Predizione delle caratteristiche biochimiche delle proteine

Caratteristiche biochimiche predicibili

CHIMICO-FISICHE

- Composizione e sequenza
- Peso molecolare
- Punto isoelettrico
- Coefficiente di estinzione
- Regioni a bassa complessità
- Regioni ripetute

STRUTTURALI

- Secondaria
- Terziaria
- Regioni Coiled coil
- Accessibilità al solvente
- Regioni transmembrana
- Topologia transmembrana

FUNZIONALI

- "Traducibilità"
- Destinazione ai compartimenti cellulari
- Turnover
- Modificazioni post-traduzionali
- Siti catalitici
- Legami a coenzimi
- Legami a metalli
- Legami a DNA o proteine

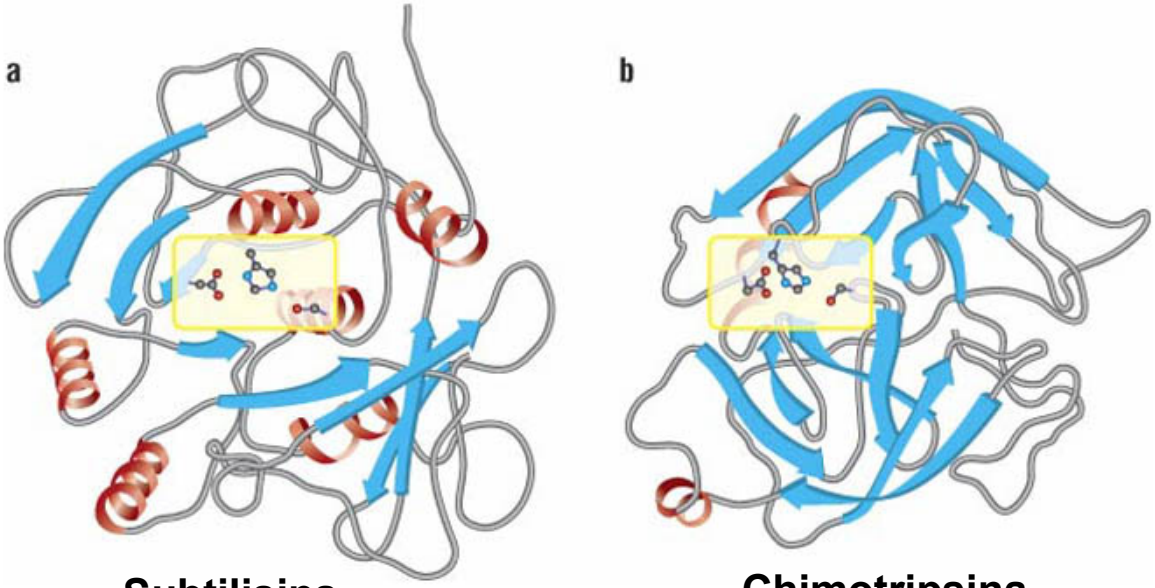
Presenza di "motivi o segnali"

"Motivi/Segnali" nelle proteine

Motivo di sequenza	
:=	Una porzione limitata della proteina la cui presenza o conservazione dipende da ragioni funzionali o strutturali

La presenza di un motivo in proteine diverse può derivare da omologia o da convergenza funzionale

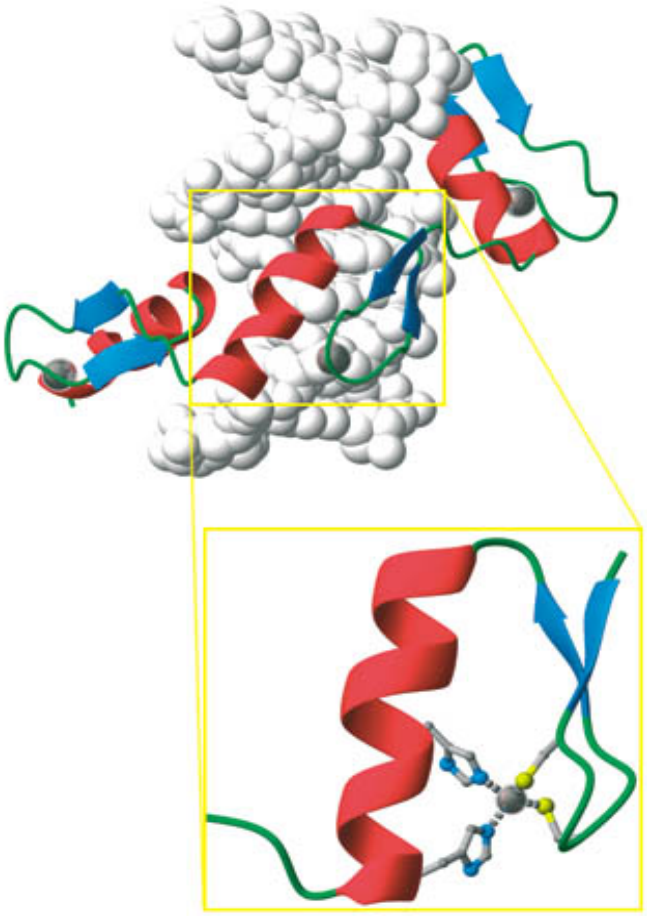
Evoluzione convergente di motivi funzionali



Subtilisina **Chimotripsina**

La triade catalitica di acido aspartico, istidina e serina nella subtilisina, proteasi batterica e nella chimotripsina, proteasi di mammifero. Le due strutture sono radicalmente diverse e i tre amino acidi compaiono con un ordine differente nella sequenza primaria (D,H,S nella subtilisina e H,D,S nella chimotripsina) . Ciò nonostante, l'arrangiamento del sito attivo è molto simile per convergenza evolutiva.

Motivi in sequenza ed in struttura



Zinc finger:

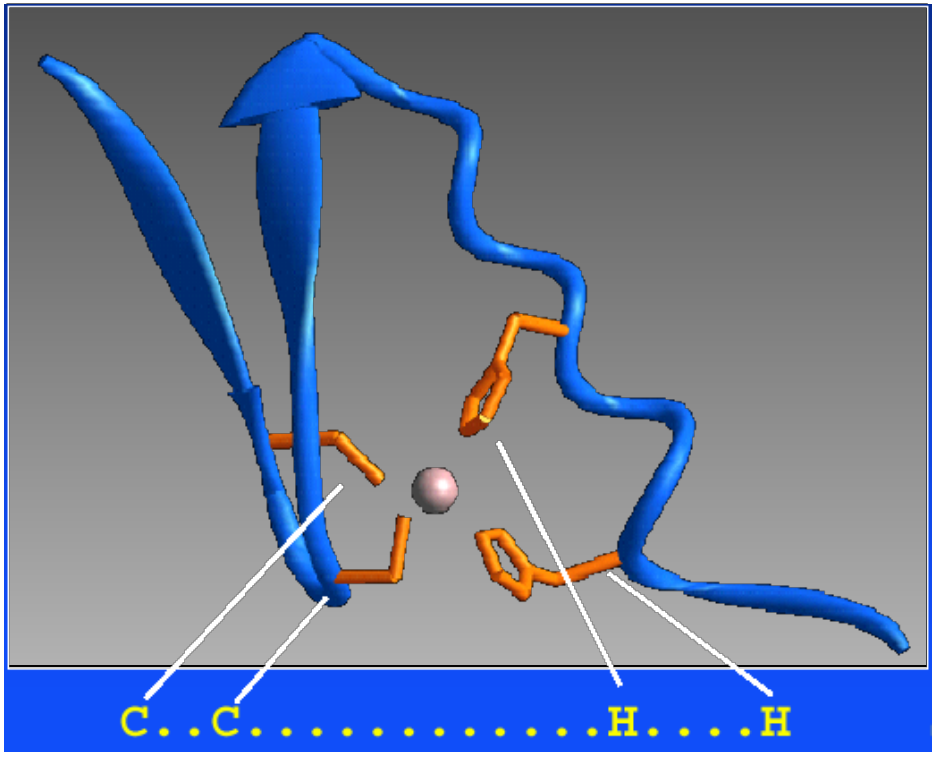
xxxx**C**xx**C**xxxxxxxxxxxxx**H**xxx**H**xxxx

Chimotripsina:

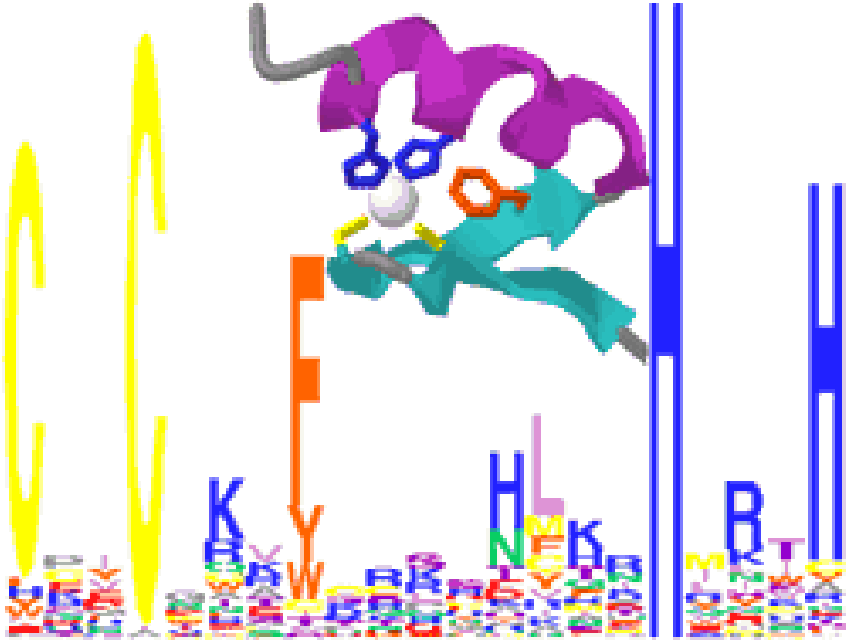
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx**H**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxx**D**xxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxx**S**xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxx

Descrizioni di un classico motivo

C_2H_2 Zinc finger



Sequence pattern



Sequence logos

Prosite patterns

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

Intorno della serina catalitica delle proteasi trypsin-like

G-D-S-G-G

Legame al piridossal-fosfato delle fosforilasi

E-A-[SC]-G-x-[GS]-x-M-K-x(2)-[LM]-N

Zinc Finger

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

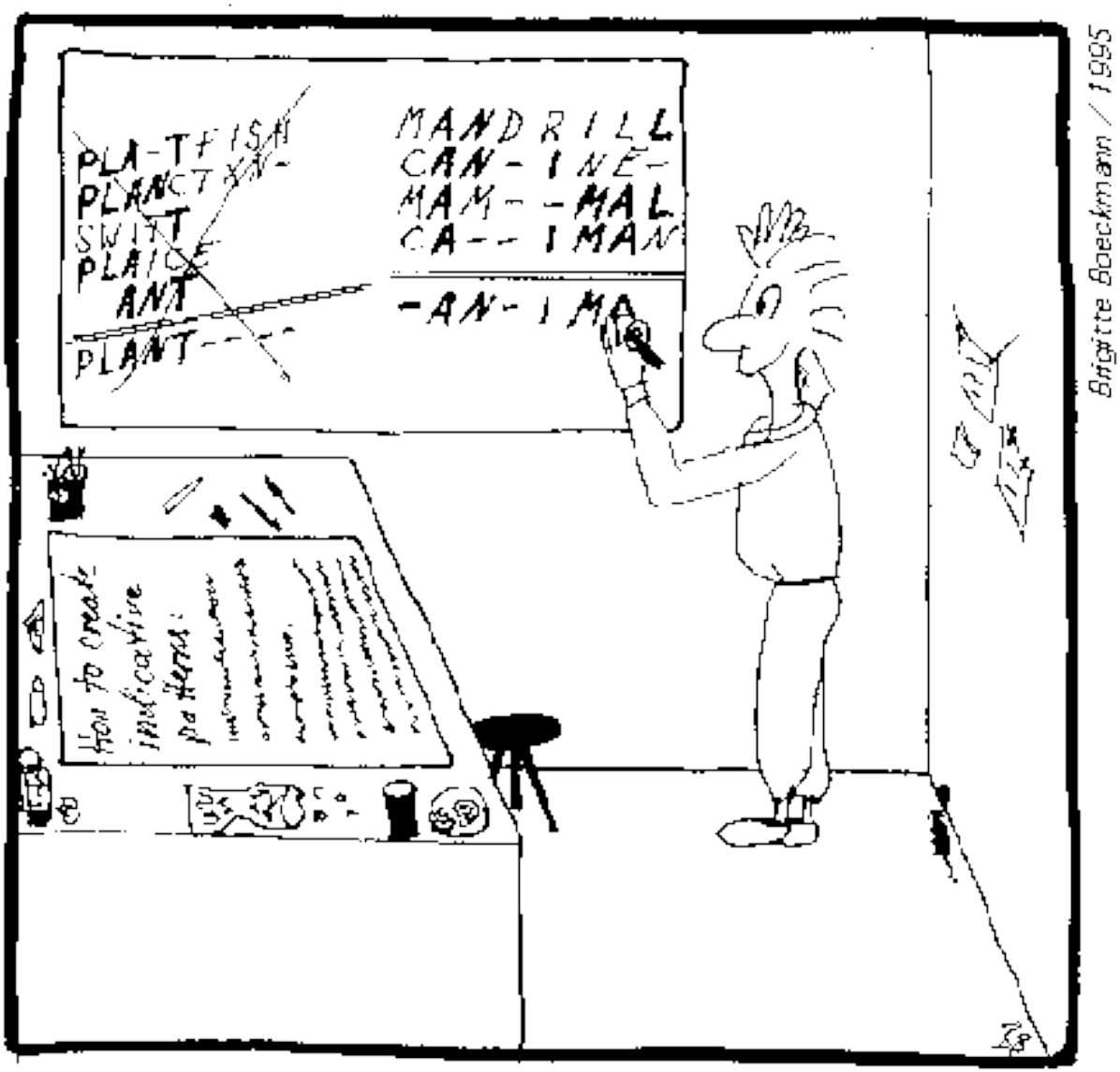
N-Glicosilazione

N-{P}-[ST]-{P}

Homeobox domain

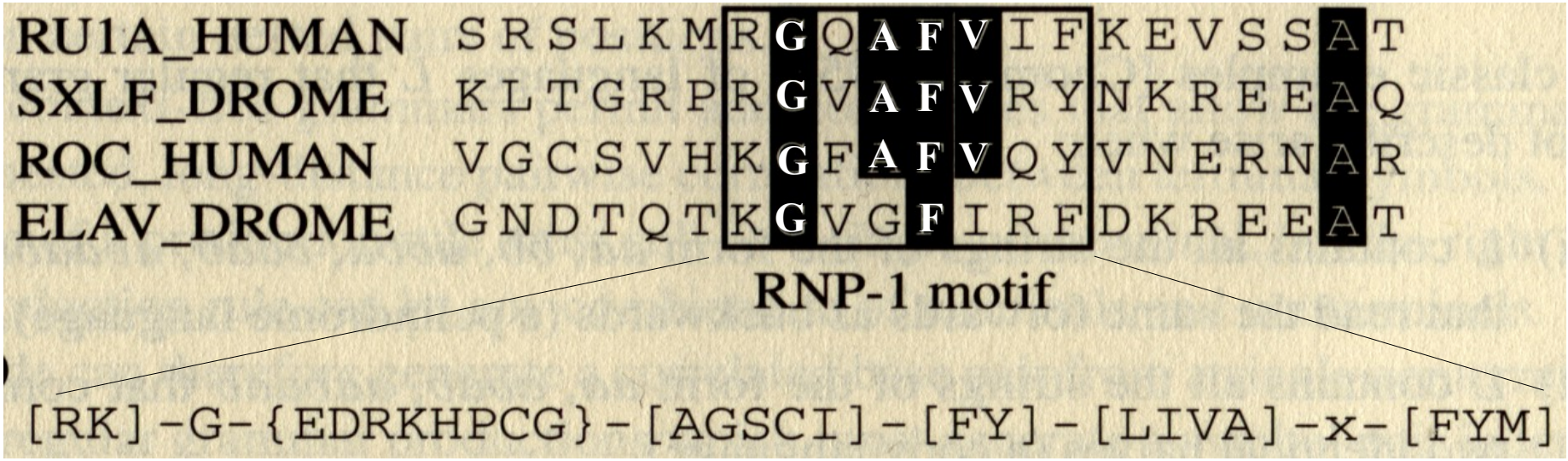
[LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)- [RKNAIMW]

Come si costruisce un Prosite pattern



Come si costruisce un pattern prosite

RNA Recognition Motif (RRM) RNP-1



La presenza di un motivo può essere casuale

NiceSite View of PROSITE: PS00030

General information about the entry	
Entry name	**RRM_RNP_1**
Accession number	**PS00030**
Entry type	PATTERN
Date	APR-1990 (CREATED); DEC-2001 (DATA UPDATE); MAY-2004 (INFO UPDATE).
PROSITE documentation	PDOC00030
Name and characterization of the entry	
Description	Eukaryotic RNA Recognition Motif (RRM) RNP-1 region signature.
Pattern	[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYLM].

Numerical results

- Swiss-Prot release number: **43.4**, total number of sequence entries in that release: **152040**.
- Total number of hits in Swiss-Prot: **587 hits in 470 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **355 hits in 238 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **232 hits in 232 different sequences**

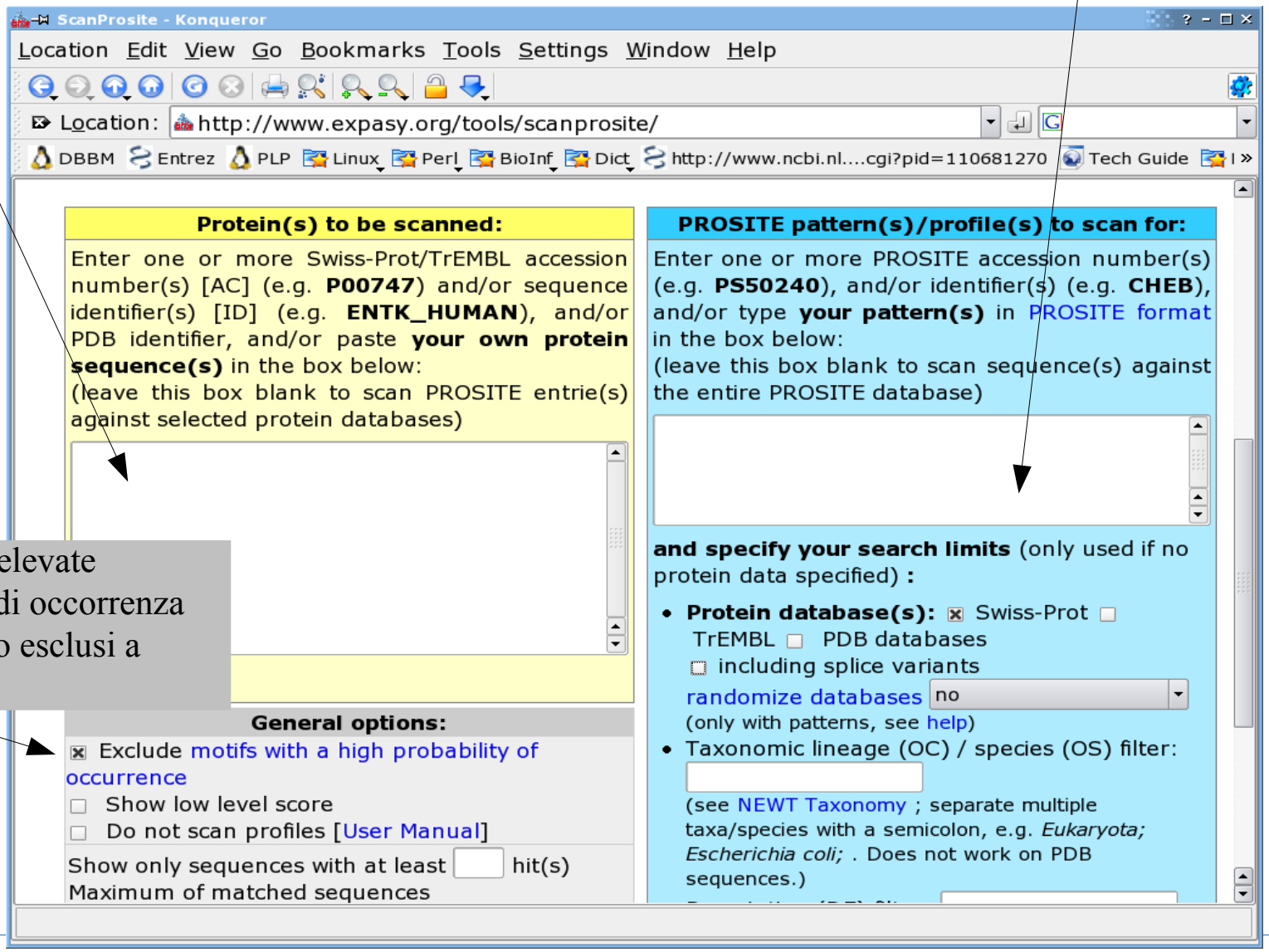
Motivi possono ricorrere nelle proteine per effetto del caso, **senza determinare una data proprietà funzionale**.

Più il motivo è corto e ambiguo, maggiori sono le probabilità di occorrenza casuale

Ricerca di motivi con Scanprosite

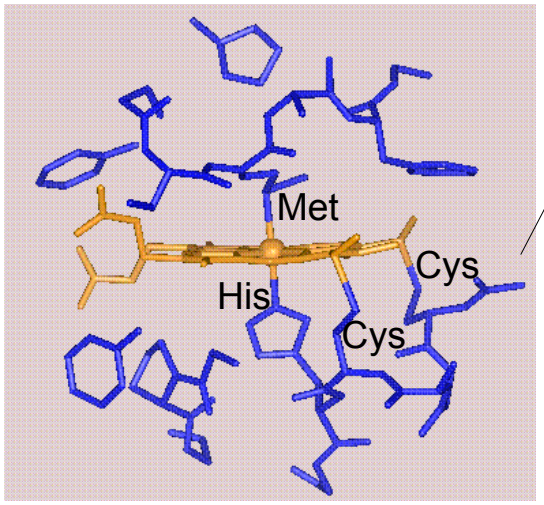
Ricerca motivi contenuti un una sequenza

Ricerca sequenze contenenti un motivo



Motivi con elevate probabilità di occorrenza casuale sono esclusi a default

Conservazione di un pattern in una famiglia di proteine



Cytochrome C Pattern: **C-X-X-C-H**

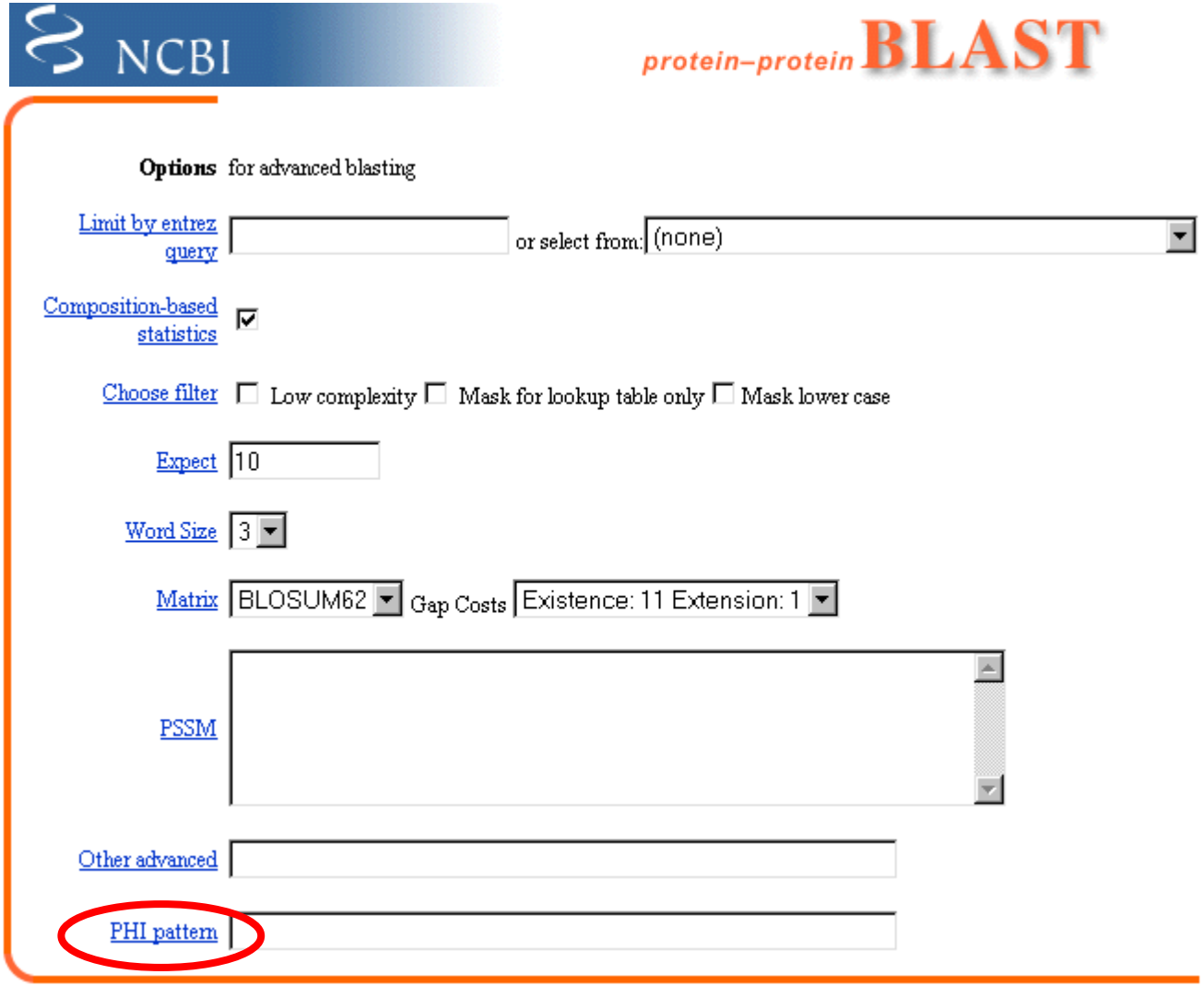
>Atu2314 hypothetical protein Atu2314
MYEY.....AHFTKARDVVQGRCSMCHAAEPVW...

Occorrenza casuale o motivo funzionale?

```
BCAN_A0361 :  RDTVLGRCSMCHAIEPGWEGIIAPPKGVVLDTDRNIAAHAREIYLQAGRSHAMPPANVTG  :  60
mlr5140    :  RDTVLGRCSMCHAQEPVYEGIYHAPKGVMLDTDAEIANHAREIYLQAGRSHAMPPANVSQ  :  60
Atu2314    :  RDVVQGRCSMCHAAEPVWEGVPFTPKSVKLETDEQIAAHAREIYLQAGRSHAMPPGNITA  :  60
Rb_muox    :  MNVVPGRCAMCHSEPYYDGIRRAPKNVLLETPADVAKYAKEIYLQAGATHAMPPANVTF  :  60
Xaut_3294  :  QNIVMGRCSMCHASEPLYEGIVIAPRGVLLDTPEHIARNRDLIVVQAGLTHAMPPNNITE  :  60
Daci_2128  :  APFFQQHCVVCHGAALQQK-------NVRLDSPEQVKAHAQQIYQQVVQLRKMPFGNPGA  :  53
BpseN_1318 :  APILQQRCVACHSAKPTLMGS--APAGVMFDTPEEIARNAPRIYQQAVQLKAMPIGNVTH  :  58
PputW619_3 :  HTVIQQRCTVCHSSKPTSPLFSTAPGGVVFDTPQQIQAQAARIQAQAVASQIMPLGNITQ  :  60
                    rC  CH   p   g    p  V ld3   6a  a  Iy Qa    aMP  N
```

Un profilo di conservazione coerente con il pattern suggerisce la presenza non casuale del motivo di sequenza

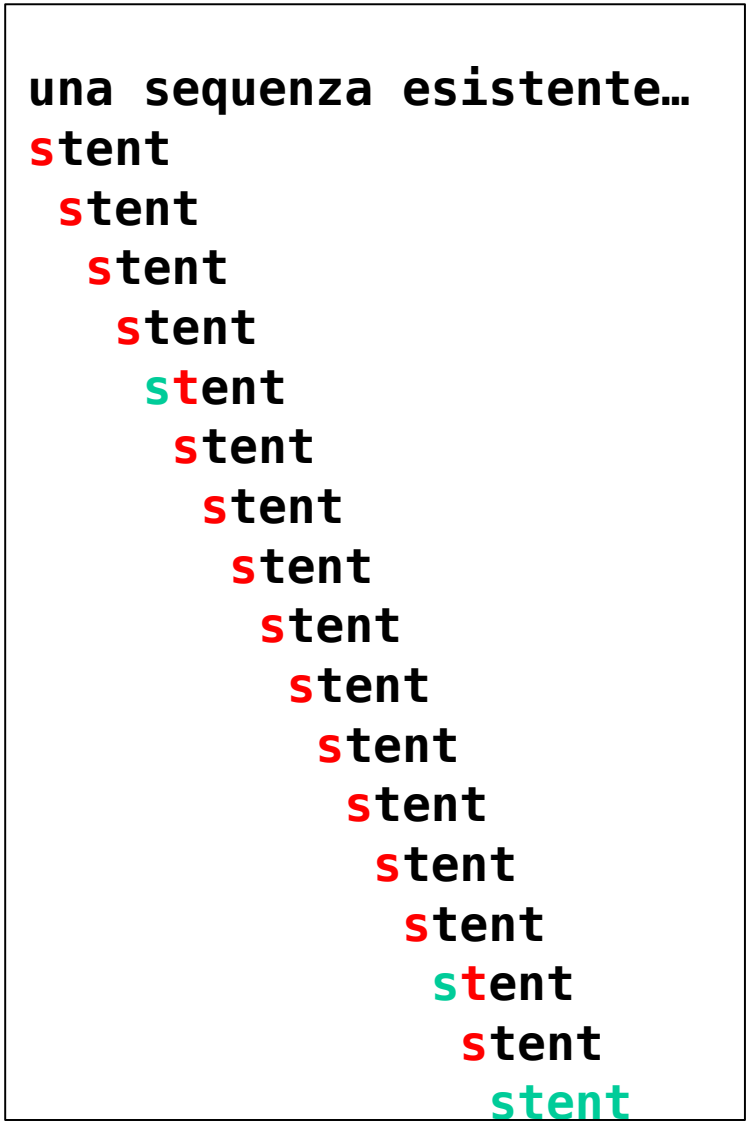
PHI Blast



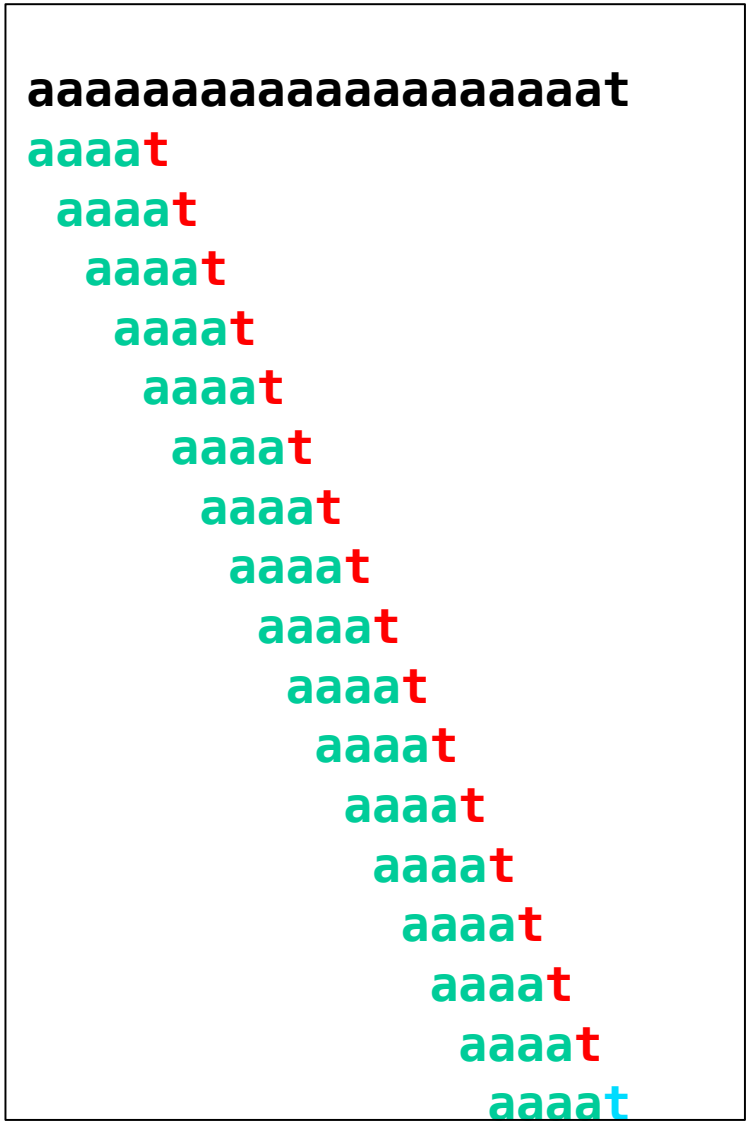
Pattern Hit Initated Blast trova sequenze che hanno somiglianza locale ad una data sequenza in input ed in più contengono un motivo specificato in sintassi Prosite

Ricerca di una stringa “brute force”

Caso tipico m + n confronti



Caso peggiore m*n confronti



Ricerca Boyer-Moore

Caso tipico: m/n + n confronti

```
una sequenza esistente…
stent
    stent
      stent
           stent
                stent
una sequenza esistente…
```

Provare prima l'ultima lettera. Se non va bene spostare la stringa in modo da far corrispondere le lettere o del tutto se la lettera non è presente nella sequenza da cercare.

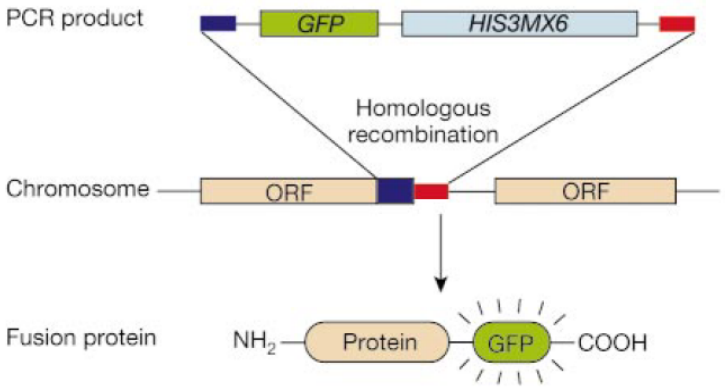
Caso peggiore: m confronti

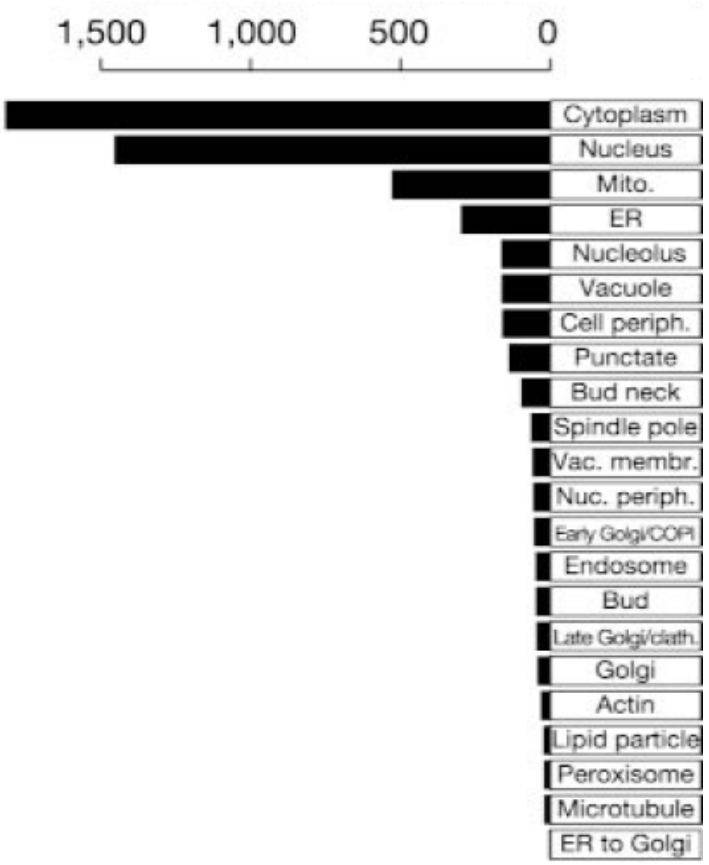
```
aaaaaaaaaaaaaaaaaaaa
taaaa
     taaaa
          taaaa
               taaaa
aaaaaaaaaaaaaaaaaaaa
```

	Caso tipico	Caso Peggiore
Brute force	m+n	m*n
Boyer Moore	m/n + n	m

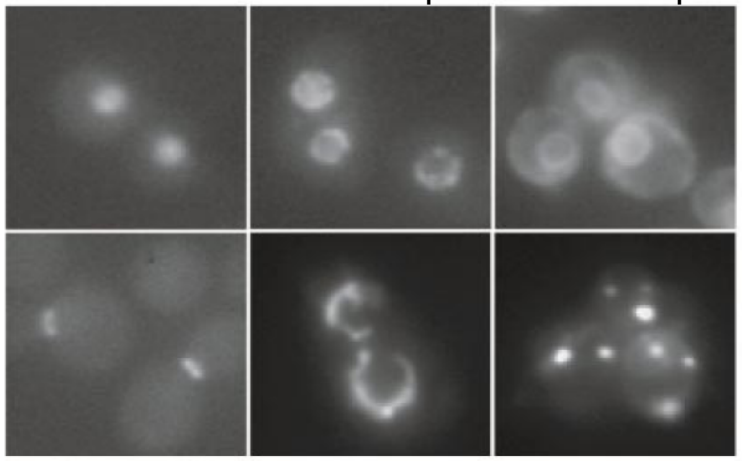
localizzazione cellulare delle proteine

Global analysis of protein localization in budding yeast.
Huh et al Nature 2003





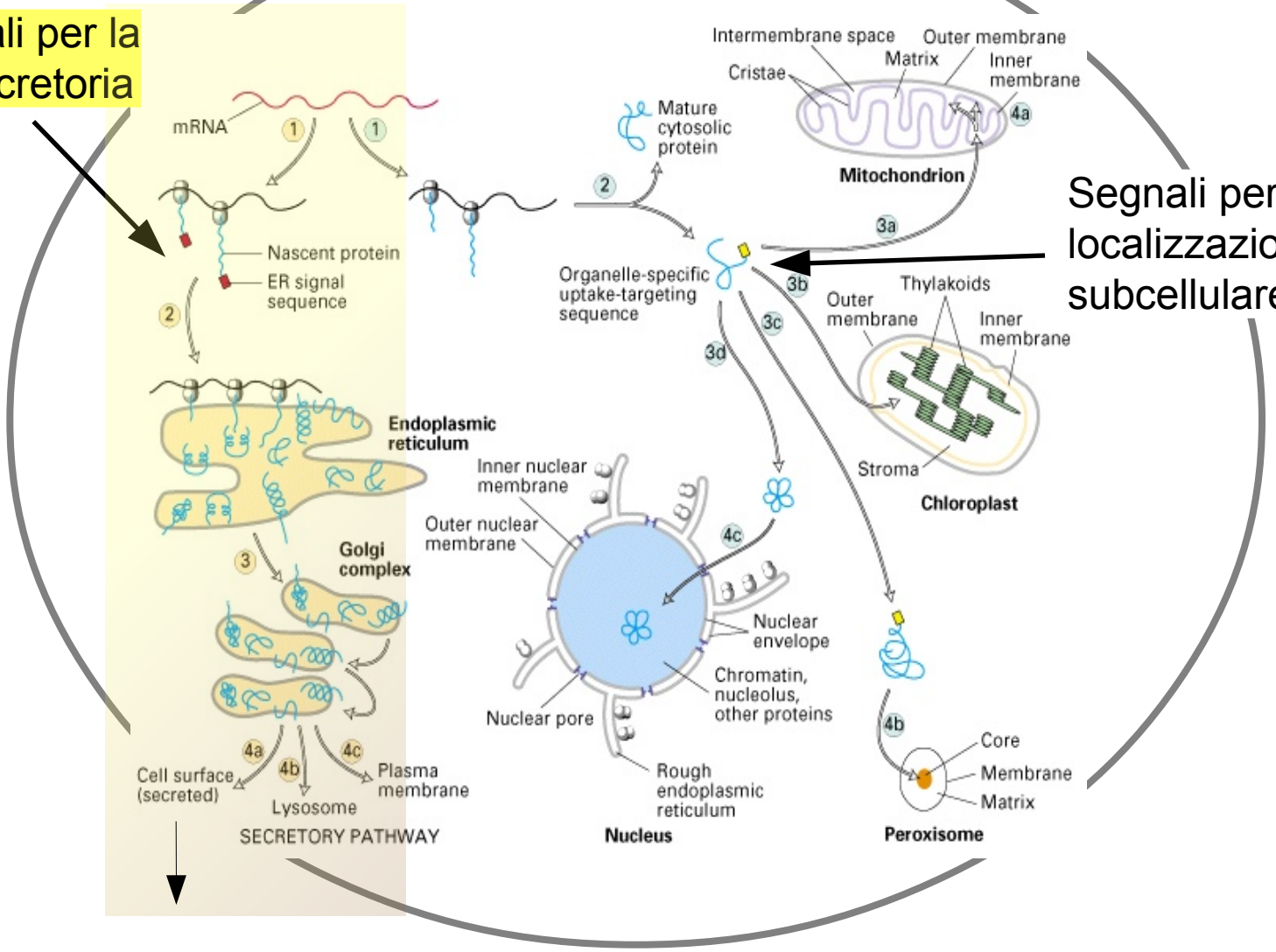
nucleo | nucleo perif. | r. endoplasm



Bud neck | mitocondrio | liposoma

Subcellular protein sorting: via secretoria

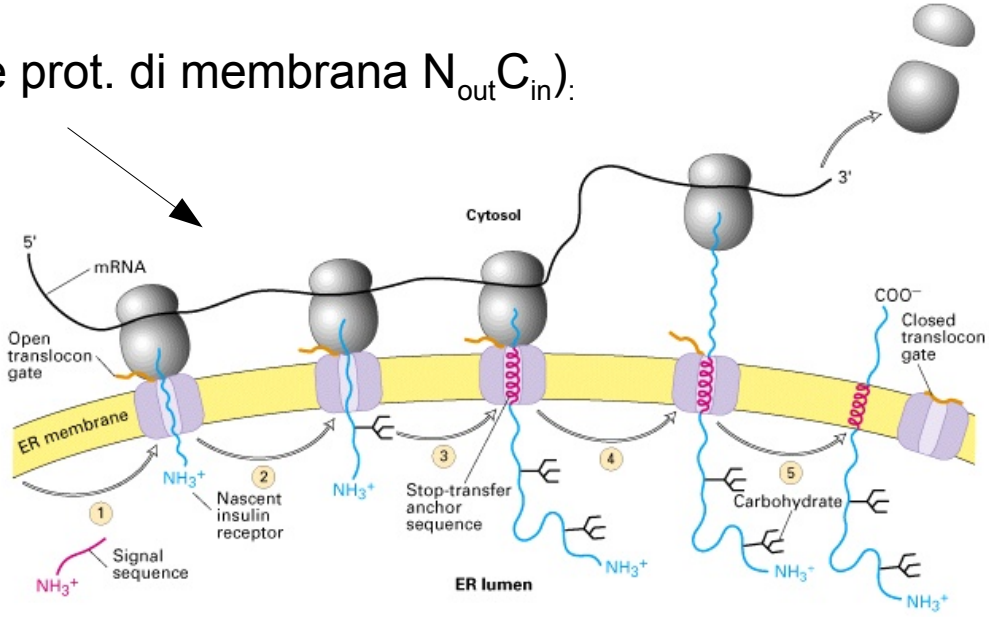
Segnali per la via secretoria

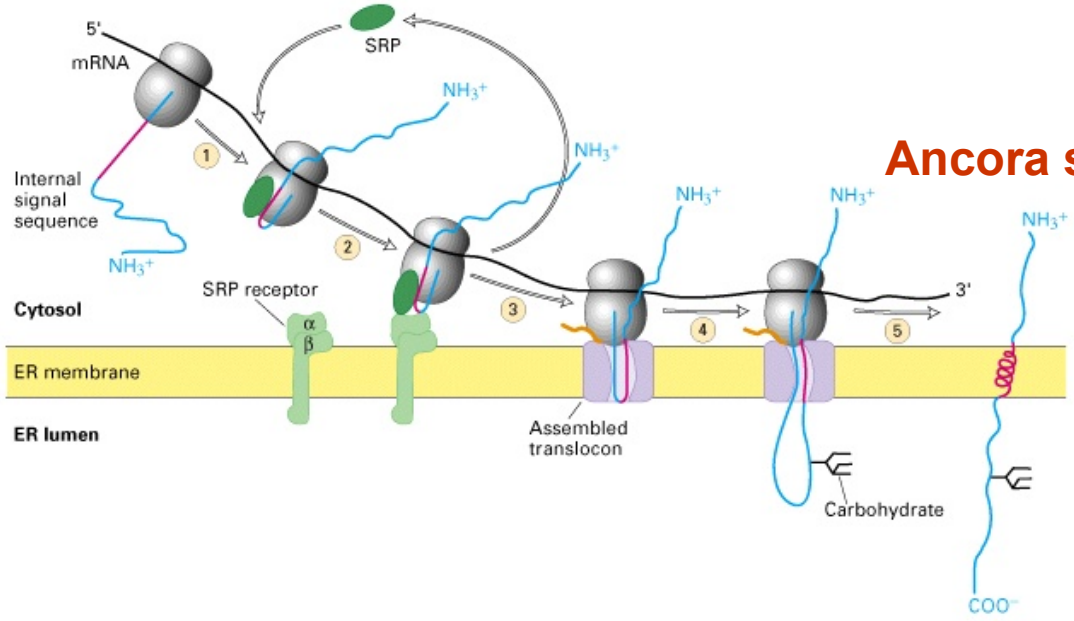


Segnali per localizzazione subcellulare

Peptide segnale (proteine secrete e prot. di membrana $N_{out}C_{in}$):

Le proteine secretorie entrano nel reticolo endoplasmico grazie alla presenza di una sequenza segnale all'estremità N-terminale che viene successivamente tagliata ("signal peptide"). Anche le proteine di membrana con una topologia $N_{out}C_{in}$ posseggono una sequenza segnale che viene tagliata nel reticolo

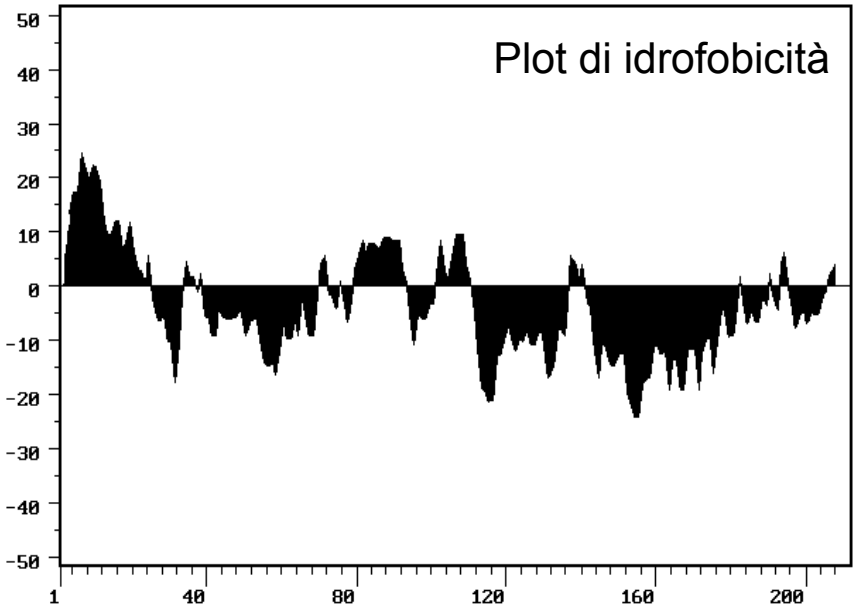




Ancora segnale (Prot. di membrana $N_{in}C_{out}$)

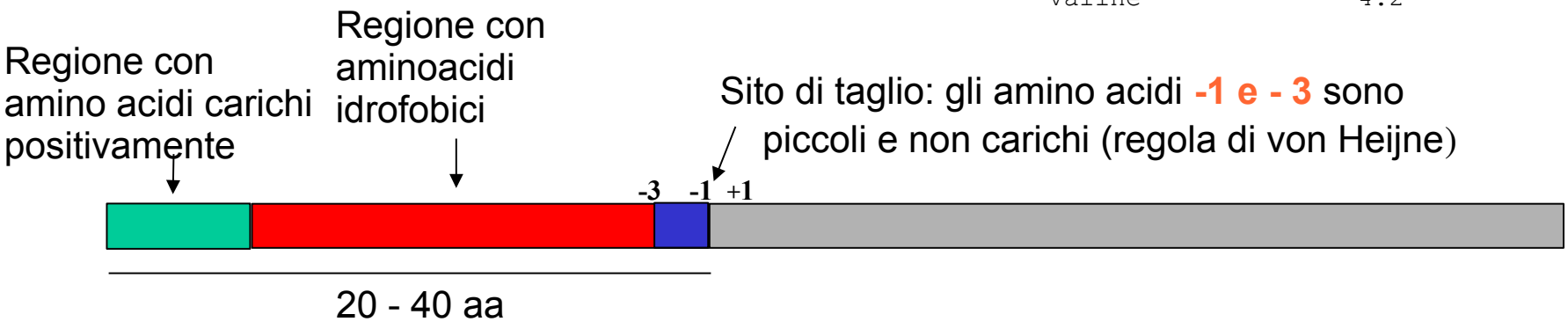
Altre proteine di membrana, di solito quelle con topologia $N_{in}C_{out}$, non posseggono una sequenza segnale all'N-terminale rimossa ma è la stessa regione transmembrana a funzionare come segnale ("signal anchors")

Peptidi segnale



scala Kyte-Doolittle

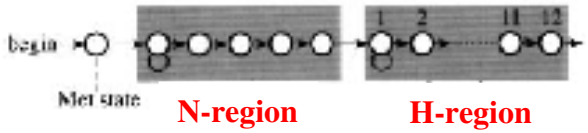
Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2



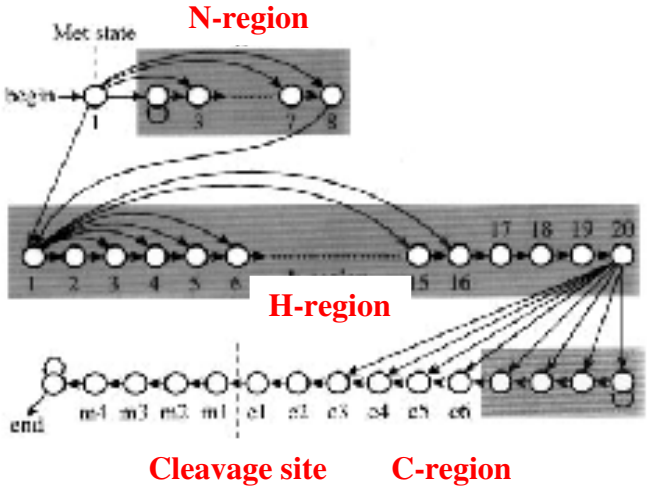
SignalP

Brunak et al

Signal anchor model



Signal peptide model



SignalP è un programma basato su reti neurali e modelli markoviani allenati su set di proteine a destinazione nota per predire signal peptides e signal anchors.

Vengono valutati 3 punteggi:
1)Signal ; 2)Cleavage ; 3)Combined

Il valore massimo del punteggio 3) è predetto come sito di taglio se lo score medio di “Signal” dal primo amino acido fino a quel punto è maggiore di 0.5

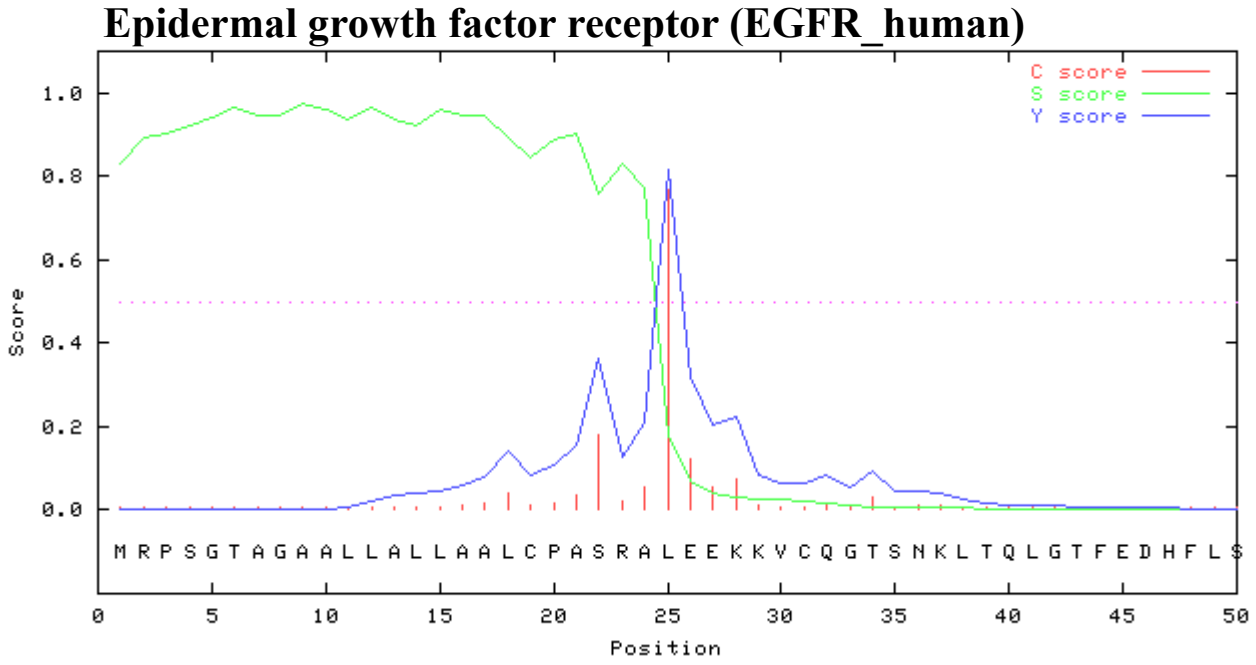
Accuratezza della predizione: ~90%

SignalP 3.0– reti neurali

C-score (raw cleavage site score)

S-score (signal peptide score)

Y-score (combined cleavage site score)



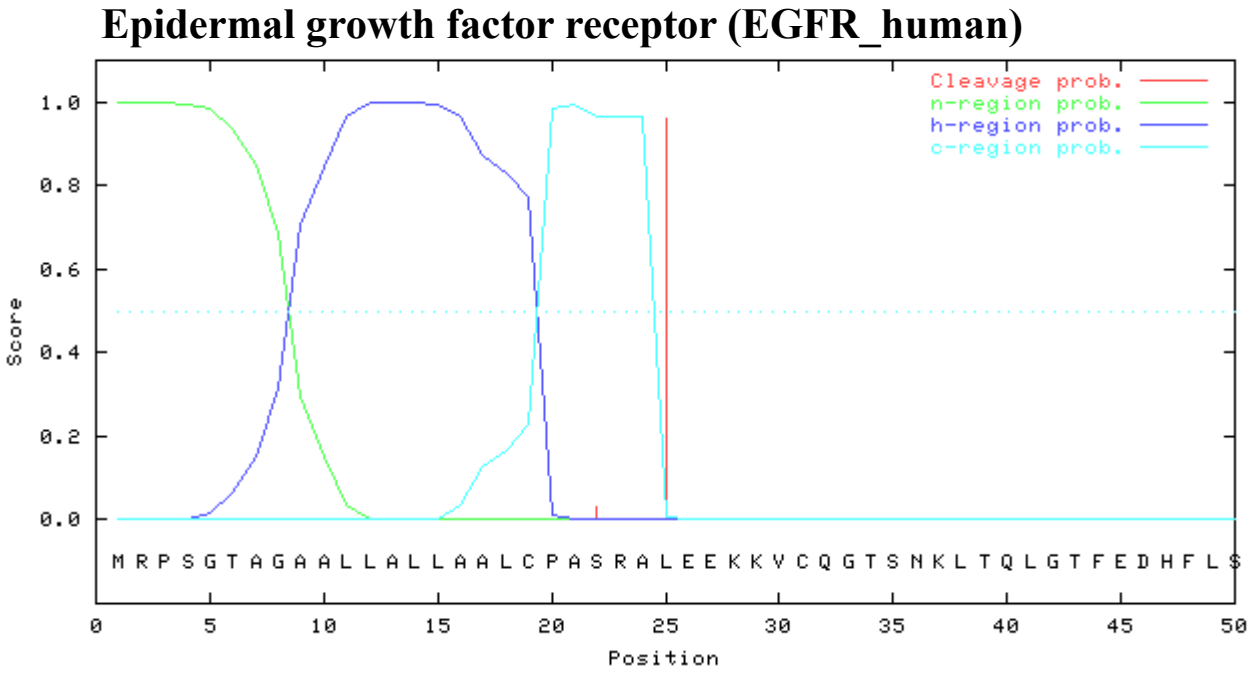
SignalP 3.0 – modelli markoviani

Cleavage probability

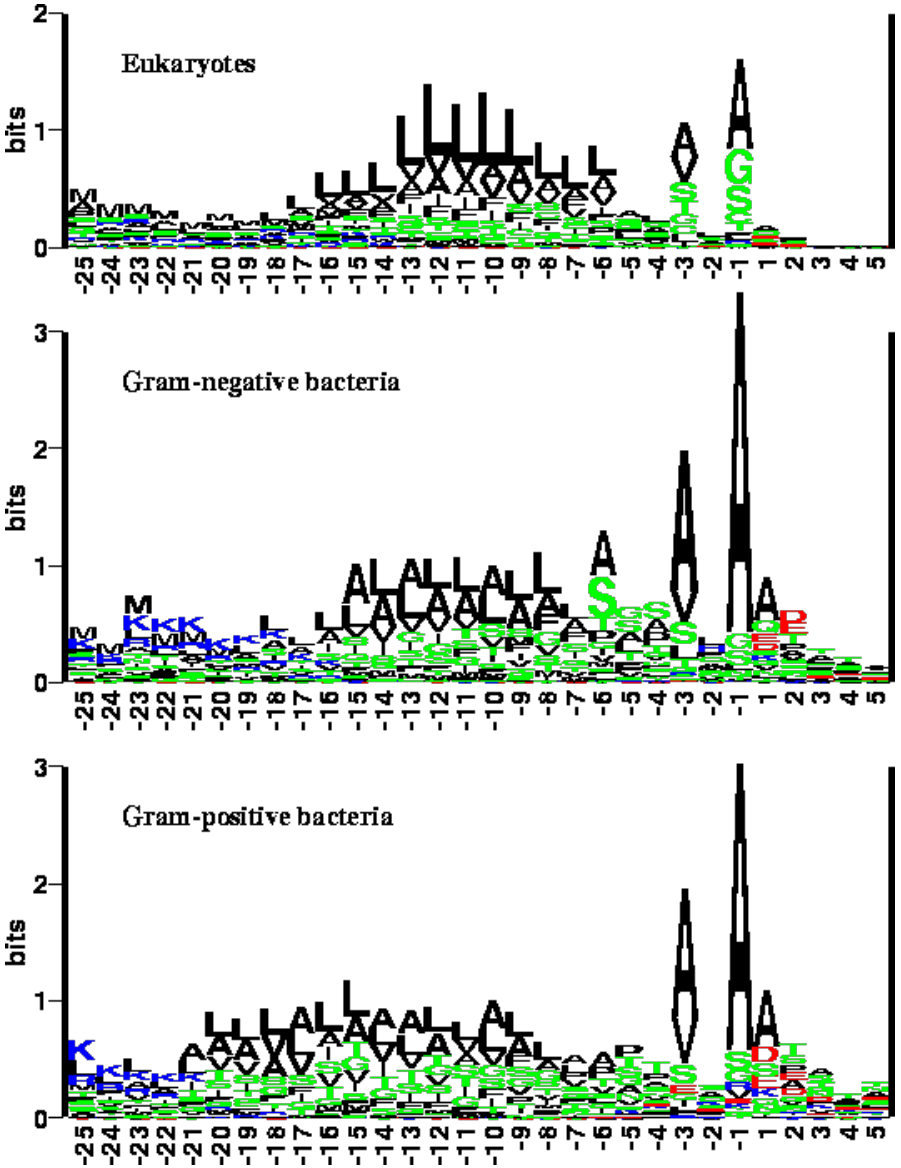
N-region probability

H-region probability

C-region probability



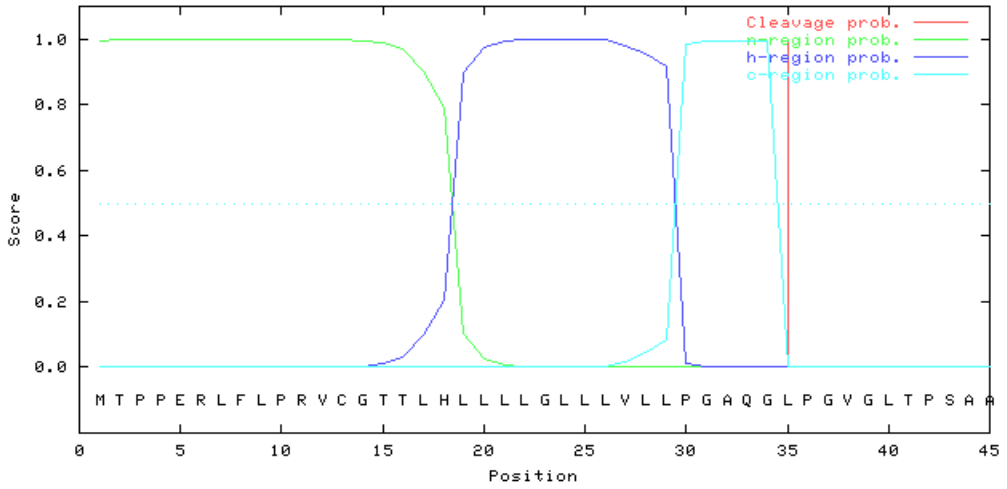
Peptidi segnale in eucarioti e procarioti



Sequence logos dei peptidi segnale in eucarioti, batteri gram negativi e gram positivi. Le sequenze sono allineate rispetto al sito di taglio predetto da SignalP. Le sequenze segnale dei gram positivi sono mediamente più lunghe. La regione idrofobica è dominata da Leu negli eucarioti mentre nei procarioti Leu e Ala sono ugualmente rappresentati. La regola di von Heijne (-1,-3) è seguita sia nei procarioti che negli eucarioti, ma in quest'ultimi Ala non è così prevalente. Residui carichi positivamente all'N-terminale sono soprattutto presenti nei procarioti. Nei procarioti, il primo residuo è sempre N-formil metionina.

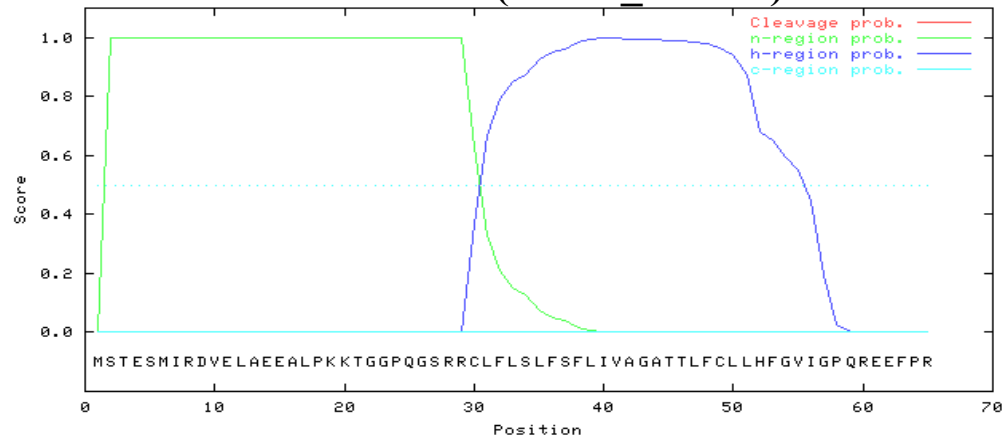
Discriminazione Peptide segnale – Áncora segnale SignalP 3.0

Lymphotoxine alpha (TNFB_human)



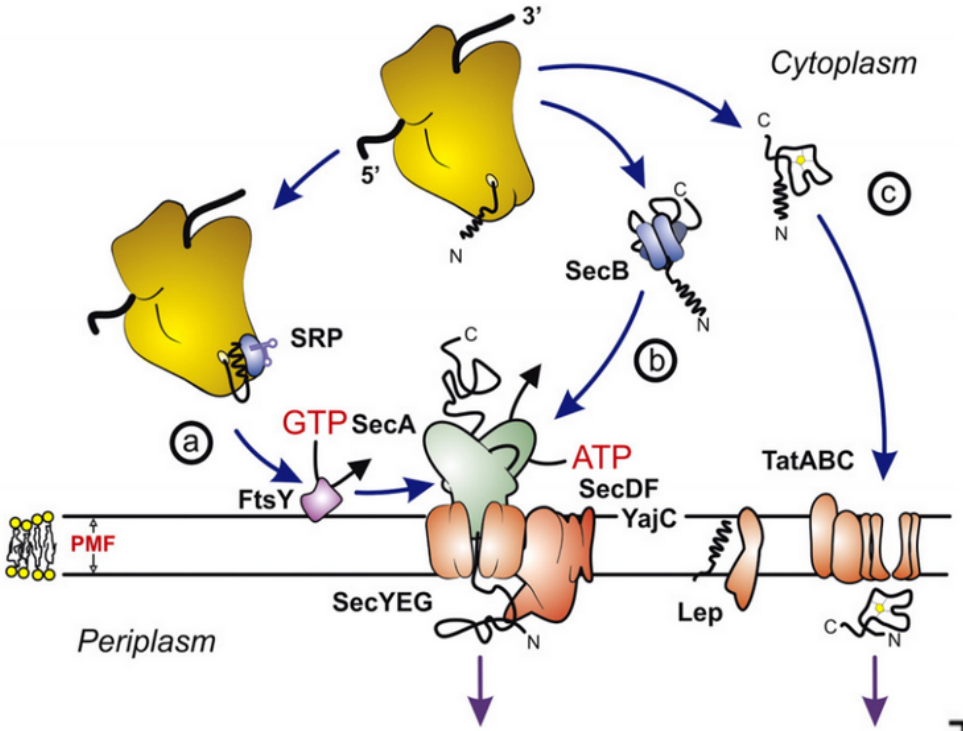
Prediction: **Signal peptide**
Signal peptide probability: 0.996
Signal anchor probability: 0.004
Max cleavage site probability: 0.989
between pos. 34 and 35

Tumor necrosis factor (TNFA_human)



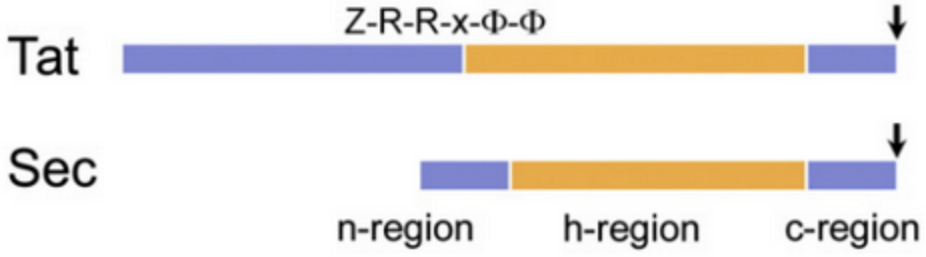
Prediction: **Signal anchor**
Signal peptide probability: 0.001
Signal anchor probability: 0.998
Max cleavage site probability: 0.001
between pos. 46 and 47

Twin-Arginine translocation (TAT)



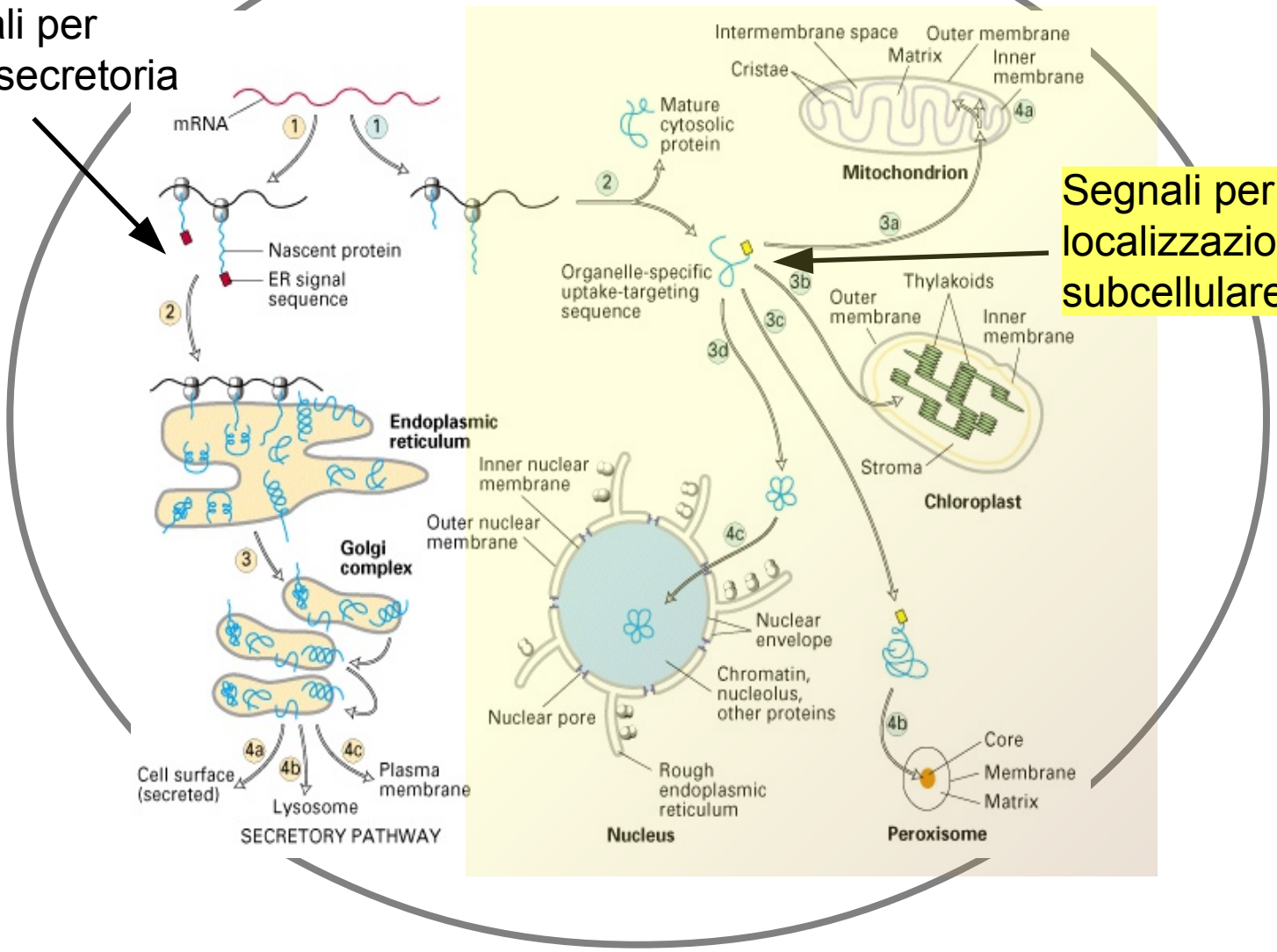
TAT: exporto di proteine 'foldate'.

Peptide segnale TAT con n-region più estesa e coppia di arginine (RR)



Subcellular protein sorting: via citoplasmatica

Segnali per
la via secretoria



Segnali per
localizzazione
subcellulare

Predizione della localizzazione delle Proteine: Psort

Nakai, K. and Kanehisa, M

Psort è un sistema 'esperto' che valuta un insieme di regole per predire il destino cellulare delle proteine. Le regole sono differenti per procarioti ed eucarioti .

Via secretoria

Extra-cellula: nessun (altro) segnale

Reticolo endoplasmico: alcune proteine sono trattenute nel reticolo in presenza di un **segnale di retenzione**: KDEL o HDEL all'estremità carbossi-terminale

Via citoplasmatica

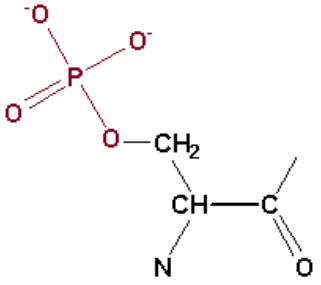
Citosol: nessun segnale

- **Mitocondrio**: un segnale (spesso bipartito) all'**N-terminale** non chiaramente identificabile con un consenso
- **Cloroplasto**: un segnale (spesso bipartito) all'**N-terminale**. I primi 30 aa sono ricchi in Ser e Ala. Il secondo residuo è spesso Ser
- **Nucleo:** Il motivo classico di nuclear localisation signal (**NLS**) segue una delle due regole: 1) 'pat4': **[KR](4)** o **[KR](3)[PH];** 2) 'pat 7' **PX(1,2)[KR](3,4).** Alcune proteine entrano nel nucleo per cotrasporto con altre proteine che hanno un NLS.
- **Perossisoma**: due tipi di segnali di localizzazione (PTS). **PTS1** è un segnale C-terminale (**[SAC][KRH]L)**. Il segnale **PTS2** è N-terminale (**[RK][LI]X(5)[HQ]L)**

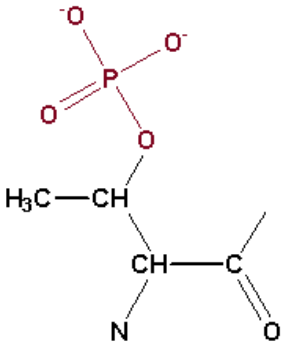
Predizione delle modificazioni post-traduzionali nelle proteine

- **FOSFORILAZIONE**
- **GLICOSILAZIONE**
- **UBIQUITINAZIONE**
- ACILAZIONE
- PRENILAZIONE
- SULFATAZIONE
- SUMOILAZIONE
- ...

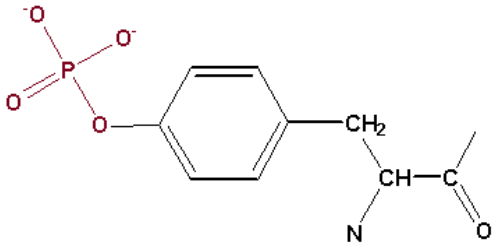
Fosforilazione



fosfoserina



fosfotreonina



fosfotirosina

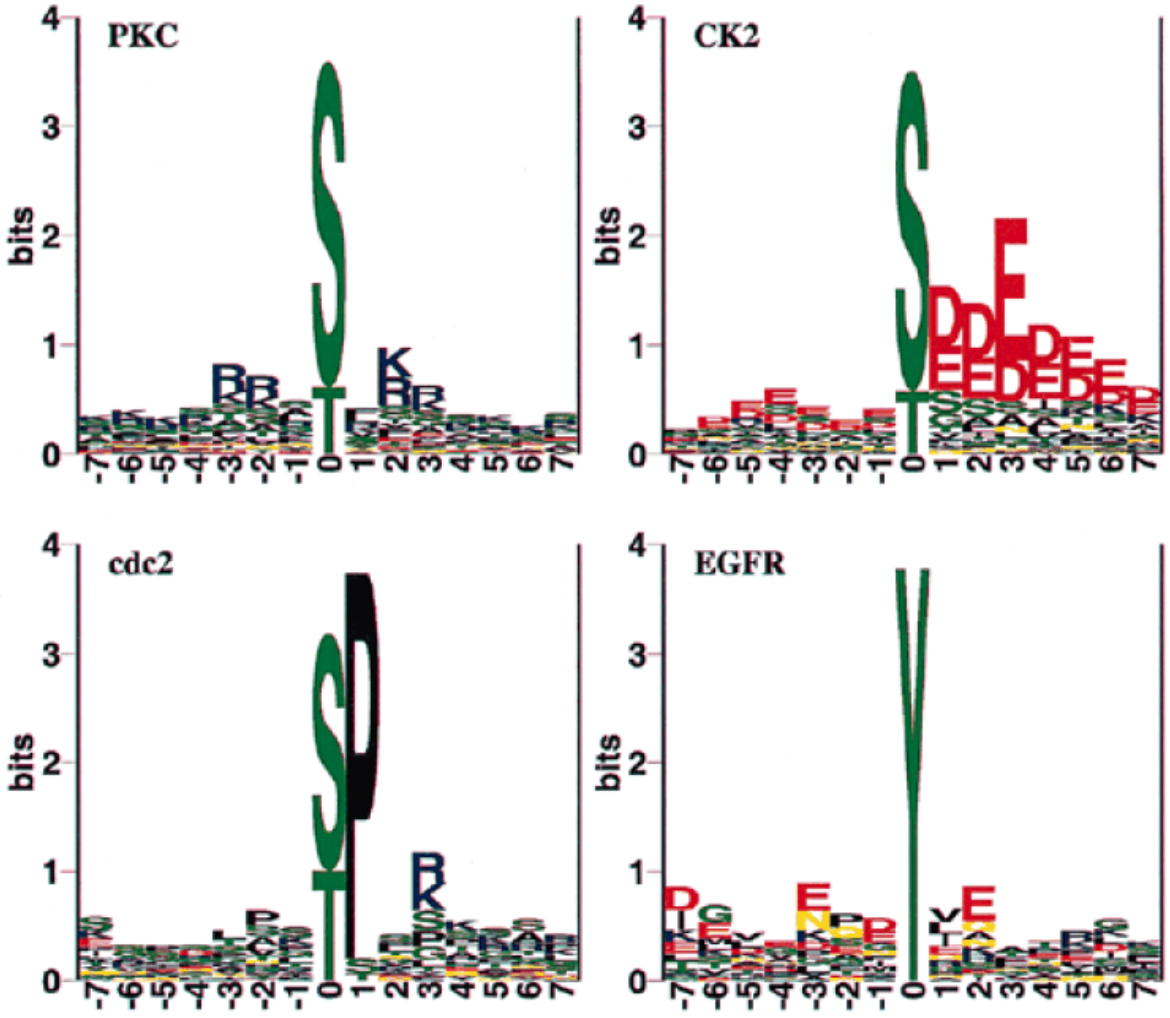
Prosite patterns

[RK](2)-x-[ST] cGMP-dependent protein kinase

[ST]-x-[RK] protein kinase C

[RK]-x(2)-[DE]-x(2,3)-Y Tyrosine kinase phosphorylation site

Phosphobase



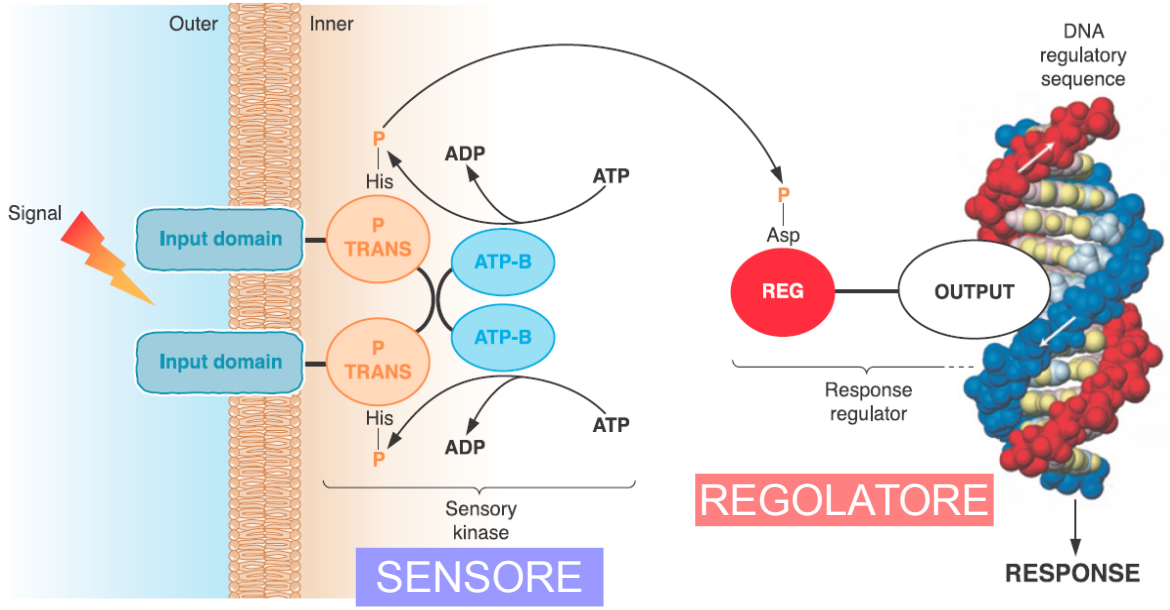
Phosphobase è un database di informazioni sui sui residui fosforilati in proteine. L'allineamento delle sequenze attorno al sito fosforilato mostra come cambia il contenuto in amino acidi attorno al residuo nei bersagli delle diverse chinasi. Le chinasi acidofile (CK2) hanno una serie di D e E dopo il residuo fosforilato. Le chinasi dirette da prolina (cdc2) hanno sempre prolina dopo il residuo fosforilato. Il programma basato su Phosphobase, **NetPhos** ha una percentuale di successo tra il 69% e il 96% per le diverse chinasi.

His and Asp phosphorylation

Protein Phosphorylation in Bacteria

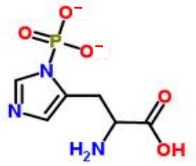
Adding or removing phosphate groups of bacterial proteins may change metabolic activity and sometimes confers virulence

Ivan Mijakovic

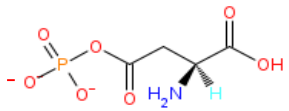


A typical two-component system.

phospho-histidine



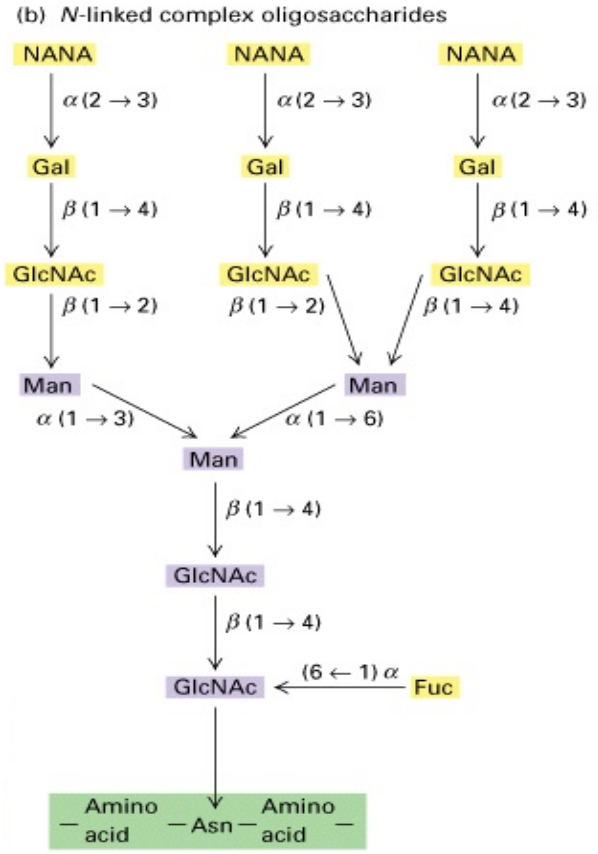
phospho-aspartate

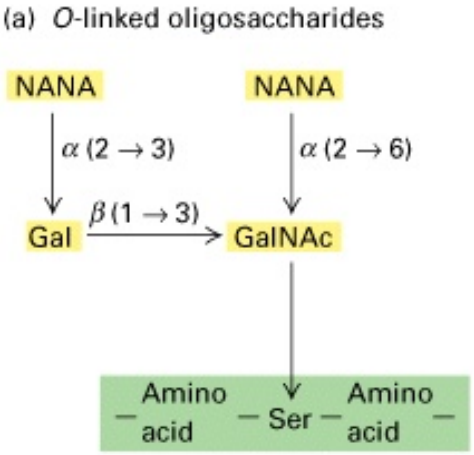


Glicosilazione

NANA = *N*-Acetylneuraminic acid (sialic acid)
GalNAc = *N*-Acetylgalactosamine
GlcNAc = *N*-Acetylglucosamine (conserved)
GlcNAc = *N*-Acetylglucosamine

Gal = Galactose
Man = Mannose
Fuc = Fucose
Glc = Glucose





Glicosilazione all'N

Asparagina
Esempio pattern: N-{P}-[ST]-{P}

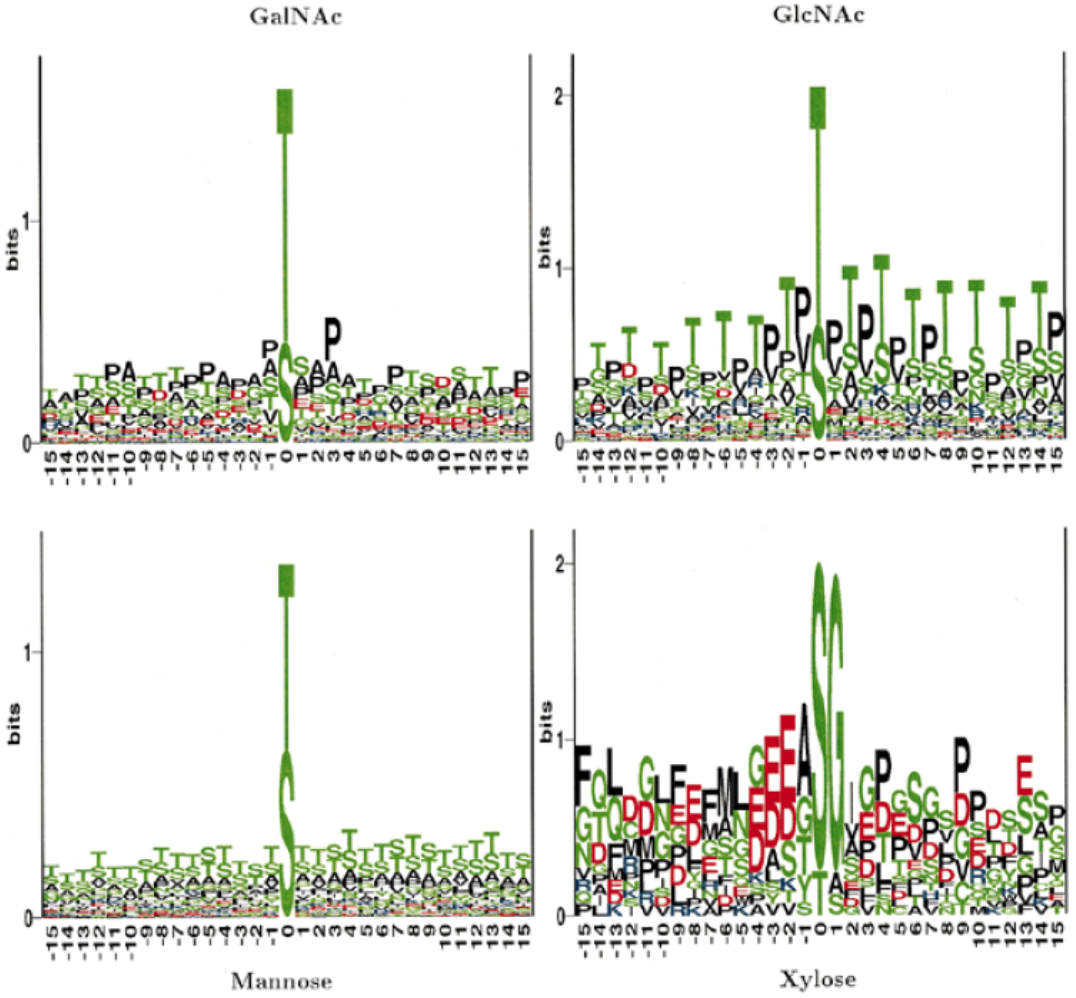
Glicosilazione all'O

Serina/Treonina
Esempio pattern: S-G-X-G (proteoglicani)

O-Glycbase

O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins

Ramneek Gupta*, Hanne Birch, Kristoffer Rapacki, Søren Brunak and Jan E. Hansen



O-linked sugar	Abbreviation	Number of sites in O-GlycBase 4.0
N-Acetylgalactosamine	GalNAc	680 (194 Ser, 486 Thr)
N-Acetylglucosamine	GlcNAc	88
Mannose	Man	158
Xylose	Xyl	16
Glucose	Glc	11
Fucose	Fuc	7
Others/Unspecified	-	31

Non vi è un consenso evidente attorno alle serine e alle treonine conservate. Si osserva solo una abbondanza di T, P e S nelle regioni fiancheggianti il residuo conservato e che per per GalNAc, GlcNAc e mannoso i residui carichi sono sfavoriti in posizione -1 e +3. Solo nel caso dello xiloso (proteoglicani) si ha un consenso chiaro, simile a quello descritto in prosite. Il programma predittivo **NetOGlyc,** basato su O-Glycbase ha una percentuale di successo attorno all'85%.

Segnali per il turnover delle proteine

Half-life (hours)	Intracellular Location: Nucleus	Cytosol	Mitochondria	Endoplasmic Reticulum and Plasma Membrane
<2	Oncogene products	Ornithine decarboxylase, tyrosine aminotransferase, protein kinase C	δ-Aminolevulinic acid synthetase	HMG-CoA reductase
2–8	—	Tryptophan oxygenase, cAMP-dependent protein kinase	—	γ-Glutamyl transferase
9–40	Ubiquitin	Calmodulin, glucokinase	Acetyl-CoA carboxylase, alanine aminotransferase	LDL receptor, cytochrome P450
41–200	Histone H1	Lactate dehydrogenase, aldolase, dihydrofolate reductase, phytochrome P670	Cytochrome oxidase, pyruvate carboxylase, cytochrome *c*	Cytochrome b_5, cyt b_5 reductase
>200	Histones H2A, H2B, H3, H4	Hemoglobin, glycogen phosphorylase	—	Acetylcholine receptor

Source: From M. Rechsteiner, S. Rogers, and K. Rote, *Trends Biochem. Sci.* (1987) 12:390–394. © 1987 with permission from Elsevier Science.

Note: This table represents just a few examples of the many proteins whose half-lives have been determined in different organisms.

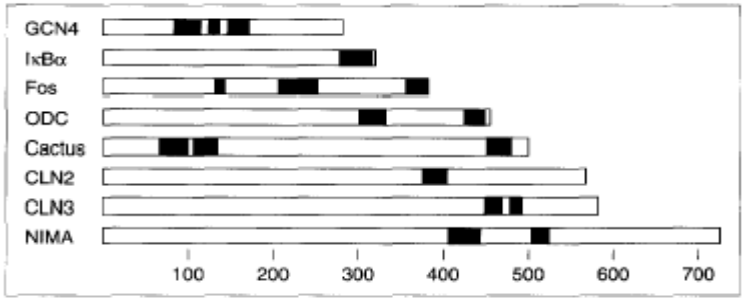
La degradazione è uno strumento importante di regolazione. Il tempo di vita media delle proteine varia da minuti a giorni. Specifici segnali determinano il tempo di vita delle proteine, attraverso l'**ubiquitinazione** e l'azione successiva di **complessi proteosomali** o attraverso il trasferimento al **lisosoma**

PEST

Rogers and Rechsteiner, 1986

Le proteine a vita media corta $t_{1/2}$ <2 ore hanno di solito regioni arricchite in prolina, glutammato, serina, treonina (PEST) e aspartato (D). Le regioni sono di solito fiancheggiate da K,R o H
Segnale: **[KRH]-([P]-[ED]-[ST])(n)-[KRH]**

Queste sequenze tendono a essere presenti in modo significativo all'estremità carbossi-terminale delle proteine a vita media corta.



Distribuzione di sequenze arricchite in P,E,D,S,T in proteine rapidamente degradate.

Table I. PEST sequences known to function as proteolytic signals

Protein	Sequence	PEST-FIND score
GCN4	KTVLPIPELDDAVVESFFSSSTDSTPMFEYENLEDNSK	5.3
	KENTSLFDNDIFVTTDDVSLADK	1.7
	KAIESTEEVSLVPSNLEVSTTSFLPTPVLEDAK	3.5
IκBα	RIQQQLGQLTLENLQMLPESEDEESYDTESEFTEFTEDEL-PYDDCVFGGQR	5.9
Fos	KVEQLSPEEEEK	10.1
	KIPDDLGFPEEMSVASLDLTGGLPEVATPESEEAFT-LPLLNDPEPK	5.7
	GSSSNEPSSDSLSSPTLLAL	4.6
Ornithine decarboxylase	KEQPCSDDEDESNEQTFMYYVNDGVYGSFNCILYDH	3.7
	HGFPPEVEEQDDGTLPNSCAQESGMDR	5.2
Cactus	KEFAVFNETSDSGFISGPQSSQIFSEEIVPDSEPQDK	7.7
	KPQPVVLDGGIIDEEEDQEEQEK	10.4
	RGAETVTFPDSDYDSSDIEDLDDTK	17.9
CLN2	KLTISTPSCSFENSNSTSIPSPASSSQSH	7.0
CLN3	KDSISPPPAPTPTSSSSSPSPPNSPYK	8.2
	KTSSSMTTPDSASH	10.6
NIMA	KSSQMSSSNSEDSDFPSSTDISQLSLESPTNK	12.6
	KFEPTLAYSDDEDDTPELPSPTR	16.4

La presenza di possibili sequenze PEST nelle proteine viene predetta con il programma **PEST-find**

Caratteristiche fisico-chimiche

ProtParam

Data una proteina a sequenza nota è possibile calcolare:

- **Massa molecolare (Mw):** somma della massa dei singoli aa. Espressa in Dalton (Da).
- **Composizione in aminoacidi:** frequenza di ogni amino acido.
- **Punto isoelettrico (pI):** pH al quale la proteina è priva di carica netta. Corrisponde al pK del peptide. Calcolato in base al pKa dei singoli aminoacidi
- **Coefficiente di estinzione (E):** Assorbimento della luce a 280 nm di una soluzione 1 molare della proteina attraverso 1 cm di cammino ottico (M^{-1} cm^{-1}). Permette di calcolare l'assorbimento delle proteine conoscendo la concentrazione o di calcolare la concentrazione conoscendo l'assorbimento. Il valore di E dipende soprattutto dal numero di triptofani e tirosine.