

Predict Subcellular Localization for Proteins in all Kingdoms

Tatyana Goldberg

Supervisor: Burkhard Rost

Advisor: Tobias Hamp

4.04.2011

Motivation

- NCBI RefSeq: more than 12Mio. protein sequences from 11.700 genomes (March 11, 2011)
- Swiss-Prot: 526.000 entries (March 08, 2011)
=> Big sequence to function gap
- Need for reliable automatic predictions of protein function
- Gene Ontology (GO): hierarchical classification of functions
- Cellular component is one of the three main classes used to organize protein function within the GO

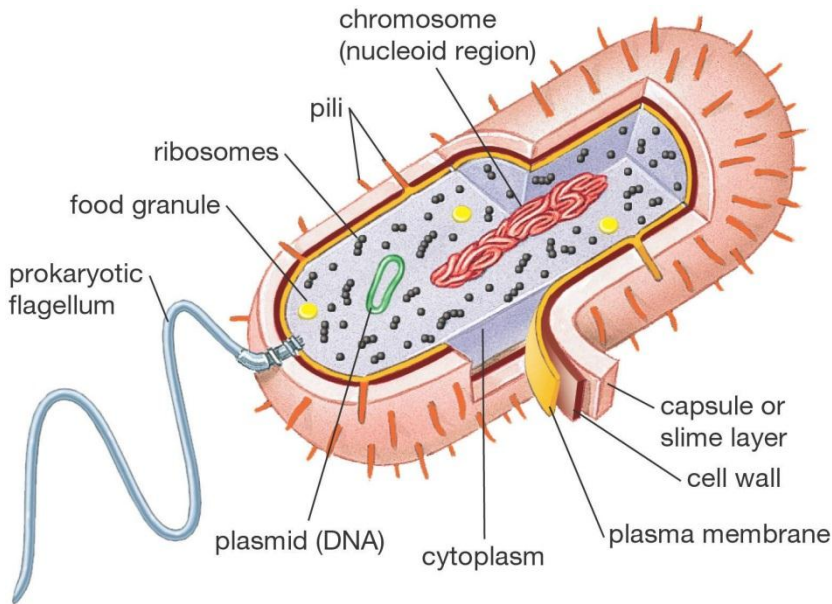
Pruitt K., Maglott D., et al. (2007) *Nucleic Acids Res*

Bairoch A., Apweiler R. (1997) *J. Mol. Med.*

Ashburner M., Sherlock G., et al. (2000) *Nature Genetics*

Cellular Compartments

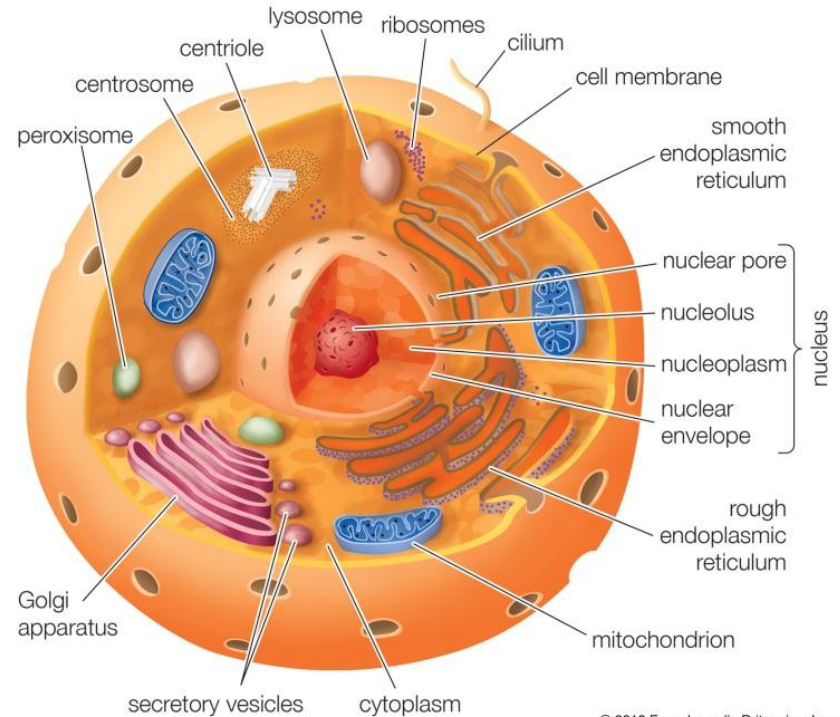
Prokaryotic Cell



Copyright © 2005 Pearson Prentice Hall, Inc.

<http://mwsu-bio101.ning.com>

Eukaryotic Cell

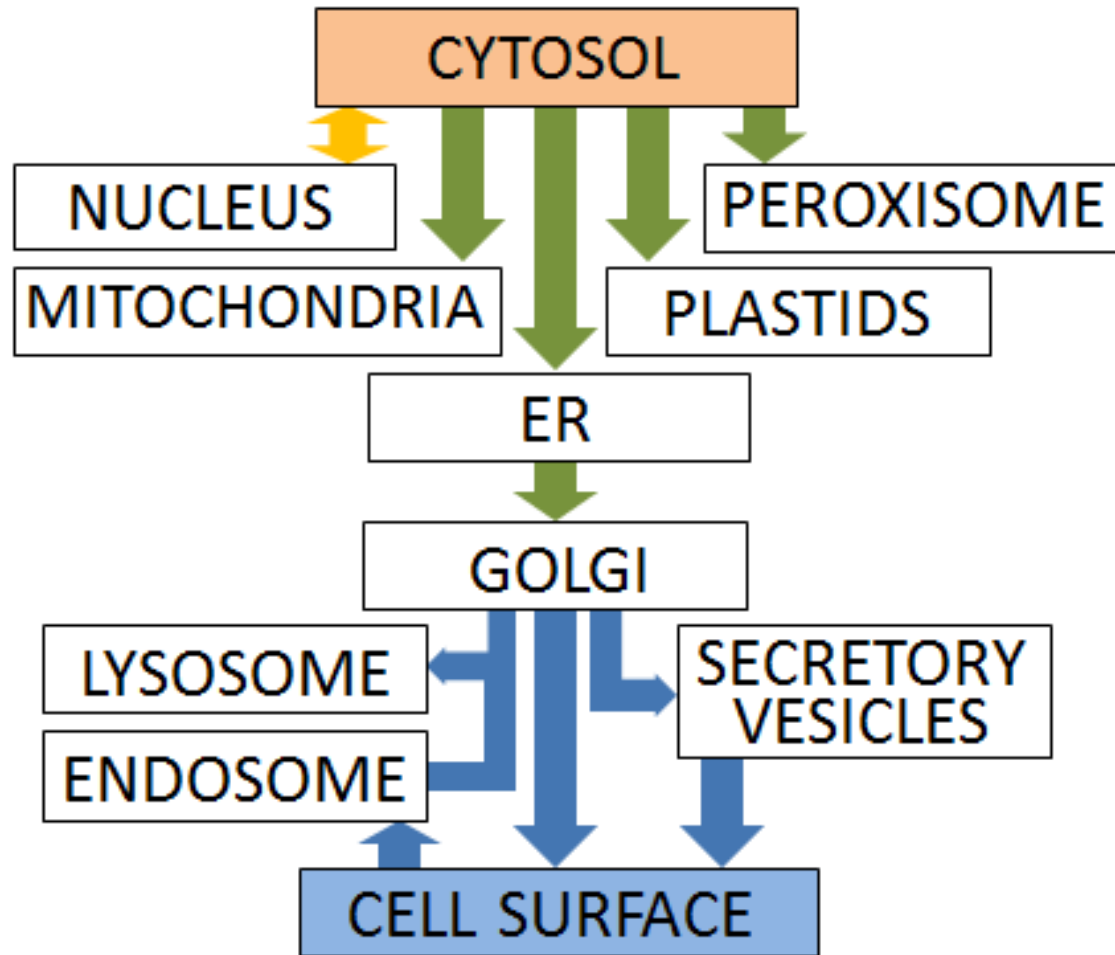


© 2010 Encyclopædia Britannica, Inc.




<http://www.britannica.com>

Common compartment → Common physiological function

Trafficking in the Cell



Key:

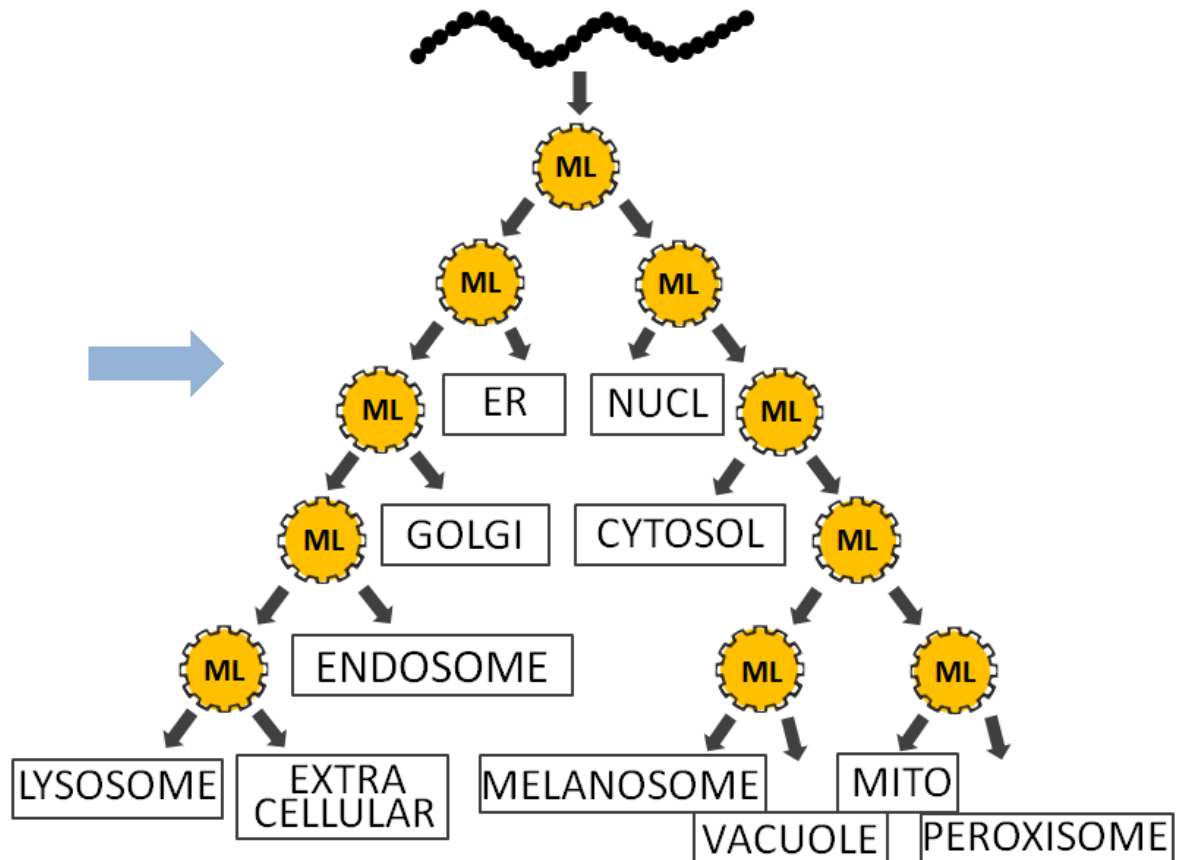
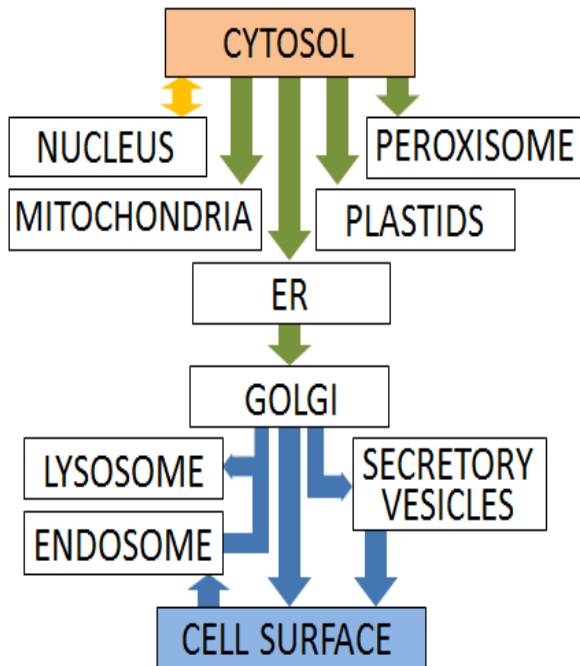
-  gated transport
-  transmembrane transport
-  vesicular transport

Existing Prediction Methods Are Based On...

- Sorting Signals
 - signalP** (Nielsen H, von Heijne G., et al. 1997 *Protein Eng.*),
 - chloroP** (Emanuelsson O., von Heijne G., et al. 1999 *Protein Sci.*),
 - PSORT** (Nakai, K. and Horton P. 1999 *Trends Biochem Sci*)
- Sequence Homology
 - Nair R. and Rost B. 2002 *Protein Sci*
- Text Analysis
 - LOCkey** (Nair R. and Rost B. 2002 *Bioinformatics*)
- *Ab initio* methods: amino acid composition & structural and evolutionary information
 - SubLoc** (Hua S., Sun Z. 2001 *Bioinformatics*),
 - CELLO** (Yu J., Lemmon M., et al. 2004 *Mol. Cell*),
 - LOCtree** (Nair R., Rost B. 2005 *JMB*),

Our Approach

TM Protein prediction using TMHMM
(Krogh A., Sonnhammer E., et al. 2001 *J Mol Biol*)



Data Set

UniProt > UniProtKB

Search Blast * Align Retrieve ID Mapping *

Search in Protein Knowledgebase (UniProtKB) Query Search

P00846 (ATP6_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot
Last modified March 8, 2011. Version 121. [History...](#)

[General annotation \(Comments\)](#)

Function	Mitochondrial membrane ATP synthase (F ₁ F ₀ ATP synthase or Complex I) domains, F ₁ - containing the extramembraneous catalytic core and F ₀ - subunits to proton translocation. Key component of the proton channel
Subunit structure	F-type ATPases have 2 components, CF ₁ - the catalytic core - and CF ₀
Subcellular location	Mitochondrion inner membrane; Multi-pass membrane protein.

- Exclusion of non-experimental proteins and proteins with multiple localizations
- Redundancy reduction at
HVAL=0 and BLAST E -value $\leq 10^{-3}$

Data Set

	Archaea		Bacteria		Fungi		Plants		Animals	
	Soluble	TMP	Soluble	TMP	Soluble	TMP	Soluble	TMP	Soluble	TMP
Secreted	✓		✓		✓		✓		✓	
Cell Membrane		✓		✓		✓		✓		✓
Cytoplasm	✓		✓		✓		✓		✓	
Nucleus					✓	✓	✓	✓	✓	✓
ER					✓	✓	✓	✓	✓	✓
Golgi					✓	✓	✗	✓	✓	✓
Endosome					✓	✓	✗	✗	✓	✓
Lysosome									✓	✓
Peroxisome					✓	✓	✓	✓	✓	✓
Melanosome									✓	✓
Vacuole					✗	✓	✓	✓	✓	✓
Mitochondria					✓	✓	✓	✓	✓	✓
Plastid							✓		✓	
Chloroplast							✓	✓		
Cyanelle							✓			
Chlorosome			✓							
Inner Membrane				✓						
Outer Membrane				✓						
Periplasm			✓							
Fimbrium			✓							

Major Tasks

- Appropriate Machine Learning approach applicable at every decision node
- Incorporation of evolutionary profiles
- Use of TMHMM at highest level node or every leaf
- Comparison to other methods for localization prediction
- Benchmark using fully sequenced genomes

**Thank
You!**

