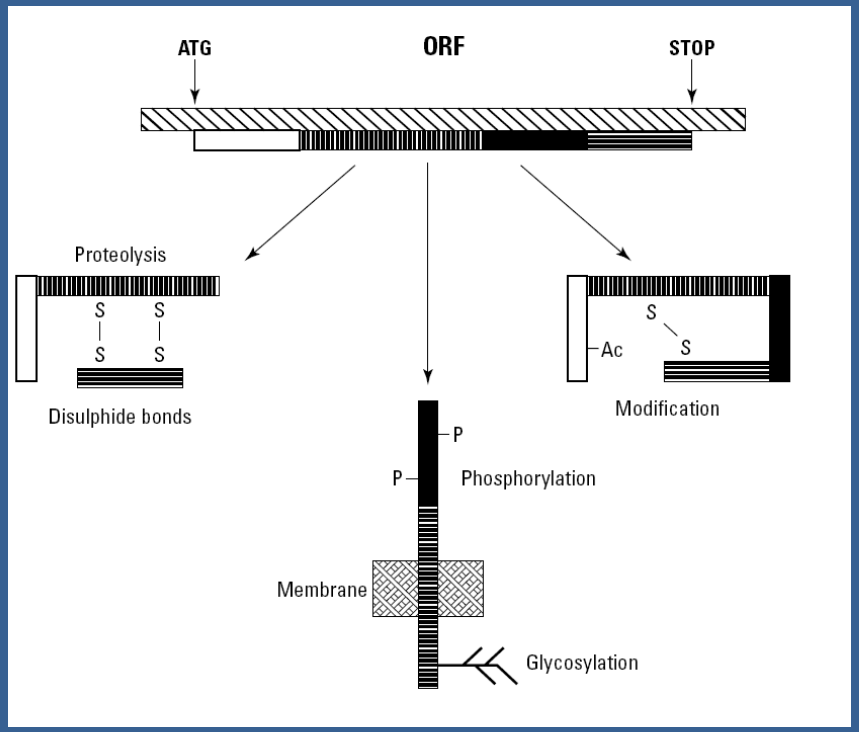
Predição de modificações em proteínas

Modificações de proteína



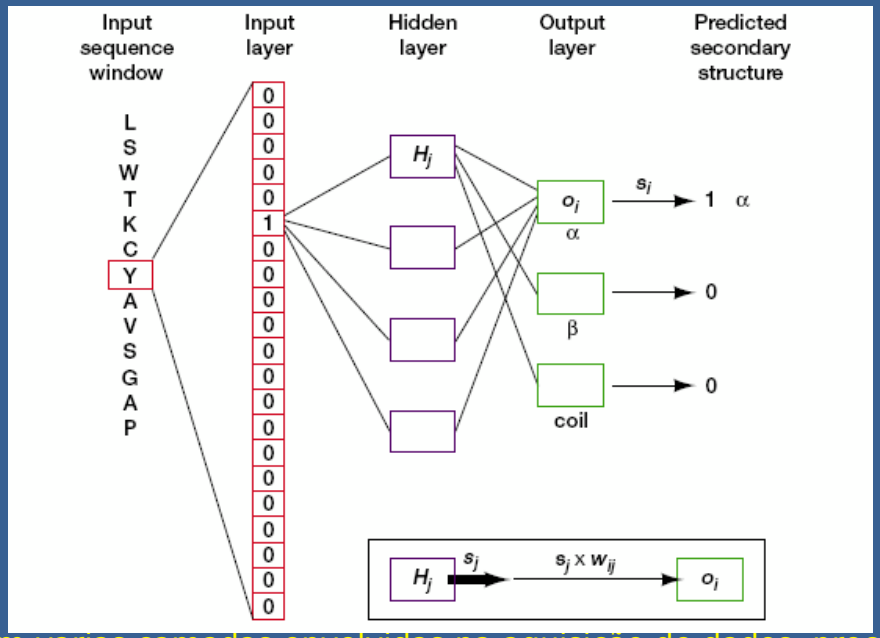
Ao representarmos uma proteína como uma seqüência de caracteres representando os seus aminoácidos (estrutura primaria) estamos representando apenas parte dos elementos que compões diversas proteínas que podem sofrer diversas modificações como proteólise, glicosilação, fosforilação, etc...

Em vários casos tais modificações podem fornecer indícios importantes sobre localização celular e função

Modificações de proteína

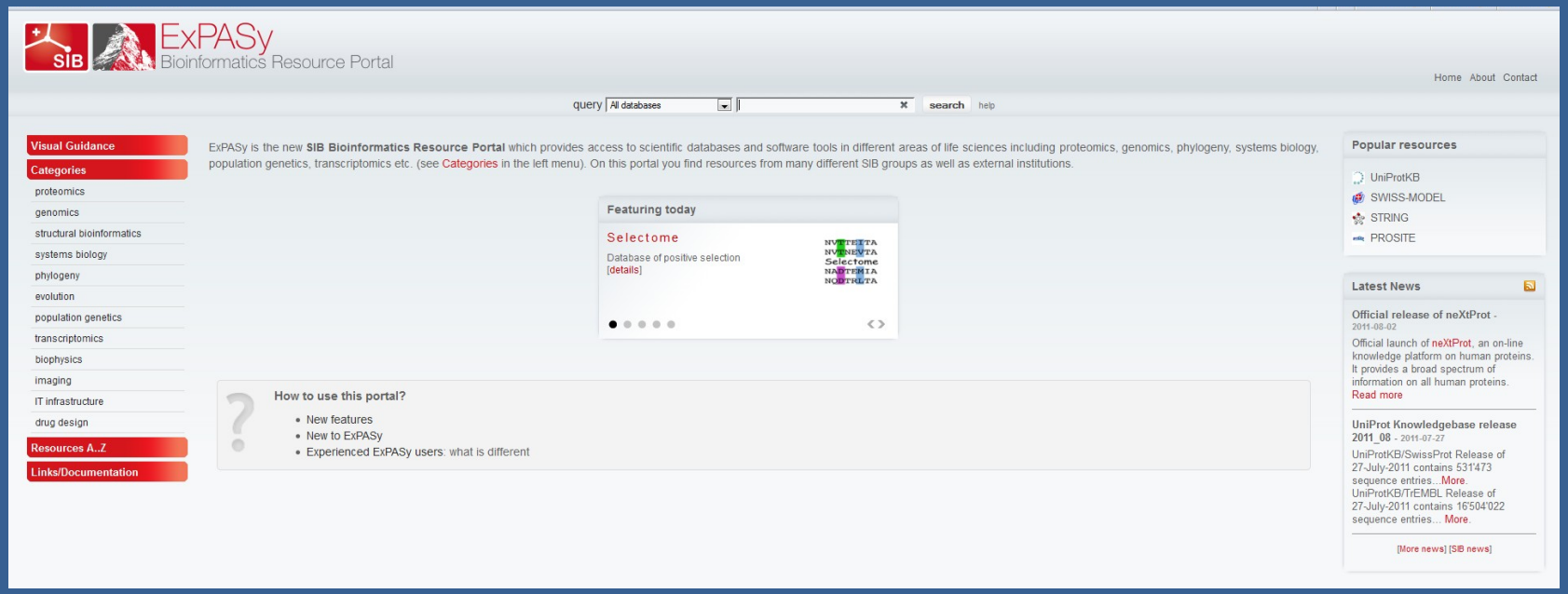
- Diversas das modificações ocorridas nas proteínas acontece através da ação enzimática de outras proteínas que reconhecem motivos nas seqüências alvo
- Deste modo é possível utiliza algoritmos de reconhecimento de padrões para reconhecer eventuais sítios de modificação em uma proteína estudada

Redes neurais



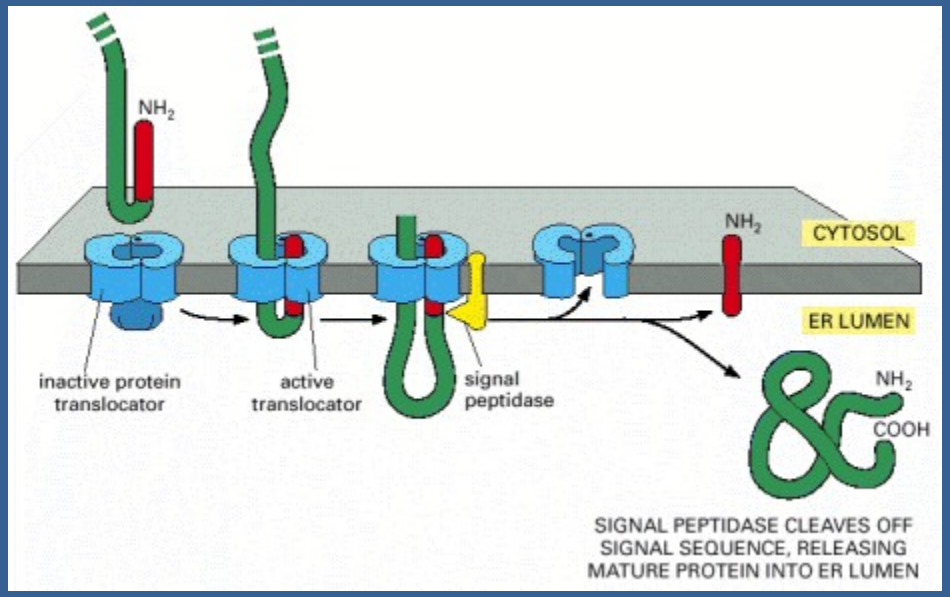
- Possui uma rede com varias camadas envolvidas na aquisição de dados, processamento e resposta. O dados da camada de aquisição é enviado a camada de processamento, onde funções irão gerar valores 1 ou 0 simulando o disparo ou não de um neurônio. A camada de resposta irá receber os diferentes sinais da camada de processamento e por sua vez irá produzir um sinal 1 ou 0.
- Processos de treinamento para ajustar o peso de cada função são realizados com seqüências com resposta conhecida

ExPASy



- Pagina contem links para diversos programas de analise de proteínas

peptídeo sinal

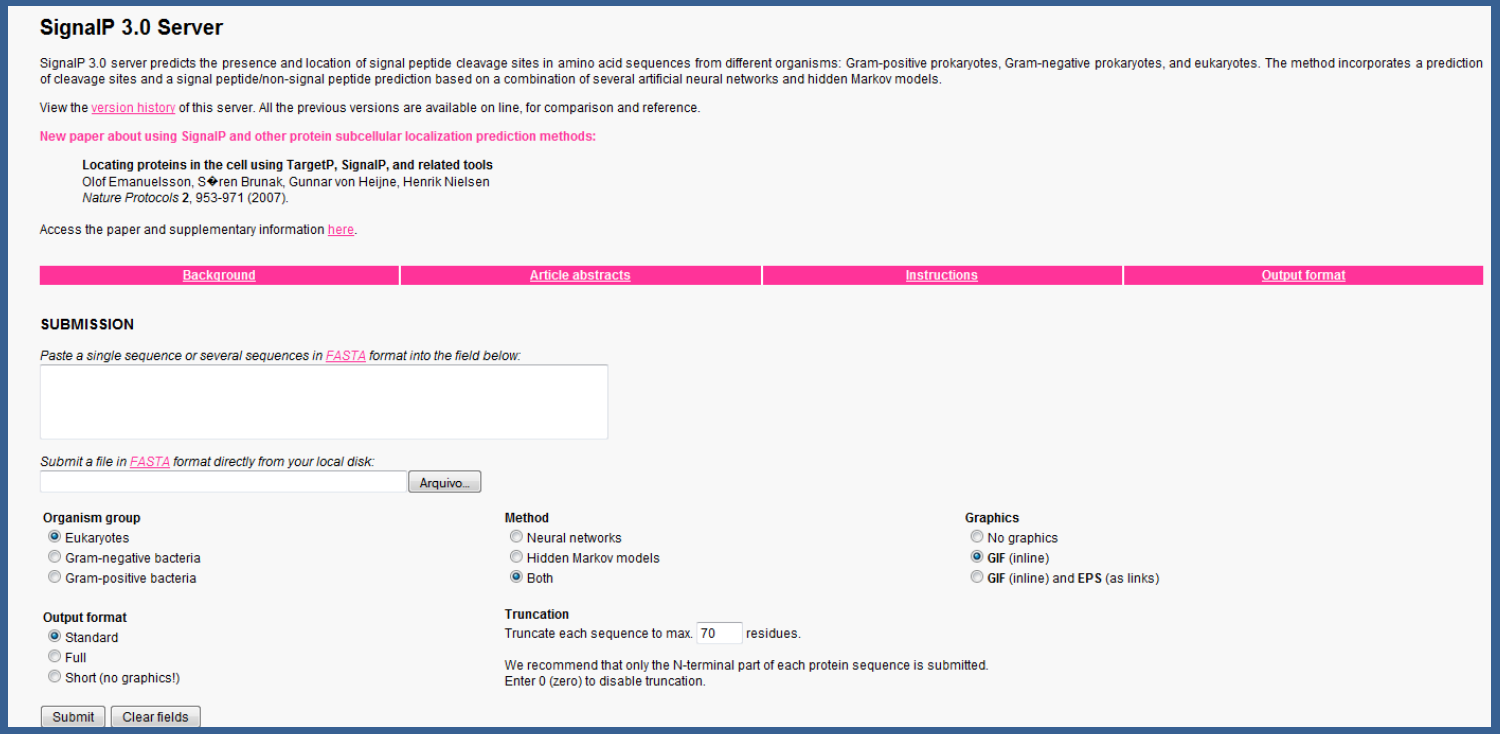


O transporte de proteínas para o lúmen do reticulo endoplasmático é realizado por um aparato celular que reconhece um peptídeo sinal na ponta amino-terminal de uma proteína e transporta esta proteína através da membrana separando o citosol do lúmen do RE.
As proteínas transportadas para o RE serão secretada para fora da célula, salvo se tiverem algum sinal de retenção.
Além disso, é somente no RE e golgi que ocorre a glicosilação de proteínas.

SignalP

- Detecção de peptídeo sinal por dois algoritmos diferentes: HMM e Neural Networks
- Utilização de um conjunto de treinamento real para estabelecimento de parâmetros de detecção
- Oferece ferramentas treinadas em eucariotos e bactérias gram-positiva e negativas

SignalP



SignalP 3.0 Server

SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

New paper about using SignalP and other protein subcellular localization prediction methods:

Locating proteins in the cell using TargetP, SignalP, and related tools
Olof Emanuelsson, S◆ren Brunak, Gunnar von Heijne, Henrik Nielsen
Nature Protocols **2**, 953-971 (2007).

Access the paper and supplementary information here.

Background | Article abstracts | Instructions | Output format

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Arquivo...

Organism group
- (•) Eukaryotes
- ( ) Gram-negative bacteria
- ( ) Gram-positive bacteria

Method
- ( ) Neural networks
- ( ) Hidden Markov models
- (•) Both

Graphics
- ( ) No graphics
- (•) **GIF** (inline)
- ( ) **GIF** (inline) and **EPS** (as links)

Output format
- (•) Standard
- ( ) Full
- ( ) Short (no graphics!)

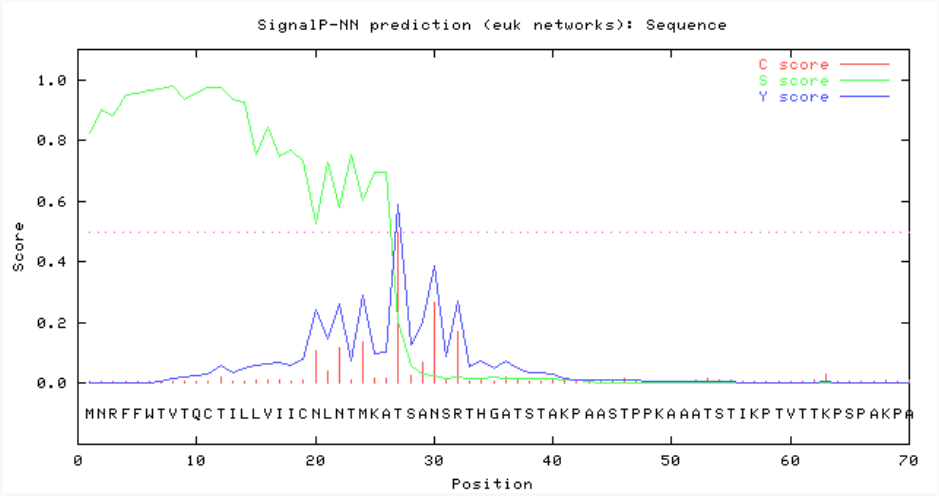
Truncation
Truncate each sequence to max. 70 residues.

We recommend that only the N-terminal part of each protein sequence is submitted.
Enter 0 (zero) to disable truncation.

Submit | Clear fields

SignalP-Neural networks

SignalP-NN result:



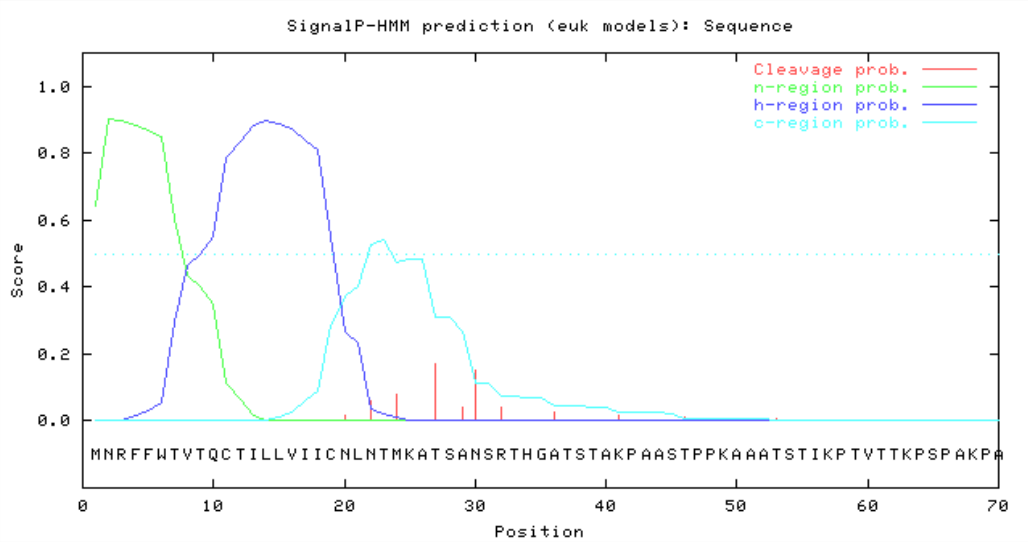
data

```
>Sequence                length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C    27       0.497   0.32   YES
  max. Y    27       0.589   0.33   YES
  max. S     8       0.979   0.87   YES
  mean S     1-26    0.829   0.48   YES
       D     1-26    0.709   0.43   YES
# Most likely cleavage site between pos. 26 and 27: MKA-TS
```

- Escore S - relacionado a probabilidade do aminoácido pertencer a seqüência sinal
- Escore C - probabilidade de ser o sitio de clivagem. Posição indicaria a primeiro aminoácido da proteína madura
- Escore Y - deriva da combinação dos escores S e C. Tende a ser máximo no ponto de clivagem
- Media de S- media do valor do escore S no peptídeo sinal
- Media D- Media do valor médio de S e valor máximo de Y

SignalP- HMM

SignalP-HMM result:



data

```
>Sequence
Prediction: Signal peptide
Signal peptide probability: 0.642
Signal anchor probability: 0.259
Max cleavage site probability: 0.173 between pos. 26 and 27
```

- Detecção de pontas amino e carboxi terminais e região hidrofóbica.

Performance do SignalP

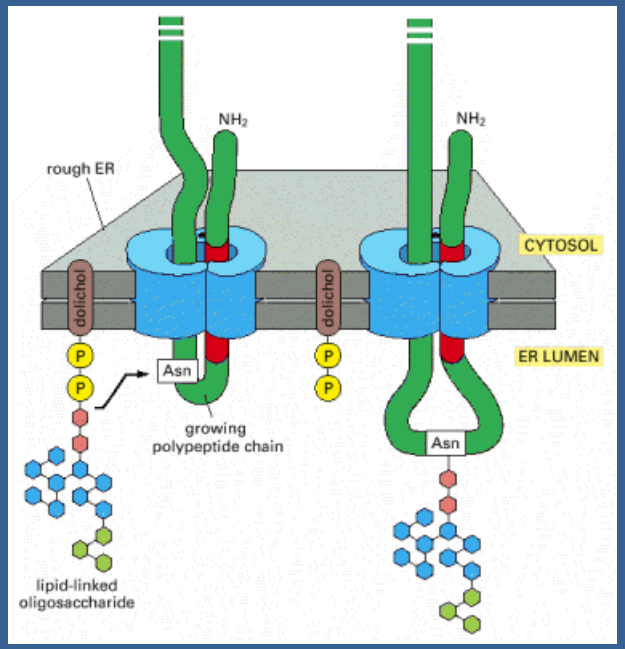
Version	Cleavage site (Y-score)			Discrimination (SP/non-SP)		
	Euk	Gram−	Gram+	Euk	Gram−	Gram+
SignalP 1 NN	70.2	79.3	67.9	0.97	0.88	0.96
SignalP 2 NN	72.4	83.4	67.4	0.97	0.90	0.96
SignalP 2 HMM	69.5	81.4	64.5	0.94	0.93	0.96
SignalP 3 NN	79.0	92.5	85.0	0.98	0.95	0.98
SignalP 3 HMM	75.7	90.2	81.6	0.94	0.94	0.98

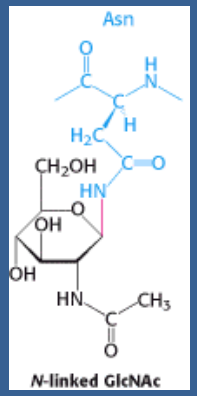
Table 1: **Performances of three different SignalP versions.** The most significant improvement was for the cleavage site predictions. Cleavage site performances are presented as % and discrimination values (based on D-score) as correlation coefficients. NN and HMM indicate neural network and hidden Markov model, respectively. Results are based on five-fold cross validation for all SignalP versions

Glicosilação de proteínas

- Adição de oligossacarídeos a cadeia protéica em posições definidas
- Processo é especifico ocorrendo em apenas algumas proteínas e somente no RE e Golgi (portanto proteínas citoplasmáticas não estão a princípio sujeitas a este tipo de modificação)
- Ligação de carboidratos pode ter diversos propósitos: Reconhecimento celular, adesão, entre outras

Glicosilação N-ligada de proteínas





Glicosilação N- ligada liga-se no grupo NH_2 do asparagina

Ocorre apenas em motivos Asn-Xaa-Ser/Thr, onde Xaa é qualquer aminoácido menos prolina

Neste tipo de glicosilação um oligossacarídeo inteiro contendo 14 resíduos de açúcar é adicionados a cadeia lateral do aminoacido

Glicosilação ocorre no reticulo endoplasmático e modificações na cadeia de oligossacarídeos ocorre no RE e Golgi

NetNGlyc

NetNGlyc 1.0 Server

The NetNglyc server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.

Instructions | Output format | Abstract

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Arquivo...

Alternatively, type in Swiss-Prot ID/AC (e.g. CBG_HUMAN)

Generate graphics **Show additional thresholds (0.32, 0.75, 0.90) in the graph(s)**

By default, predictions are done only on the Asn-Xaa-Ser/Thr sequons (incl. Asn-Pro-Ser/Thr)

Predict on all Asn residues - *use this only if you know what you are doing!*

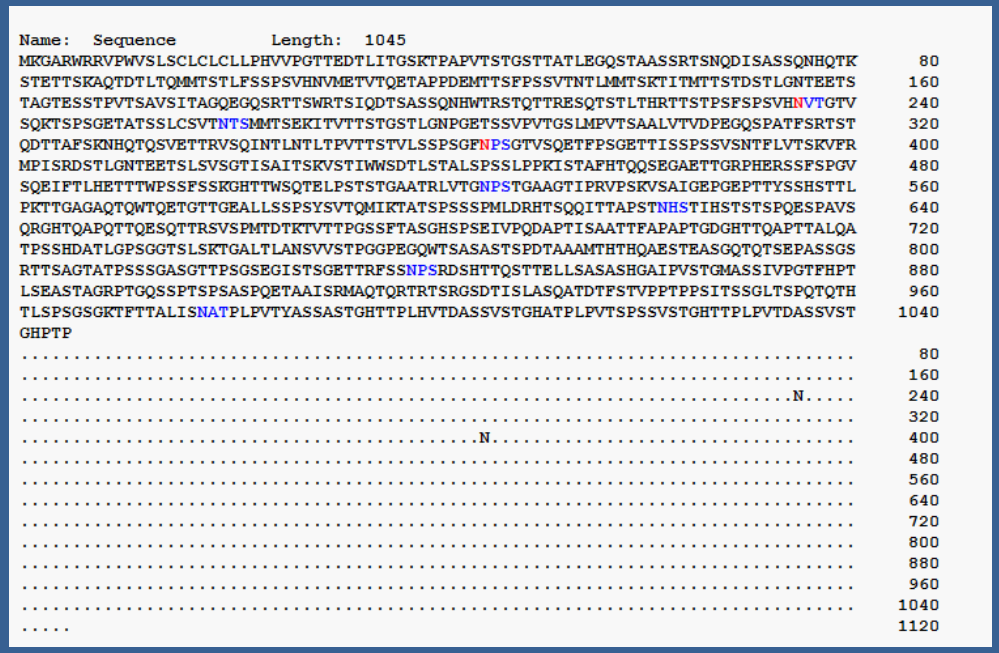
Submit Clear fields

Notes: SignalP is automatically run on all sequences. A warning is displayed if a signal peptide is not detected. In transmembrane proteins, only extracellular domains may be N-glycosylated. This is currently not checked by the NetNGlyc server. Cytoplasmic and transmembrane sequence regions may be predicted to be glycosylated - this should, of course, be ignored. One transmembrane region predictor is TMHMM.

Restrictions: *At most 2,000 sequences and 200,000 amino acids per submission; each sequence not more than 5,000 amino acids.*
Confidentiality: *The sequences are kept confidential and will be deleted after processing.*

Ferramenta de predição de sítios de N-glicosilação
Utiliza redes neurais
Não é possível prever a cadeia de oligonucleotideos resultantes
Treinada em proteínas de humano

NetNGlyc

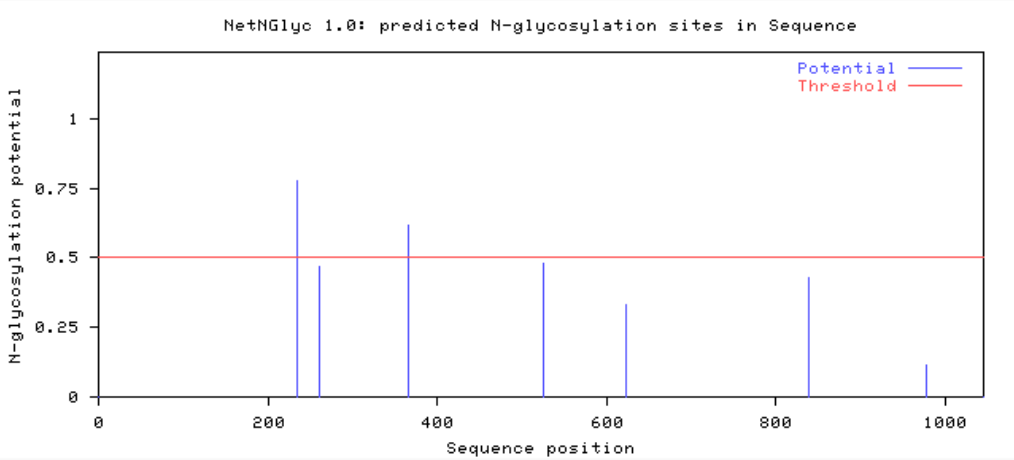


```
Name:  Sequence          Length:  1045
MKGARWRRVPWVSLSCLCLCLLPHVVPGTTEDTLITGSKTPAPVTSTGSTTATLEGQSTAASSRTSNQDISASSQNHQTK      80
STETTSKAQTDTLTQMMTSTLFSSPSVHNVMETVTQETAPPDEMTTSFPSSVTNTLMMTSKTITMTTSTDSTLGNTEETS     160
TAGTESSTPVTSAVSITAGQEGQSRTTSWRTSIQDTSASSQNHWTRSTQTTRESQTSTLTHRTTSTPSFSPSVHNVTGTV     240
SQKTSPSGETATSSLCSVTNTSMMTSEKITVTTSTGSTLGNPGETSSVPVTGSLMPVTSAALVTVDPEGQSPATFSRTST     320
QDTTAFSKNHQTQSVETTRVSQINTLNTLTPVTTSTVLSSPSGFNPSGTVSQETFPSGETTISSPSSVSNTFLVTSKVFR     400
MPISRDSTLGNTEETSLSVSGTISAITSKVSTIWWSDTLSTALSPSSLPPKISTAFHTQQSEGAETTGRPHERSSFSPGV     480
SQEIFTLHETTTWPSSFSSKGHTTWSQTELPSTSTGAATRLVTGNPSTGAAGTIPRVPSKVSAIGEPGEPTTYSSHSTTL     560
PKTTGAGAQTQWTQETGTTGEALLSSPSYSVTQMIKTATSPSSSPMLDRHTSQQITTAPSTNHSTIHSTSTSPQESPAVS     640
QRGHTQAPQTTQESQTTRSVSPMTDTKTVTTPGSSFTASGHSPSEIVPQDAPTISAATTFAPAPTGDGHTTQAPTTALQA     720
TPSSHDATLGPSGGTSLSKTGALTLANSVVSTPGGPEGQWTSASASTSPDTAAAMTHTHQAESTEASGQTQTSEPASSGS     800
RTTSAGTATPSSSGASGTTPSGSEGISTSGETTRFSSNPSRDSHTTQSTTELLSASASHGAIPVSTGMASSIVPGTFHPT     880
LSEASTAGRPTGQSSPTSPSASPQETAAISRMAQTQRTRTSRGSDTISLASQATDTFSTVPPTPPSITSSGLTSPQTQTH     960
TLSPSGSGKTFTTALISNATPLPVTYASSASTGHTTPLHVTDASSVSTGHATPLPVTSPSSVSTGHTTPLPVTDASSVST    1040
GHPTP
.....     80
.....     160
.....N.....     240
.....     320
.....N.....     400
.....     480
.....     560
.....     640
.....     720
.....     800
.....     880
.....     960
.....    1040
.....    1120
```

- Seqüências em azul indicam motivos NXS/T
- Resíduos em vermelho indicam aquele acima do limiar que são preditos como glicosilados

NetNGlyc

SeqName	Position	Potential	Jury agreement	N-Glyc result	
Sequence	235 NVTG	0.7770	(9/9)	+++	
Sequence	260 NTSM	0.4706	(5/9)	-	
Sequence	365 NPSG	0.6197	(8/9)	+	WARNING: PRO-X1.
Sequence	525 NPST	0.4774	(5/9)	-	
Sequence	622 NHST	0.3310	(9/9)	--	
Sequence	838 NPSR	0.4310	(7/9)	-	
Sequence	978 NATP	0.1136	(9/9)	---	



- São consideradas positivas sequencias com potencial acima de 0.5
- Jury agreement- lista o numero de redes neurais que apresentam resultado concordante com resultado
- Pro-X1 –indica que existe uma prolina na posição X, e que portanto existe baixa probabilidade deste ser um sitio autentico

NetNGlyc

glycosylated sites:

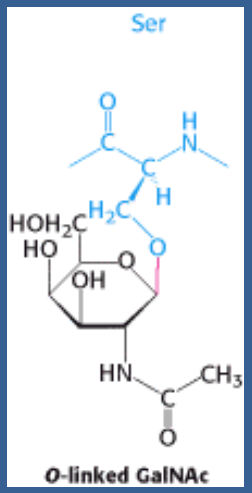
```
   + Potential > 0.5
  ++ Potential > 0.5 AND Jury agreement (9/9)  OR Potential>0.75
 +++ Potential > 0.75 AND Jury agreement
++++ Potential > 0.90 AND Jury agreement
```

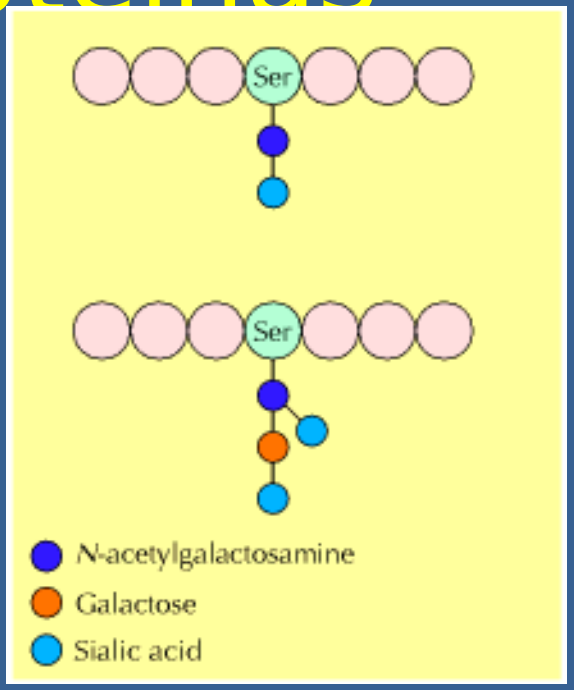
and *non-glycosylated* sites:

```
   - Potential < 0.5
  -- Potential < 0.5 AND Jury agreement (all nine < 0.5)
 --- Potential < 0.32 AND Jury agreement
```

- Sistema de escore do NetNGlyc

Glicosilação O-ligada de proteínas



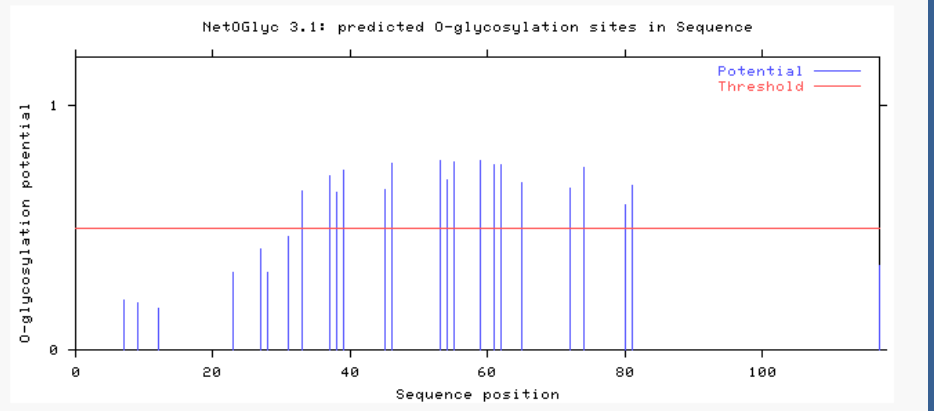


Glicosilação O- ligada liga-se no grupo OH da Serina ou Treonina
Adição de carboidratos acontece no Golgi
Apenas um monossacarídeo adicionado por vez.
Resulta em cadeias que normalmente são curtas

NetOGlyc

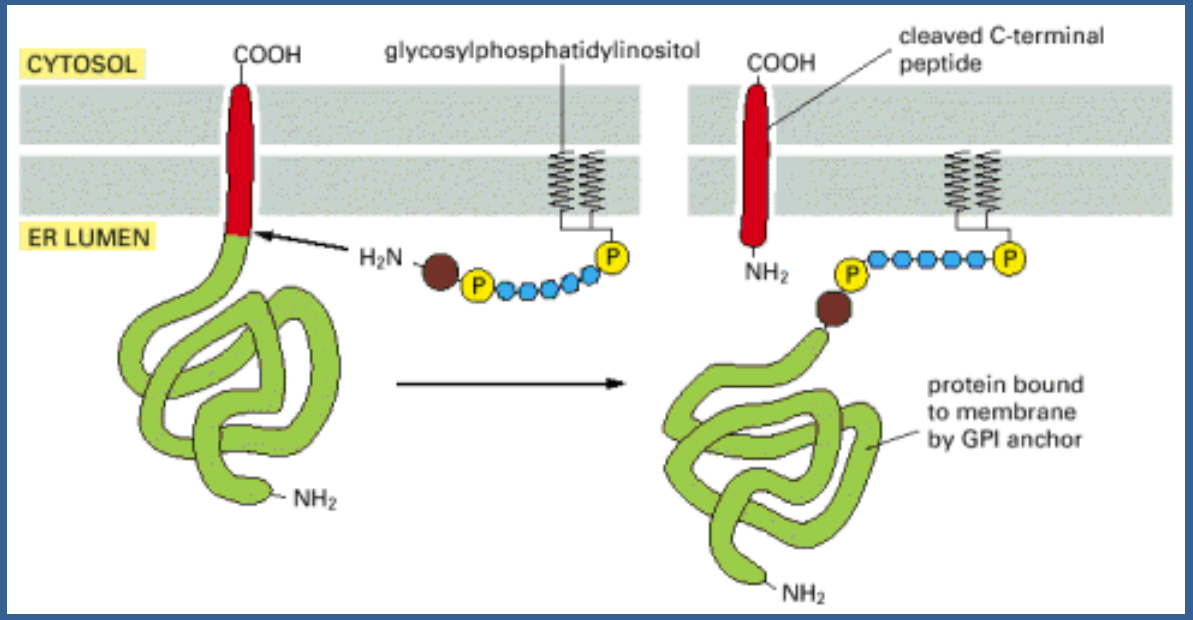
```
Name:  Sequence          Length:  117
MNRFFWTVTQCTILLVIICNLNTMKATSANSRTHGATSTAKPAASTPPKAAATSTIKPTVTTKPSPAKPAASNTAKPAAS
TPKKPHDERAVLAAAAVPIVLGVIGEVIGFILQYIAS
_____________________________.....T...TST.....ST......TST...T.TT..S......S.T.....S
T.....................................
```

Name	S/T	Pos	G-score	I-score	Y/N	Comment
Sequence	T	7	0.205	0.050	.	-
Sequence	T	9	0.190	0.041	.	-
Sequence	T	12	0.170	0.040	.	-
Sequence	T	23	0.317	0.052	.	-
Sequence	T	27	0.415	0.027	.	-
Sequence	S	28	0.317	0.051	.	-
Sequence	S	31	0.462	0.021	.	-
Sequence	T	33	0.648	0.054	T	-
Sequence	T	37	0.715	0.029	T	-
Sequence	S	38	0.645	0.029	S	-
Sequence	T	39	0.735	0.474	T	-
Sequence	S	45	0.659	0.466	S	-
Sequence	T	46	0.762	0.341	T	-
Sequence	T	53	0.775	0.279	T	-
Sequence	S	54	0.697	0.025	S	-
Sequence	T	55	0.771	0.434	T	-
Sequence	T	59	0.776	0.297	T	-
Sequence	T	61	0.757	0.574	T	-
Sequence	T	62	0.758	0.561	T	-
Sequence	S	65	0.686	0.178	S	-
Sequence	S	72	0.661	0.083	S	-
Sequence	T	74	0.746	0.071	T	-
Sequence	S	80	0.592	0.247	S	-
Sequence	T	81	0.674	0.516	T	-
Sequence	S	117	0.344	0.028	.	-



- Neste algoritimo a acessibilidade do resíduo é computado para predição de glicosilação
- G -score é utilizado como parâmetro de decisão para presença de glicosilação o-ligada

Predição de ancoras GPI



Processos no RE ligam covalentemente uma âncora de glicosilfosfatidilinositol (GPI) a região C-terminal de algumas proteínas de membrana

Proteína é exportada a membrana plasmática onde ficam expostas ao ambiente externo

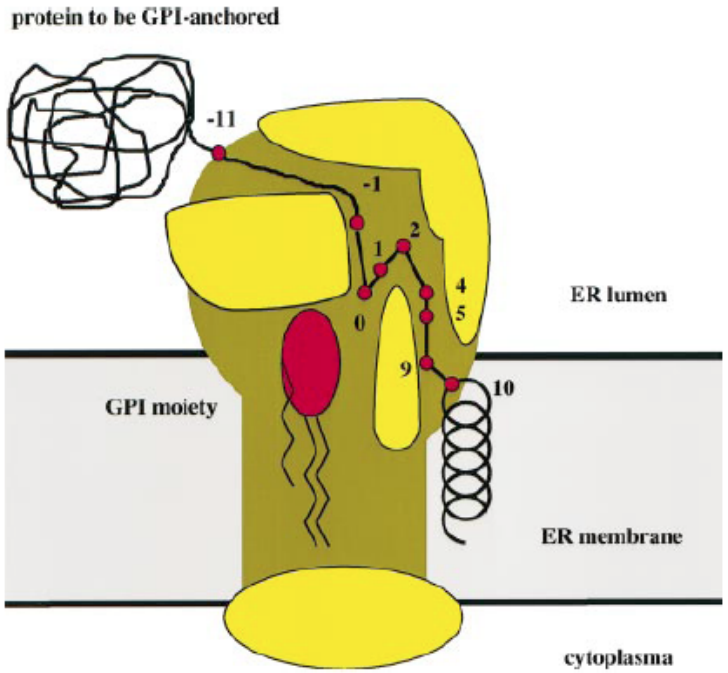
Em alguns casos sob a ação de uma fosfolipase especifica a proteína pode ser liberada da membrana

big-PI Predictor

big-PI Predictor

Table III. Sequence properties of the C-terminal propeptide near the ω-site

Position relative to the ω-site	Protozoa	Metazoa
–11...–1	Unstructured region	Unstructured region
0	Ser (44%), Asp, Asn, Ala, Gly	Ser (48%), Gly, Asn, Asp, Cys
1	Similar to position 0	Tiny (Gly, Ala, Ser)
2	Ser + Ala (94%)	Ala + Gly (70%)
4...5	Hydrophobic	Hydrophobic
9	Onset of hydrophobic region	Onset of hydrophobic region
	up to about ω + 25, mostly Leu	up to about ω + 31, mostly Leu
15	–	Ser + Thr + Ala (60%)



big-PI Predictor

Output of the prediction tool:

~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

Query sequence Query (length 148 amino acids):
MLSNAKLLLS LAMASTALGL VSNSSSSVIV VPSSDATIAG NDTATPAPEP SSAAPIFYNS
TATATQYEVV SEFTTYCPEP TTFVTNGATF TVTAPTTLTI TNCPCTIEKP TSETSVSSTH
DVETNSNAAN ARAIPGALGL AGAVMMLL

Best predicted site is shown in red.

~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

Prediction of potential C-terminal GPI-Modification Sites

~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

Use of the prediction function for METAZOA

Potential GPI-modification site was found.
Quality of the site ..... : P
Sequence position of the omega-site : 127
Score of the best site ..... : 8.96 (PValue = 3.017322e-04)
~~~~~~~~~~

big-PI Predictor

	Best Site
Total Score	8.96
Components of the Score Function:	
Profile Score	8.96
Term 0 Contents and Windows of DE in Region [-11..1]	0.00
Term 1 Hydrophilicity of N-terminal Region [-11..1]	0.00
Term 2 Penalty for low Profile Score in Region [0..2]	0.00
Term 3 Volume Limitation [-1..+2]	0.00
Term 4 Volume Compensation (-1, 1, 2)	0.00
Term 5 Volume Compensation (-1, 2)	0.00
Term 6 Backbone Flexibility [-1..2]	0.00
Term 7 Propeptide Length	0.00
Term 8 Hydrophilicity of Spacer Region [3..8]	0.00
Term 9 Volume Limitation [3..8]	-0.00
Term 10 Penalty for charged AAs in Spacer Region [3..10]	0.00
Term 11 Backbone Flexibility [3..8]	0.00
Term 12 Penalty for low Profile Score in Region [10..end]	0.00
Term 13 Hydrophobicity of Tail [10..end]	0.00
Term 14 Hydrophobicity of Tail [26..end]	0.00
Term 15 Even Distribution of Hydrophobicity [9..end]	0.00
Term 16 Penalty for polar Windows in Region [10..end]	0.00
Term 17 Penalty for SGC-Windows in Region [10..end]	0.00
Term 18 LVI Contents [10..end]	0.00
Term 19 Penalty for FYHW - Sections in Region [10..end]	0.00
Term 20 Penalty for Windows with small Volume [10..end]	-0.00
Profile independent Score	-0.00