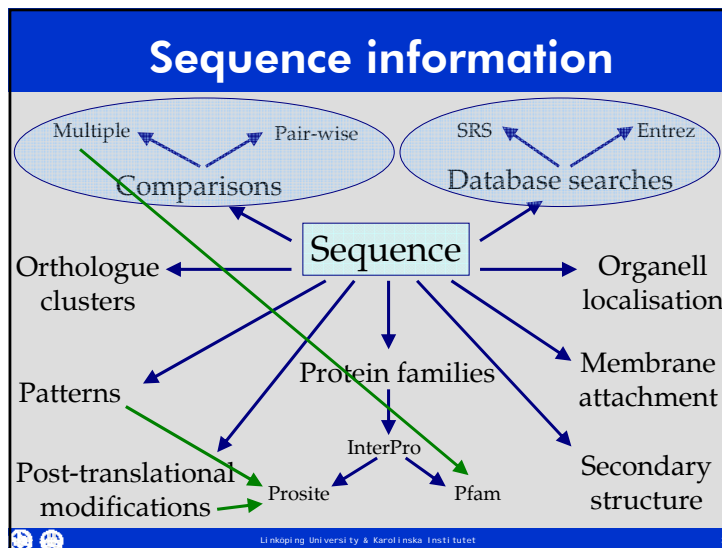
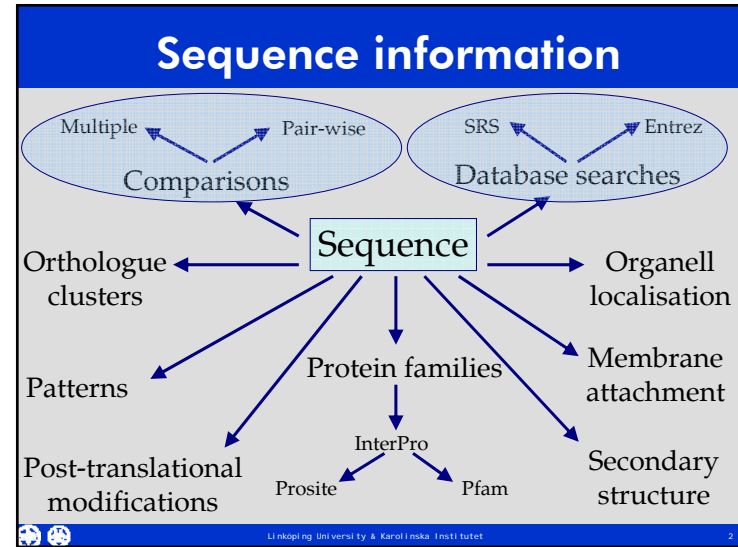


Sequence Information

Bengt Persson



- ### Good web sites
- ★ www.expasy.org
 - ★ www.ebi.ac.uk
 - ★ www.ncbi.nlm.nih.gov
 - ★ www.cbs.dtu.dk
- Linköping University & Karolinska Institutet

Protein family databases

Protein families, nomenclature

- ★ Super-family
 - Family
 - Sub-family

InterPro

- ★ Prosite
 - Amos Bairoch, Genève
- ★ Pfam
 - Erik Sonnhammer, KI and Sanger Institute, UK
- ★ PRINTS
 - Terri Attwood, UCL, London, UK
- ★ ProDom
 - Daniel Kahn, INRA, Toulouse, France
- ★ SMART
 - Peer Bork, EMBL
- ★ Swissprot+TrEMBL

InterPro entry

The screenshot shows the InterPro website interface. The main content area displays the entry for IPR001304, titled "C-type lectin domain". The entry details include:

- Database:** InterPro
- Accession:** IPR001304 (matches 751 proteins)
- Name:** C-type lectin domain
- Type:** Domain
- Dates:** 08-OCT-1999 (created), 28-JUN-2000 (last modified)
- Signatures:**
 - PS00615: C_TYPE_LLECTIN_1 (401 proteins)
 - PS50041: C_TYPE_LLECTIN_2 (672 proteins)
 - PF00055: lectin_c (483 proteins)
- Found in:** IPR01491: Thrombomodulin (13 proteins)
- Children:**
 - IPR002352: Eosinophil major basic protein (9 proteins)
 - IPR002353: Type II antitrypsin protein (44 proteins)
- Abstract:** A number of different families of proteins share a conserved domain which was first characterized in...

InterPro entry, cont.

Found in

- PF00059: lectin_c (483 proteins)
- PF001491: Thrombomodulin (13 proteins)
- PF0002352: Eosinophil major basic protein (9 proteins)
- PF0002353: Type II antifreeze protein (44 proteins)

Abstract

A number of different families of proteins share a conserved domain which was first characterized in some animal lectins and which seems to function as a calcium-dependent carbohydrate-recognition domain [1, 2, 3]. This domain is known as the C-type lectin domain (CTL) or as the carbohydrate-recognition domain (CRD). The categories of proteins in which the CTL domain has been found include:

1. type-II membrane proteins where the CTL domain is located at the C-terminal extremity of the proteins,
2. proteins, sometimes called 'collectins', that consist of an N-terminal collagenous domain followed by a CTL-domain [4].
3. selectins (or LEC-CAM), cell adhesion molecules implicated in the interaction of leukocytes with platelets or vascular endothelium [5, 6]. Structurally, selectins consist of a long extracellular domain, followed by a transmembrane region and a short cytoplasmic domain. The extracellular domain is itself composed of a CTL-domain, followed by an EGF-like domain and a variable number of SCR/Sushi repeats,
4. large proteoglycans that contain a CTL-domain followed by one copy of a SCR/Sushi repeat at the C-terminus. In addition they may also contain, in their N-terminal domain, an Ig-like V-type region, two or four link domains and up to two EGF-like repeats,
5. type-I membrane proteins, and
6. various other proteins that uniquely consist of a CTL domain.

InterPro entry, cont.

Examples

- P10716 KUCR_RAT: Kupffer cell receptor from rat - type-II membrane protein
- P20765 LEM1_PAPHA: Baboon L-selectin
- P22297 MANR_HUMAN: Mannose receptor from macrophages - a type-I membrane protein
- P26874 PSPA_PIG: Pulmonary surfactant-associated protein A (SP-A)
- P31113 ABA2_TREAB: Bovine trehalase, a large proteoglycan
- P12108 LEC2_POLM: A calcium-dependent galactose-binding lectin

References

1. Drickamer K. Two distinct classes of carbohydrate-recognition domains in animal lectins. *J Biol Chem*. 263: 9557-9560(1988) [MEDLINE 83257070] [PUB00002490]
2. Drickamer K. Evolution of Ca²⁺-dependent animal lectins. *Prog Nucleic Acid Res. Mol Biol* 45: 207-232(1993) [MEDLINE 93342185] [PUB00004941]
3. Drickamer K. *Curr Opin Struct Biol* 3: 393-400(1993) [PUB00001078]
4. Weiss WI, Kahn R, Faume R, Drickamer K, Hendrickson WA. Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science* 254: 1608-1615(1991) [MEDLINE 92066855] [PUB00005149]
5. Siegelman M. *Curr Biol* 1: 125-128(1991) [PUB00001011]
6. Leblay A. Selectins: interpreters of cell-specific carbohydrate information during inflammation. *Science* 258: 984-986(1992) [MEDLINE 93068290] [PUB00005159]

Database links PROSITE doc: P0000537; Blocks: E0001304

Matches [Table](#) [all](#) [Graphical](#) [all](#)

InterPro -- protein matches

InterPro - Proteins matching IP001304

Table [Graphical](#)

Item 1-20 of 905 < 1 2 3 4 5 Next >

	P800816	P850041	PF00059	SM00034
000274 000274		32-165 T	51-180 T	25-181 T
00448 00448	259-281 T	165-282 T	184-281 T	158-282 T
001263 001263				56-194 T
001336 001336			114-252 T	83-253 T
001620 001620				90-202 T
001622 001622				15-179 T
001701 001701			166-270 T	135-273 T
001782 001782			3-108 T	75-211 T
			111-210 T	

InterPro -- protein matches, graphical

InterPro - Proteins matching IP001304

Graphical

Old shows 10aa intervals, first mark at position 0. Move the mouse over a match to see more information in the status bar of your browser window.

Item 1-20 of 905 < 1 2 3 4 5 Next >

Protein Match Display

Protein	Match Display	Match
TRMBL000274	ER000152 ER000110	ASX_HYDROX
000274	ER000561 ER01186	EGF_2
	ER000561 EF00008	EGF
	ER000561 SM00001	EGF_Mce
	TR001304 EF50041	C_TYPE_LLECTIN
	TR001304 PF00059	lectin_c
	TR001304 SM00034	CLBCT
	TR001881 TR01187	EGF_CA
	TR001881 EF50034	EGF_CA_2
	TR001881 SM00172	EGF_CA
TRMBL000274	TR001304 R006615	C_TYPE_LLECTIN_1

Prosite

- ★ Database of protein families and domains
- ★ Release 16, September 1999
- ★ 1035 documentation entries
- ★ 1375 different patterns

<http://www.expasy.ch/prosite/>

Amos Bairoch, University of Geneva

Prosite

Prosite

ScanProsite

Pfam

Ljnköping University & Karolinska Insti tutet 21

Pfam

Domain	Start	End	Type	Score	Start	End
Smart_22081_1	49	136	signal peptide	1	1	21
Smart_22021_98	136	195	low complexity	6	11	10000
IG	137	195	low complexity	522	533	20500
EGF	222	253	low complexity	566	575	19200
EGF	313	342	low complexity	769	784	22300
IG	363	426				
IG	444	531				
IG	544	630				
IG	642	727				
Smart_2225_728	636	837				
EGF	837	1105				

Ljnköping University & Karolinska Insti tutet 22

Pfam

Ljnköping University & Karolinska Insti tutet 23

COG--Clusters of Orthologous Groups

Code	Name	Proteins	Typical component analysis of genomes
A	Archaeoglobus fulgidus	2420	18493
O	Bacillus subtilis	2058	1404
M	Methanocaldococcus jannaschii	1786	1200
T	Methanobacterium thermoautotrophicum	1873	1375
P	Thermoplasma acidophilum	1479	1176
K	Zoonospora berkeleyi	2080	1345
Z	Zoonospora alba	1767	1443
E	Escherichia coli	2722	1169
T	Thermoplasma volcanium	5954	2175
Q	Archaeoglobus	1560	1317
V	Thermoplasma	1858	1507

Ljnköping University & Karolinska Insti tutet 24

Structure predictions

- ★ Secondary structure
- ★ Hydrophilicity
- ★ Membrane-spanning regions
- ★ Antigenicity
- ★ Glycosylation
- ★ Acetylation
- ... and much more ...

29

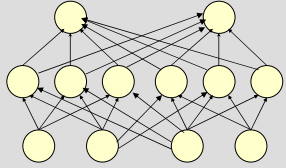
Secondary structure predictions

- ★ Chou & Fasman (CF)
- ★ Garnier, Osguthorpe & Robson (GOR)
 - http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
- ★ neural networks (e.g. PHD)
 - <http://dodo.cpmc.columbia.edu/predictprotein/>

30

Artificial Neural Networks (ANNs)

- ★ Statistical method
- ★ Pattern recognition, e. g. secondary structure predictions



Output layer


Hidden layer

Input layer

modified from Yvonne Kallberg

31

The PredictProtein server



32

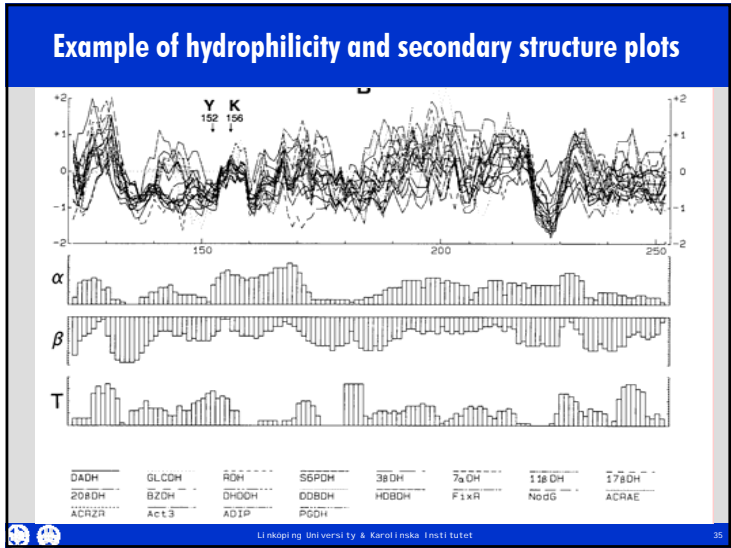
Default submission form

Liikopingi Universiti & Karoliinska Instituti

Hydrophilicity

- ★ Kyte & Doolittle
- ★ Hopp & Woods

Liikopingi Universiti & Karoliinska Instituti

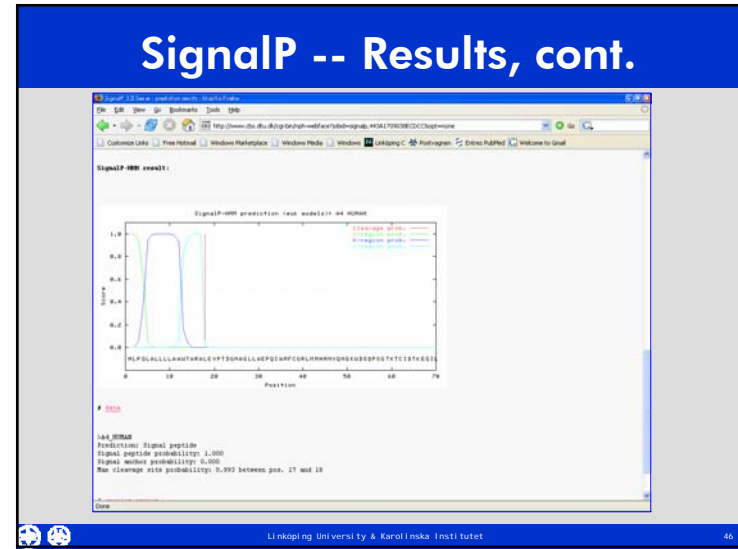
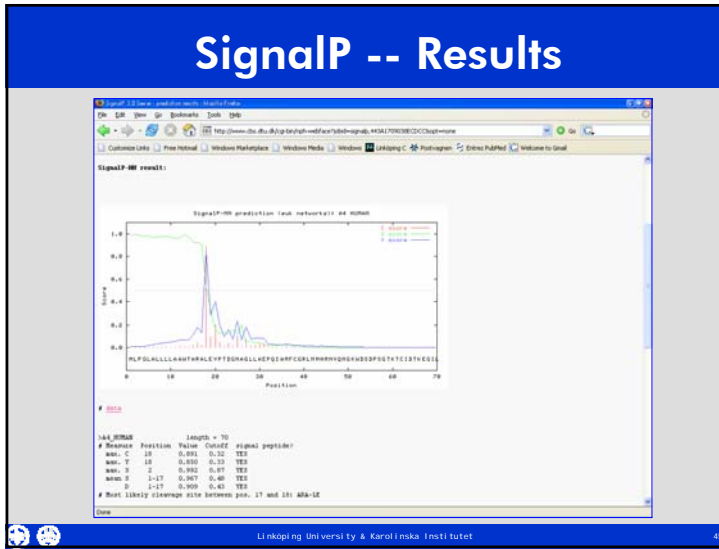


ProtScale

- ★ A general tool for plotting sequence properties, e.g. hydrophilicity

– <http://www.expasy.ch/cgi-bin/protscale.pl>

Liikopingi Universiti & Karoliinska Instituti



TargetP

TargetP 1.1 Server

TargetP 1.1 predicts the subcellular location of eukaryotic proteins. The location assignment is based on the predicted presence of any of the N-terminal presequences: mitochondrial transit peptide (MTP), mitochondrial targeting peptide (mtTP) or secretory pathway signal peptide (SP).

For the sequences predicted to contain an N-terminal presequence a potential cleavage site can also be predicted.

NOTE 1: TargetP uses [CleavPro](#) and [SignalP](#) to predict cleavage sites for TP and SP, respectively.

NOTE 2: The method has been tested on A. thaliana and H. sapiens sets, see the [results](#).

NOTE 3: This page has been rewritten recently (April 2005).

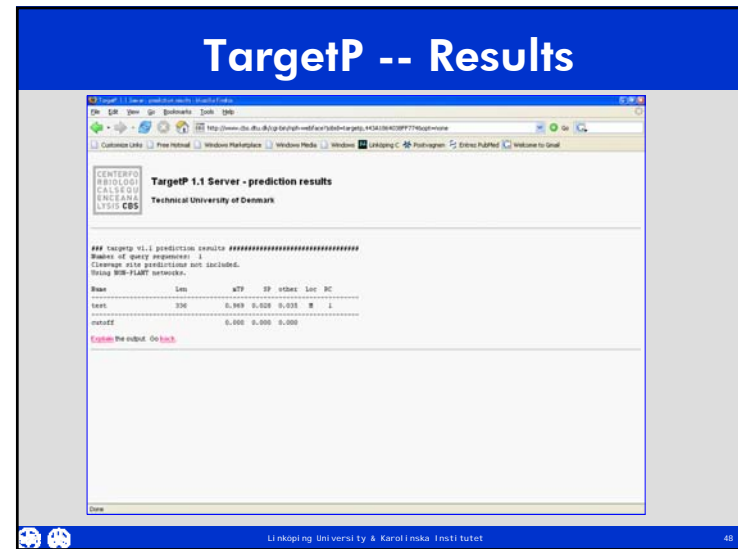
Submission options: [Both options](#) | [Output format](#) | [Article abstract](#) | [Data sets](#)

SUBMISSION

Provide a single sequence of several sequences in [FASTA](#) format into the text below:

```
>SEQ1_SEQ10
MPLGLALLLAARARALEYPTSDKALLKQIARFQKLNRRHNRHNSQSDSPSTATCTTCEKSL
</pre>

```



Post-translational modifications

The screenshot shows a web browser window with the title "ELPASy Tools - Metaspaces". The main content area is titled "Post-translational modification prediction" and lists several tools:

- SECRET** - Prediction of protein sorting signals and localization sites
- SignalP** - Prediction of signal peptide cleavage sites
- ChloroP** - Prediction of chloroplast transit peptides
- MITOPROT** - Prediction of mitochondrial targeting sequences
- PreDot** - Prediction of mitochondrial and plastid targeting sequences
- NucOglyc** - Prediction of type O-glycosylation sites in mammalian proteins
- Net-PI Predictor** - GPI Modification Site Prediction
- DOG1** - Prediction of GPI-anchor and cleavage sites
- NetPhos** - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- NetProSIA** - Prediction of protease cleavage sites in prokaryotic proteins

Below this section, there is a "Primary structure analysis" section with tools like **ProtParam**, **Compute pI/Mw**, **MW, pI, Titration curve**, **REP**, and **SAPS**.

Primary structure analysis

The screenshot shows a web browser window with the title "ELPASy Tools - Metaspaces". The main content area is titled "Primary structure analysis" and lists several tools:

- ProtParam** - Physico-chemical parameters of a protein sequence (amino-acid and atomic composition, pI, extinction coefficient, etc.)
- Compute pI/Mw** - Compute the theoretical pI and Mw from a SWISS-PROT or TrEMBL entry or for a user sequence
- MW, pI, Titration curve** - Computes pI, composition and allows to see a titration curve
- REP** - Searches a protein sequence for a repeats
- SAPS** - Statistical analysis of protein sequences at EMBOSS-CH [Also available at EBI]
- Cole** - Prediction of coiled coil regions in proteins (Lupas's method) at EMBOSS-CH [Also available at EBI]
- Target** - Prediction of coiled coil regions in proteins (Berger's method)
- Multicoil** - Prediction of two- and three-stranded coiled coils
- PEST** - Identification of PEST regions
- PESTfind** - Identification of PEST regions at EMBOSS Austria
- HLA_Bind** - Prediction of MHC type I (HLA) peptide binding
- SYFPEITHI** - Prediction of MHC type I and II peptide binding
- ProtScale** - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- ConSurf** - Draw an RGA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Protein Colorizer** - Tool for coloring your amino acid sequence
- Colorater** - Tool to highlight (in red) a selected set of residues in a protein sequence
- HelixWheel** - Representation of a protein fragment as a helical wheel
- RandomSeq** - Random protein sequence generator

Secondary structure prediction

The screenshot shows a web browser window with the title "ELPASy Tools - Metaspaces". The main content area is titled "Secondary structure prediction" and lists several tools:

- AGADE** - An algorithm to predict the helical content of peptides
- BUMPSH** - Baylor College of Medicine
- TriP** - Cascaded Multiple Classifiers for Secondary Structure Prediction
- GOR I** (Garnier et al. 1978) [At EBI, or at SBIIS]
- GOR II** (Garnier et al. 1987)
- SOE-IV** (Garnier et al. 1996)
- HN1** - Hierarchical Neural Network method (Guarnier, 1997)
- Jpred** - A consensus method for protein secondary structure prediction at EBI
- JPRED4** - University of California at San Francisco (UCSF)
- ProteinProtein** - PHDsec, PHDdist, PHDhtm, PHDtheadr, MacHom, EvalSec from Columbia University
- PREDATOR** - Protein secondary structure prediction from single or multiple sequences at EMBL (Argos' group)
- PSA** - BioMolecular Engineering Research Center (EMERCO) / Boston
- ESpred** - Various protein structure prediction methods at Brand University
- SOFP** (Georgiyev and Delage, 1994)
- SOEPA** (Georgiyev and Delage, 1995)

Below this section, there is a "Tertiary structure" section with tools like **SWISS-MODEL**, **CPHmodels**, **3D-FSSM**, **SWERT**, and **Sensi-Pho-Viewer**.

Transmembrane regions & Sequence alignments

The screenshot shows a web browser window with the title "ELPASy Tools - Metaspaces". The main content area is titled "Transmembrane regions detection" and lists several tools:

- DAS** - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method (Stockholm University)
- EMOTOP** - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
- PredictProton** - Prediction of transmembrane helix location and topology (Columbia University)
- SOSUI** - Prediction of transmembrane regions (TUAT, Tokyo Univ. of Agriculture & Technology)
- TMAP** - Transmembrane detection based on multiple sequence alignment (Karolinska Institut, Sweden)
- TMDMM** - Prediction of transmembrane helices in proteins (CBS, Denmark)
- TMPred** - Prediction of transmembrane regions and protein orientation (EMBOSS-CH)
- TopPred.2** - Topology prediction of membrane proteins (Stockholm University)

Below this section, there is a "Sequence alignment" section with tools like **SM + LALIGNVIEW**, **LALIGN**, **Dotlet**, **CLUSTALW**, **ALIGN**, **DALIGN**, **Match-Reg**, **MSA**, **Multalin**, and **MUSCA**.