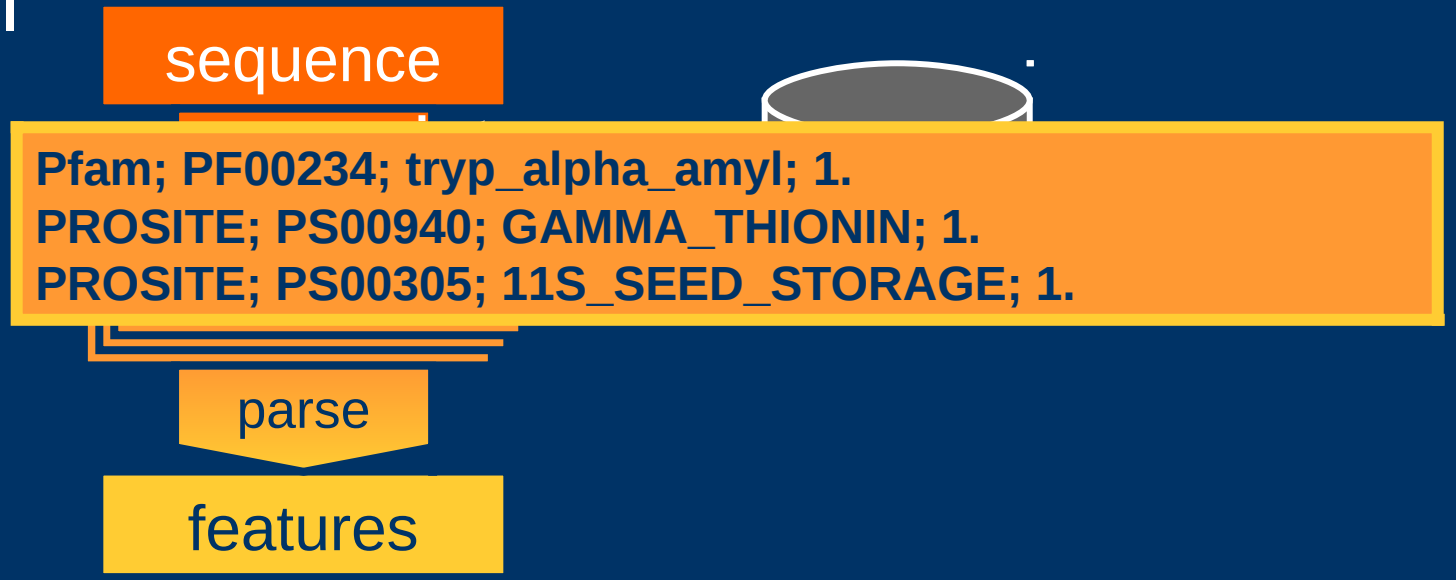
- Molecular Weight
- Isoelectric Point
- UV Absorptivity
- Hydrophobicity



Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools
- Annotation by Subcellular localization
  - requires computational tools for better subcellular localization prediction.

Feature based annotation



- PROSITE - http://www.expasy.ch/
- BLOCKS - http://blocks.fhcrc.org/
- DOMO - http://www.infobiogen.fr/services/domo/
- PFAM - http://pfam.wustl.edu
- PRINTS - http://www.biochem.ucl.ac.uk/bsm/dbrowser/PRINTS
- SEQSITE - PepTool

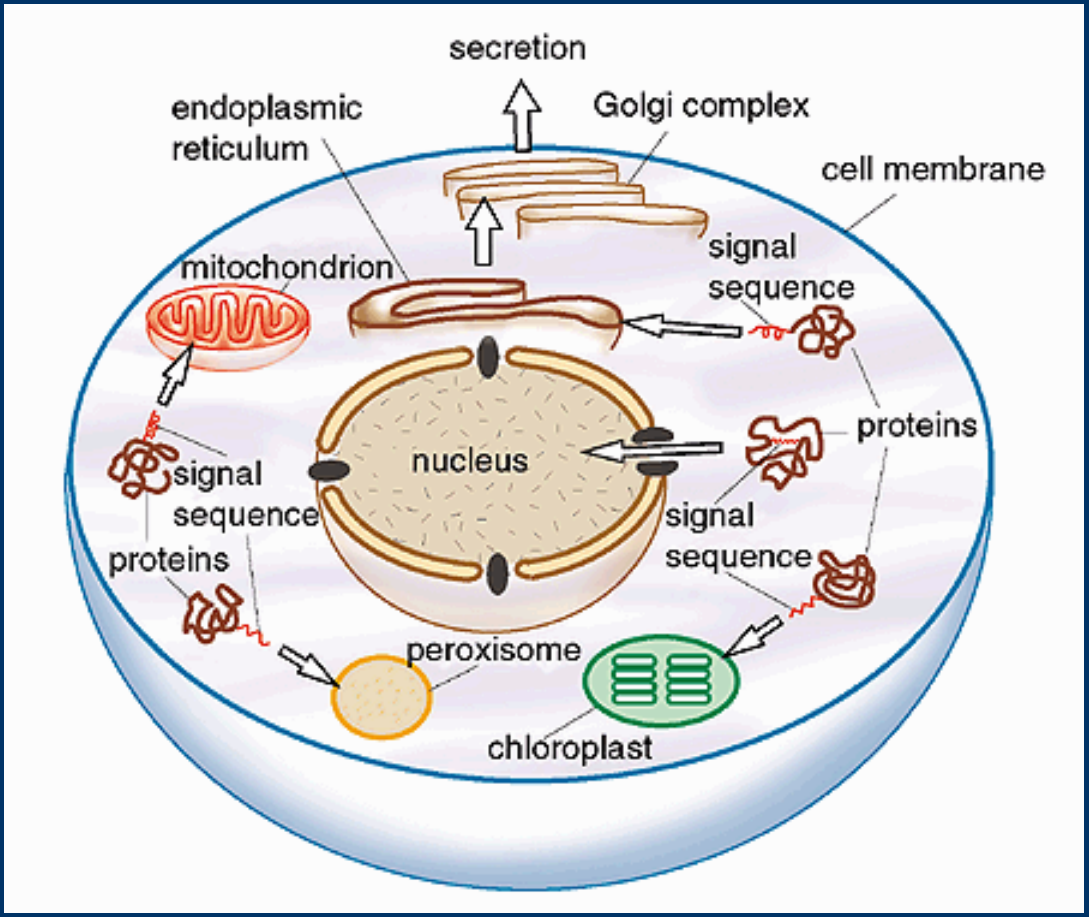
Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods

- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools

- Annotation by Subcellular localization
  - requires computational tools for better subcellular localization prediction.

- What is Subcellular Localization?
  - Organelles
  - Membranes
  - Compartments
  - Micro-environments



Gene Ontology

Cellular component contains organelles, membranes, cell regions, localized and unlocalized protein complexes

- Subcellular Localization Ontology
  - Cellular Components can be instantiated
  - Captures spatial relationships
  - Maps to GO concepts
  - Uses EcoCyc concepts: Macromolecule, Reaction, Pathway

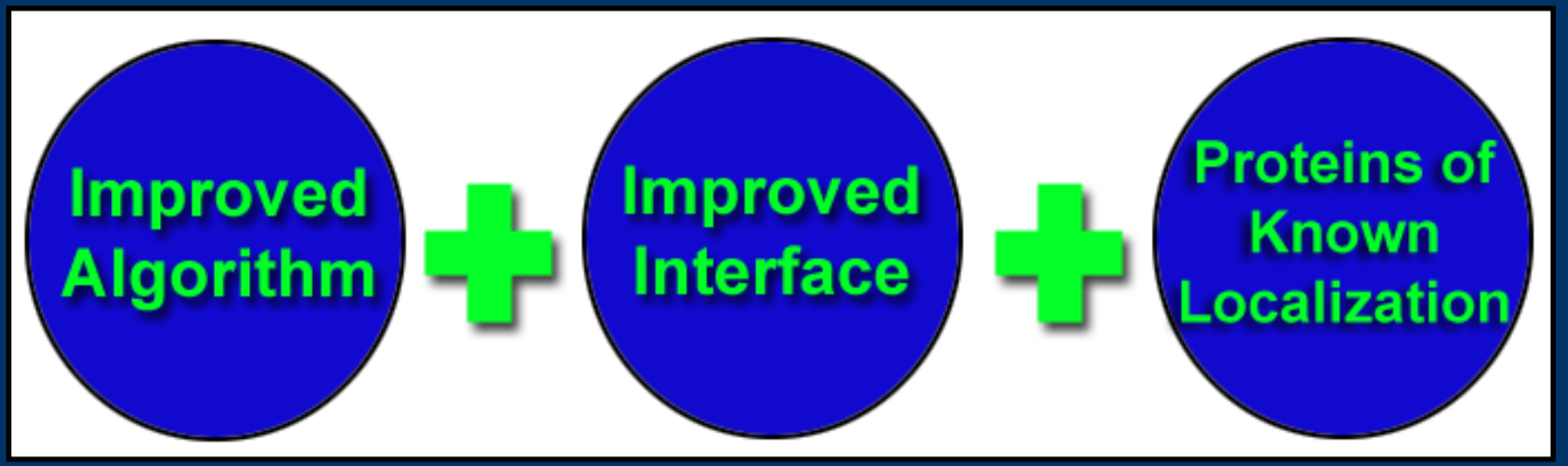
Why is Subcellular Localization Important?

- Function is dependent on context
- Localization is dynamic and changing
- Compartmentalization forms groups which allows for abstraction of concepts (i.e. mitochondria)

Specifying Subcellular Localization: Why is it difficult?

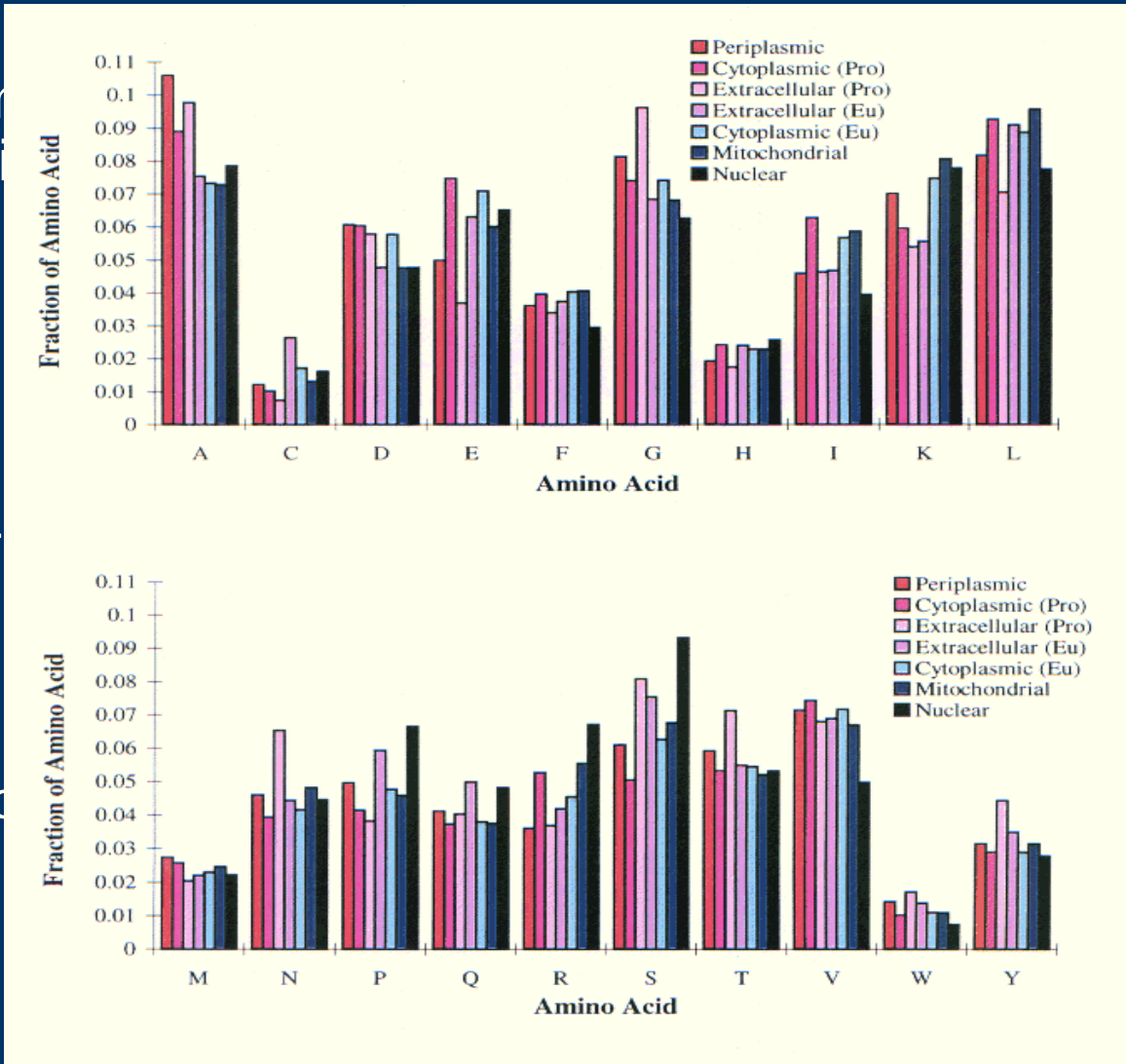
- **Biological Context**
- **Hard to define boundaries**
- **Dynamic Systems**
- **Distributions of proteins**

Our solution: Code Name bPSORT



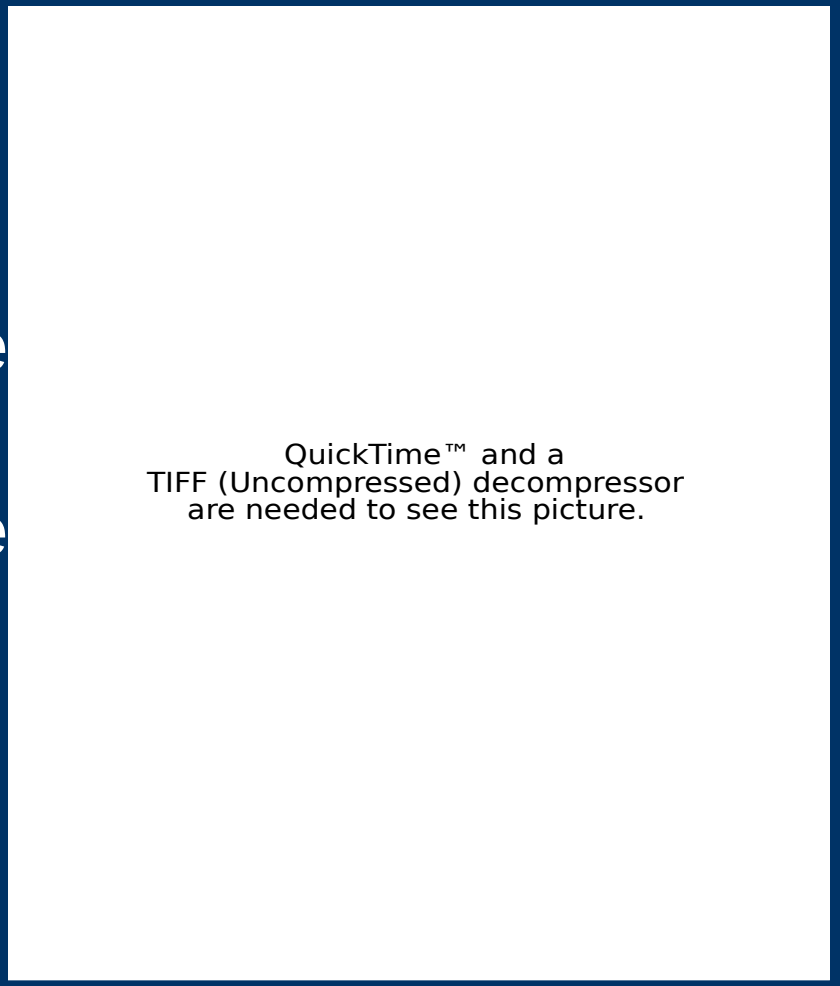
Method 1 - Amino acid composition

- Correlate amin acid compositi to subcellular location
  - Alanine - periplasm
  - Glycine - extracellular
  - Serine - nucleus
  - Leucine - mitochondri



Method 2 - Find signal sequences

- Short stretches of amino acids
- Located at either end of the protein
- Sometime in the middle of the sequence



Protein translocation across the ER membrane

- The signal peptide binds to the SRP
- The SRP complex docks on the channel
- The signal peptide is cleaved and the protein is secreted out of the cell

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Protein translocation across the ER membrane

- The signal peptide binds to the SRP
- The SRP complex docks on the channel
- The signal peptide is cleaved and the protein is secreted out of the cell

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Method 3 - Combine homologs and NLP

- A lot of well annotated databases
- Sequence alignment to find homologs
- Extract important texts (features) from homologs
- Analyze texts (features) using NLP techniques
- Training predictors based on those features
- Make predictions to new sequences using pre-built predictors

Method 4 - Integrative system

- Composed of separate modules
  - Motif analysis
  - Signal peptide detection
  - Transmembrane domains
- Each predicts one particular location
- Integrate modules by either rule-based system or probabilistic model

Future directions

- Is cross validation convincing?
- Golden datasets for fair evaluation
- Is the prediction result obvious to the users?
  - Transparency
- Is one location per protein enough?
  - Protein transport

Current prediction methods

- Eukaryotic localization predictors:

  TargetP (Emanuelsson et al, 2001)
  iPSORT (Bannai et al, 2001)
  PSORT II (Horton, P. and Nakai, 1997)
  ESLPred (Bhasin and Raghava, 2004)

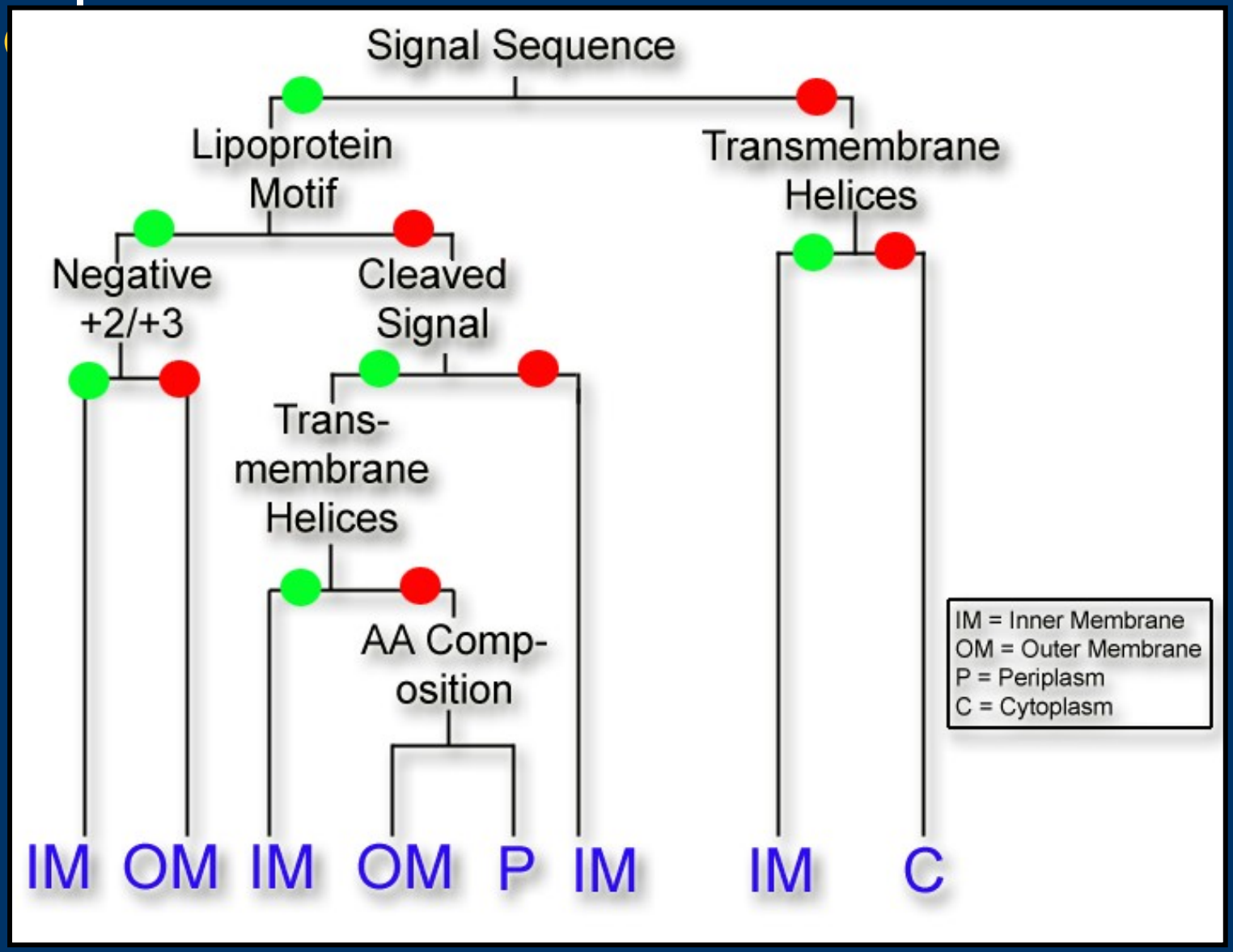
- Prokaryotic localization predictors:

  PSORT I (Nakai, K. and Kanehisa, 1991)
  PSORT-B (Nakai, K. and Kanehisa, 2001)

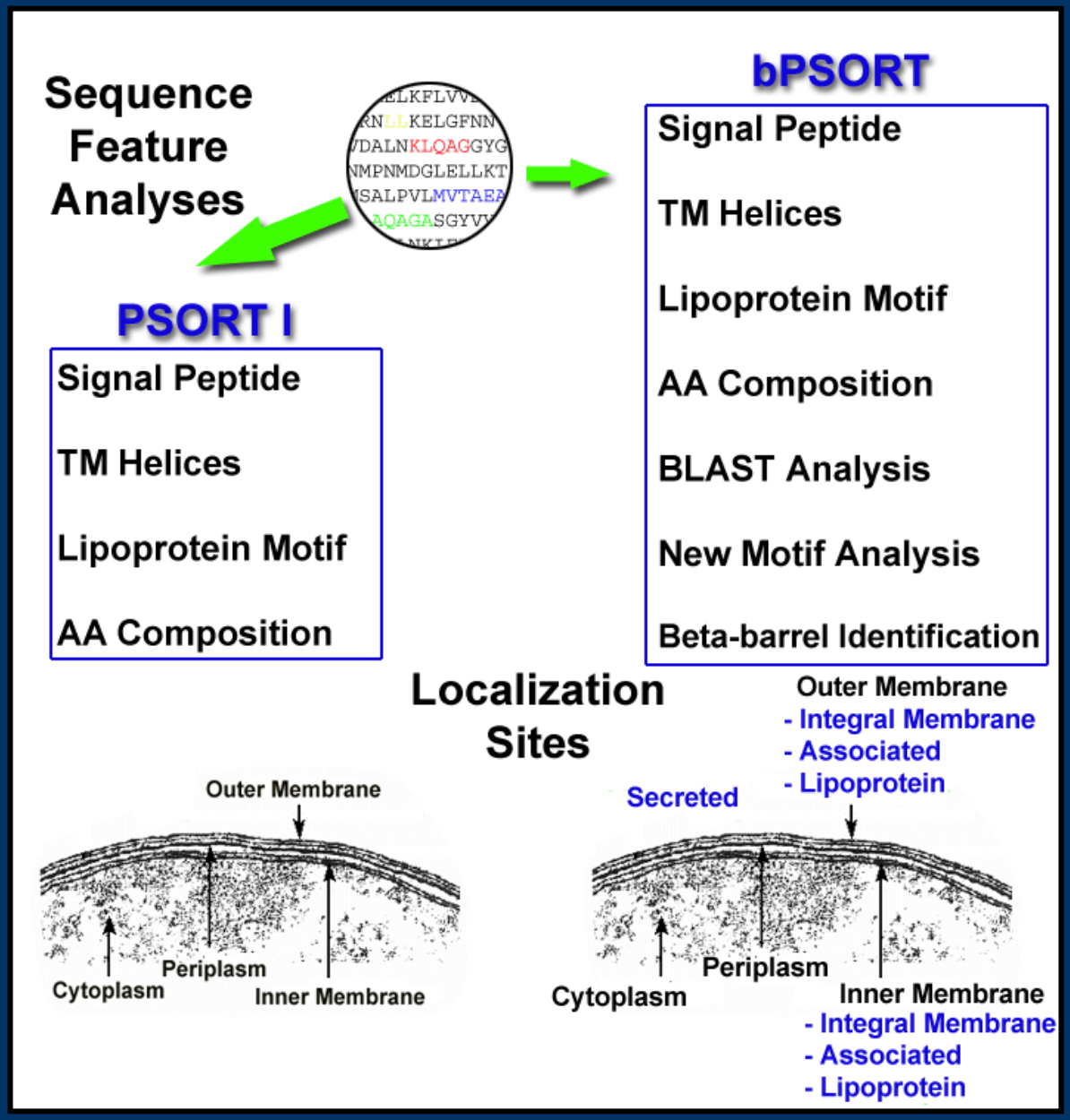
- Eukaryotic and Prokaryotic localization predictors:

  NNPSL (Reinhardt and Hubbard, 1998)
  SubLoc (Hau and Sun, 2001)

Current prediction methods – PSORT I



Improved Prediction Algorithm



NNPSL

- Use AA composition
- Use neural networks
- Prokaryotic
  - periplasm, cytoplasm and extracellular
  - 81%
- Eukaryotic
  - Extracellular, cytoplasm, mitochondrion and nuclear
  - 66%

SubLoc

- Use AA composition
- Use SVMs instead of NN in NNPSL
- Same datasets
- Different results
  - 91.4% prokaryotic
  - 79.4% eukaryotic

SignalP

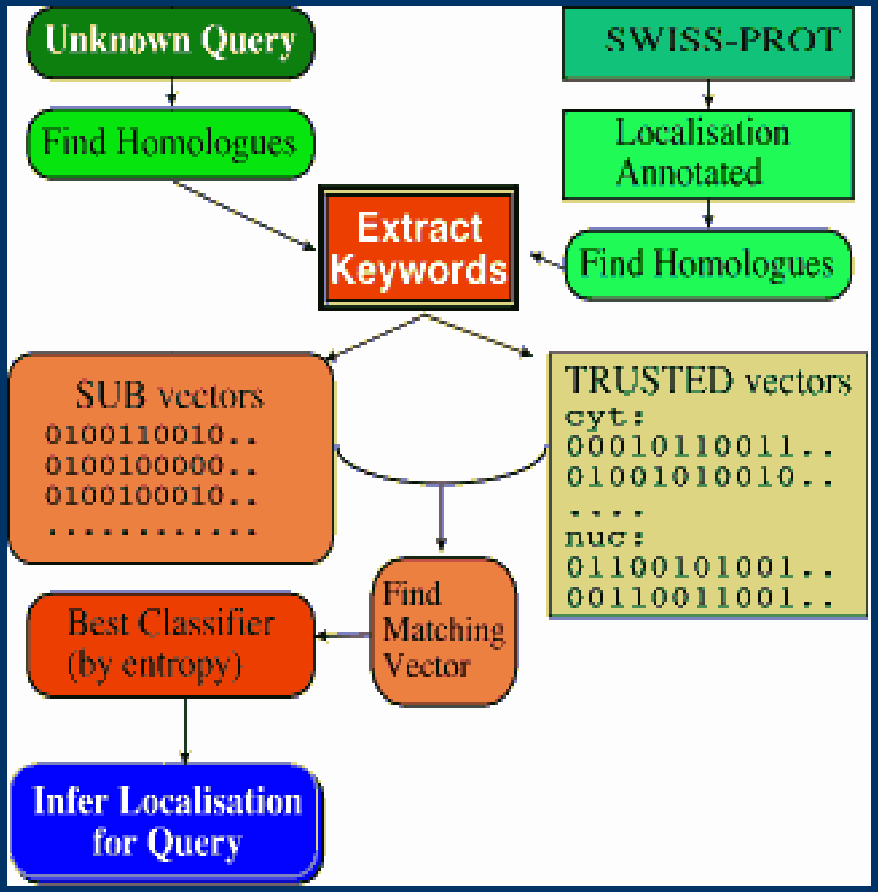
- Current version 3.0 predicts the presence and location of signal peptide cleavage sites
- Based on neural networks in its first version v1.0
- Developed SignalP-HMM in Version 2.0
- Eukaryotic and Gram positive and negative bacteria

TargetP

- Predicts the subcellular location of eukaryotic protein sequences
- Based on the predicted presence of any of the N-terminal short sequences
  - Chloroplast transit peptide (cTP)
  - Mitochondrial targeting peptide (mTP)
  - Secretory pathway signal peptide (SP)

LOCKey

- Look for proteins with known localization in Swiss-Prot
- Construct trusted vectors using reliable homologs
- Match SUB vectors to trusted vectors to make new predictions



Improved Method for Subcellular localization of Eukaryotic protein is Required

The PSORT-B is highly accurate method for prediction of subcellular localization of prokaryotic protein. The subcellular localization of eukaryotic proteins is not so accurate due to complexity of proteins.

A highly accurate method for subcellular localization of eukaryotic proteins is of immense importance for better functional annotation of protein.

Attempt for better prediction of subcellular localization of eukaryotic protein

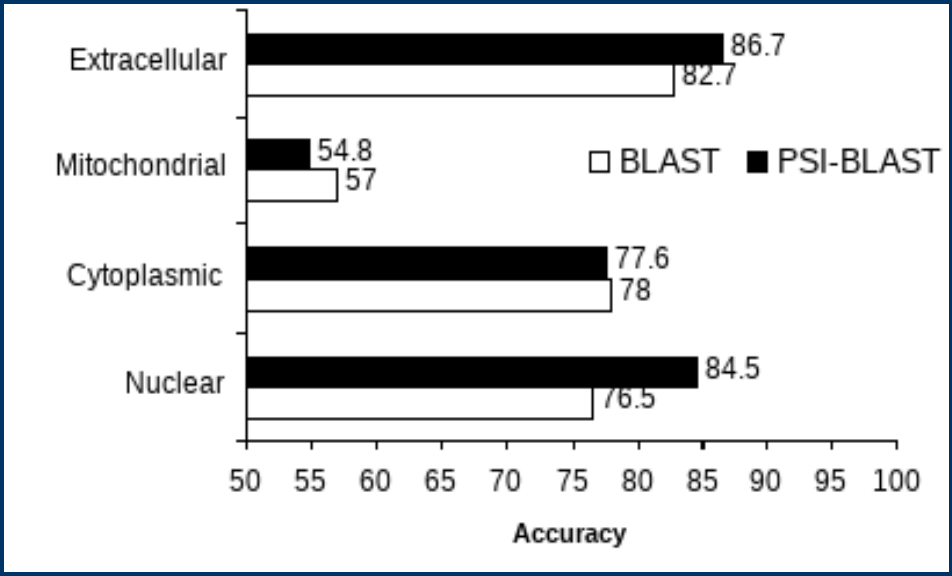
Dataset for classification

- Experimentally proven proteins .
- Complete protein
- Non-redundant proteins(90)

2427

Similarity based prediction of subcellular localization

We have generated a BLAST as well as with PSIBLAST based module for subcellular localization. It performance was evaluated by 5-fold cross-validation.



It proves that PSIBLAST is better as compared to simple BLAST in subcellular localization prediction. Out of 2427 proteins whereas it was only 362 proteins for which no significant hit was found in case of PSI-BLAST.

AS it is proved in past that machine learning techniques are elegant in classifying the biological data. Hau and sun applied the SVM for classification of prokaryotic and eukaryotic proteins and shown that it is better than statistical methods as well as other machine learning techniques such as ANN. For class classification four 1-v-r SVMs were used.

REq: Fixed Pattern length

Q: How to convert the variable length of proteins to fixed length?

Ans: Amino acid composition is most widely used for this purpose.

$$\text{fraction of } aa_i = \frac{\text{total number of amino acid i}}{\text{total number of amino acids of protein}}$$

where i can be any amino acid out of 20 natural amino acids.

RESULTS

Approach	Nuclear		Cytoplasmic		Mitochondrial		Extracellular	
	ACC	**MCC**	**ACC**	**MCC**	**ACC**	**MCC**	**ACC**	**MCC**
Composition based	86.1	0.73	76.9	0.64	55.5	0.54	76.0	0.76

Amino acid properties based prediction:

We have taken in consideration 33 physic-chemical properties for classification like hydrophobocity,hydrphilicity.

Approach	Nuclear		Cytoplasmic		Mitochondrial		Extracellular	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Properties Based	85.6	0.73	74.6	0.64	59.2	0.55	76.6	0.74

The physico-chemical properties-based SVM module predicted subcellular localization of protein with slightly lower accuracy (77.8%) than the amino acid composition based module.

What is lacking in amino acid composition and properties based classification ?

Both of the properties provide information about the fraction of residues of particular type and lack information about residue order.

So property that provide information

amino acid composition + Order =More accurate prediction ?

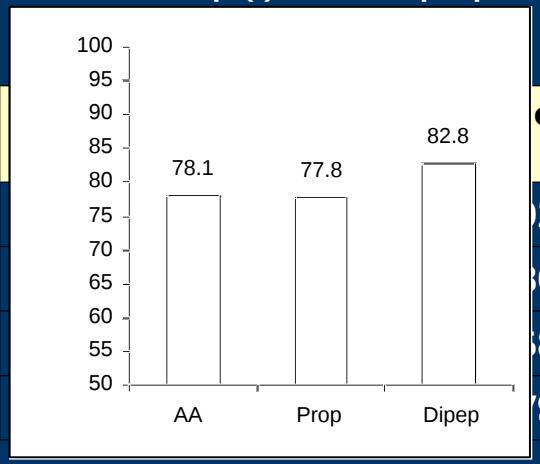
Dipeptide
Tripeptide
Tetrapeptide……

So we have used dipeptide composition for subcellular localization prediction.

$$\text{fraction of dep (i)} = \frac{\text{total number of dep(i)}}{\text{total number of all possible dipeptides}}$$

Where dep(i) is a dipeptide i out of 400 dipeptides.

Result



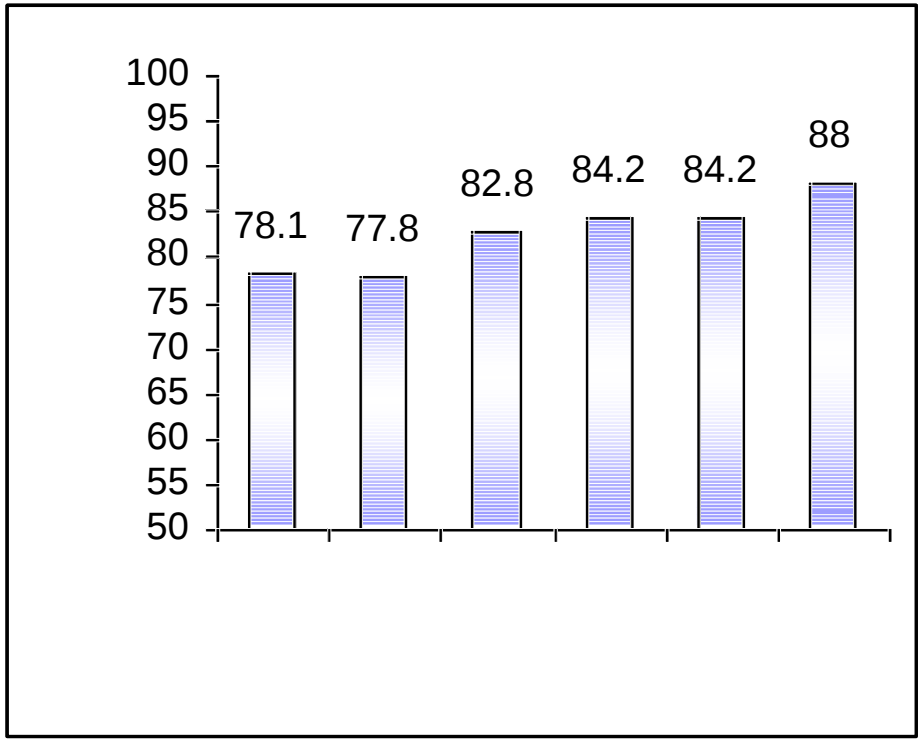
cy	
2.7	
0.2	
8.8	
79.0	0.83

It proves that dipeptide composition is better than aa composition and properties composition.

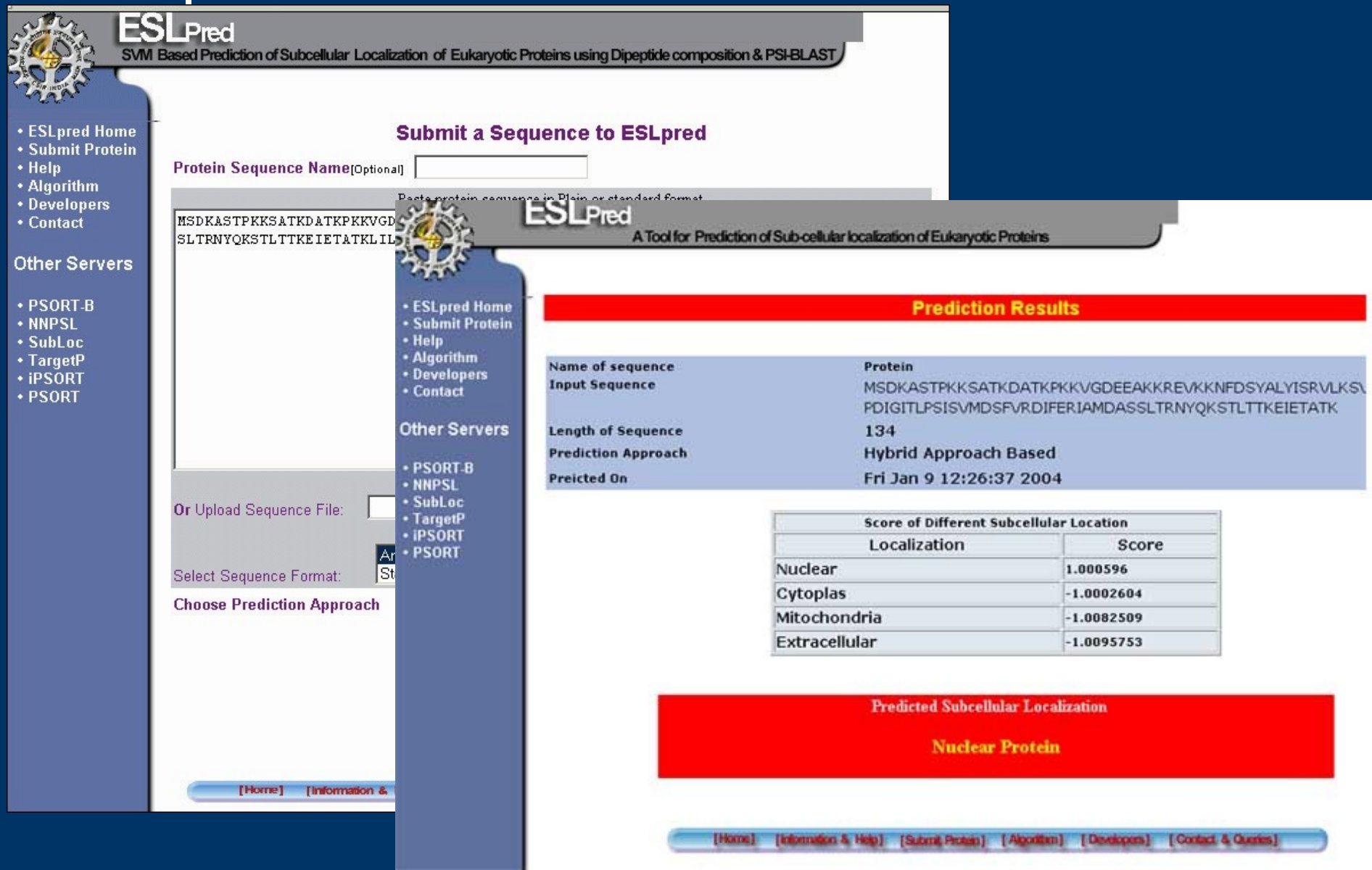
To further improve the accuracy:

Tripeptide composition: SVM fails to train due to complexity of patterns. The pattern of each protein is of 8,000 vectors with lot of noise.

Hybrid approach: Using more then one feature .



Method based on above results have been implemented online as **ESLPRED**



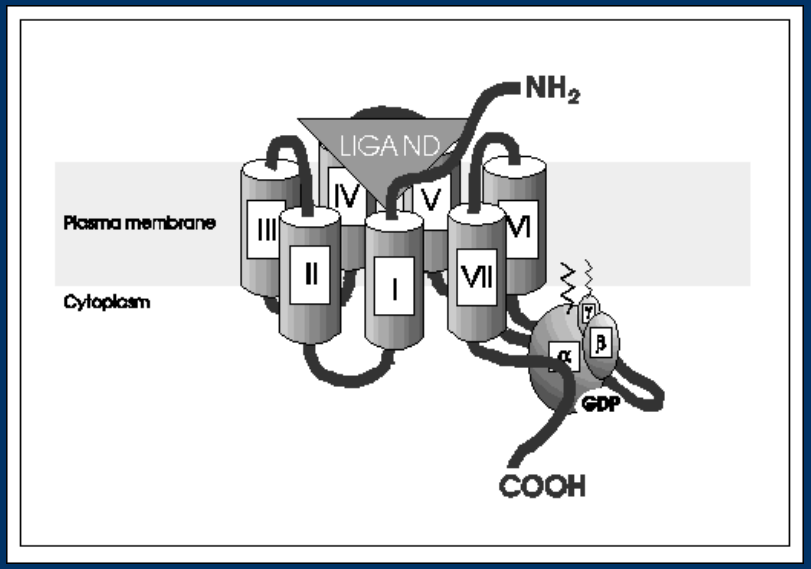
Functional annotation:Classification

G-Protein Coupled Receptor

- Membrane-bound receptors

- Transducing messages as photons, organic odorants, nucleotides, nucleosides, peptides, lipids and proteins.

- 6 different families
- A very large number of different domains both to bind their ligand and to activate G proteins.



More than 50% of drugs in the market are base on GPCRs due to their major role in signal transduction.

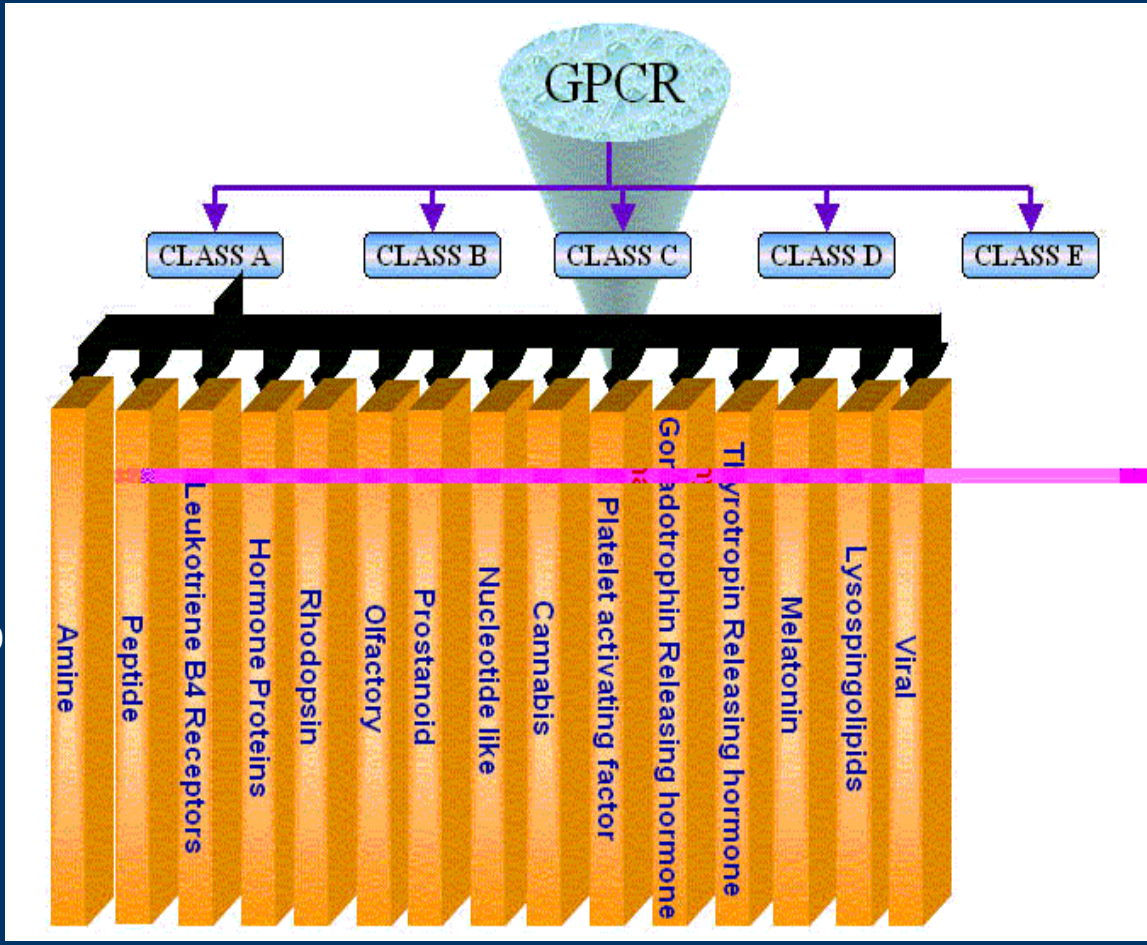
Classoification of GPCRs

Up to Subfamilies (Part 1)

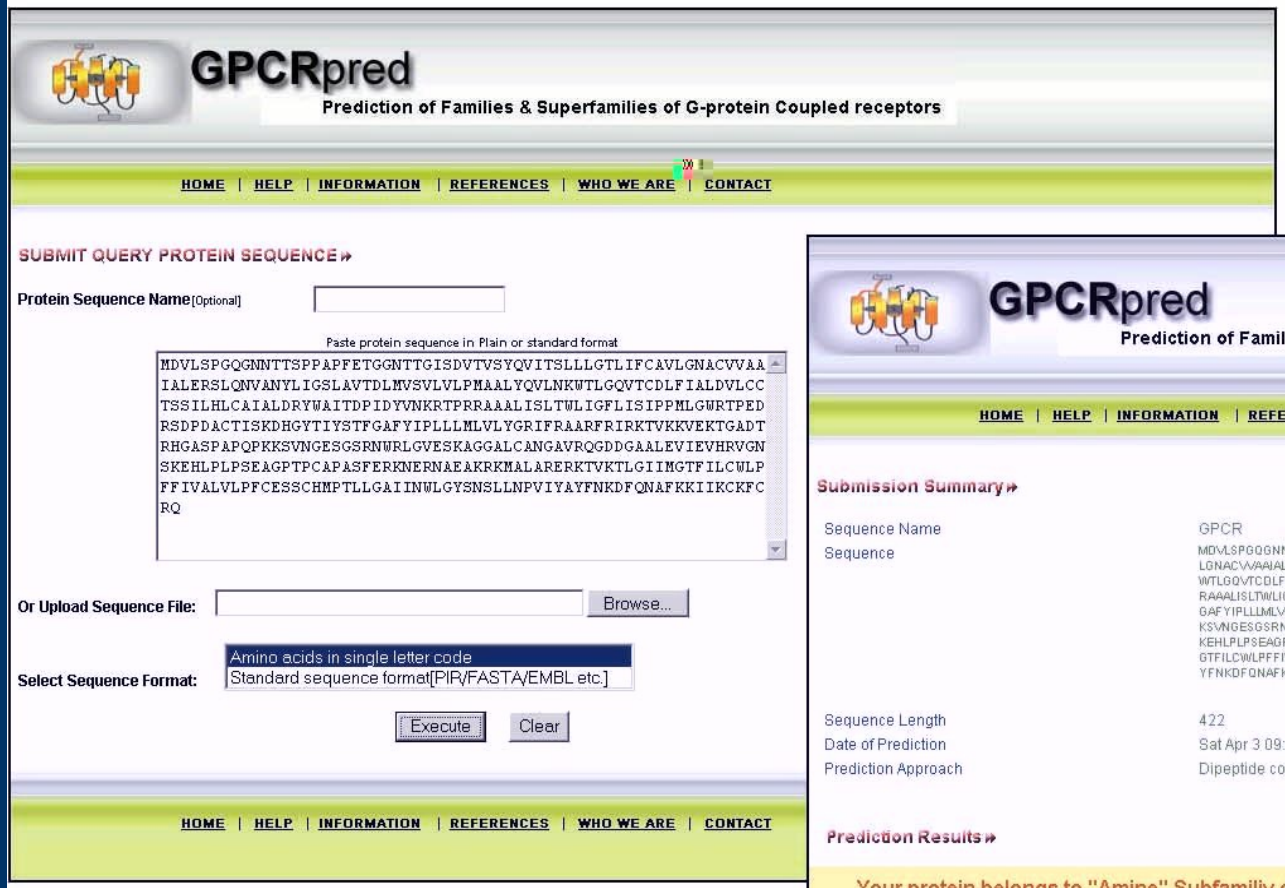
Types of Receptors of each subfamily (II)

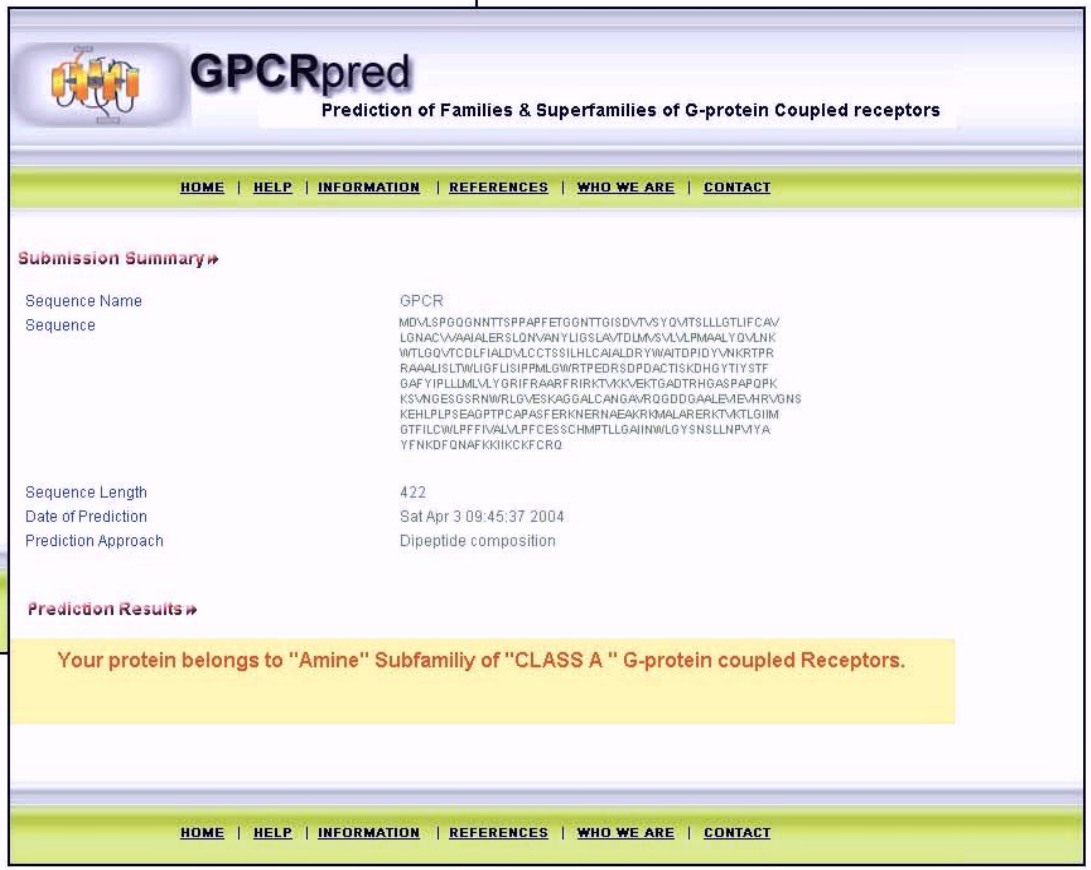
History :

- Mostly BLAST is used for the classification and recognition of novel GPCRs.
- Motifs search is also used as GPCRs have conserved structure.
- One SVM based method is also available for classification of GPCRs of Rhodopsin family. (Karchin et al., 2001)



Method based on SVM using aa and dipep composition has been implemented online as GPCRPred





Functional annotation:Classification

Nuclear Receptor

- Nuclear receptors are the key transcription factors that regulate crucial gene networks responsible for cell growth, differentiation and homeostasis.
- Potential drug targets for developing therapeutic strategies for diseases like cancer and diabetes.
- classified into seven subfamilies
- Consist of six distinct regions.