

Descubrimiento de Motivos en Secuencias

Luis Garreta

Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

14 de abril de 2021

Motivos de Secuencias

Secuencias conservadas de **patrones similares o identicos**

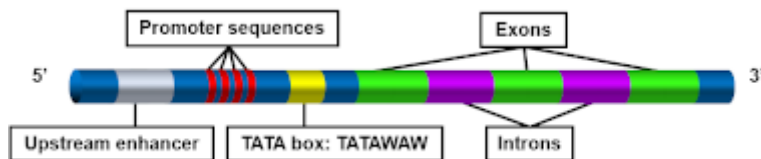
En bioinformática, es un **patrón** de secuencia de ADN/ARN/aminoácidos que está **ampliamente conservado** y se ha probado o se asume que tiene un **significado biológico**.



Una vez conocemos el patrón de secuencia del motivo, entonces podemos usar diferentes métodos de búsqueda para localizarlo en la secuencia

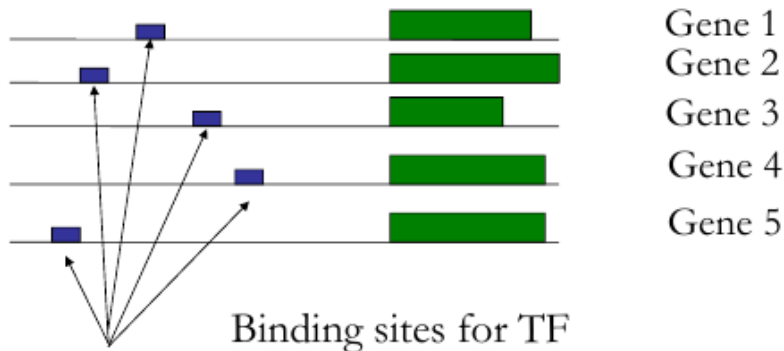
Caja TATA

La caja TATA es una secuencia de ADN que se encuentra en la región central de los promotores de eucariotas.

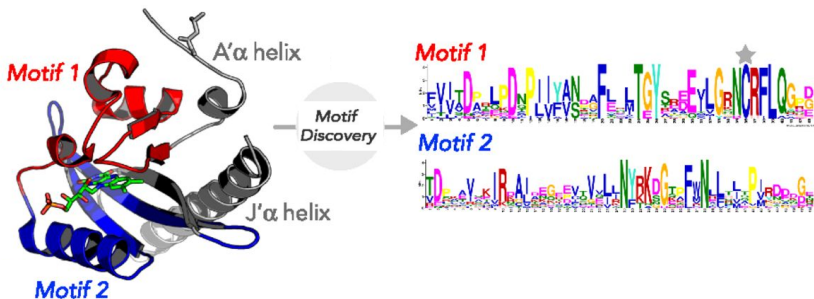


Factores de Transcripción (TF)

Otro ejemplo de motivos: Sitios de unión para los factores de transcripción (TF) cercanos a las regiones promotoras de los genes



Dominios, Familias, y Sitios Funcionales de Proteínas



Algunas notaciones para Motivos (Prosite DB)

Un ejemplo de algunos motivos en una proteína:

Seq1:	N	A	R	S	R	G
Seq2:	N	R	R	S	D	G
Seq3:	N	C	R	T	D	G
Seq4:	N	Q	R	T	L	G
Seq5:	N	H	R	S	A	G
Seq6:	N	R	R	S	C	G
Seq7:	N	I	R	T	F	G
Seq8:	N	K	R	T	Q	G
....

N - x - R - [ST] - x - G

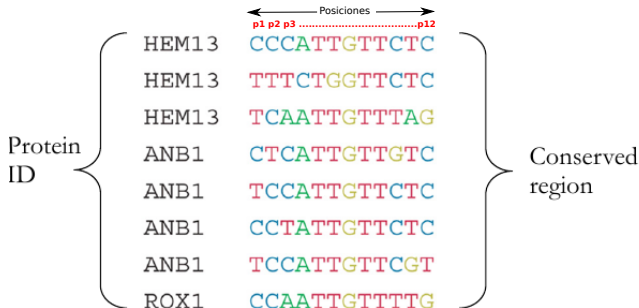
Notación PROSITE: expresiones regulares (RE):

- - : Separador
- N: Inicia con N,
- x: Seguido de cualquier amino
- R: Seguido del amino R
- {P}: Cualquier amino excepto P
- [ST]: Seguido por el amino S o T
- G: Termina en el amino G

Descubrimiento de Motivos basados en Alineamientos

Análisis de perfiles: generalmente sigue estos pasos:

1. Construir un alineamiento local de múltiples secuencias
2. Aislar las regiones conservadas basado en los puntajes de alineamiento:



Descubrimiento de Motivos basados en Alineamientos

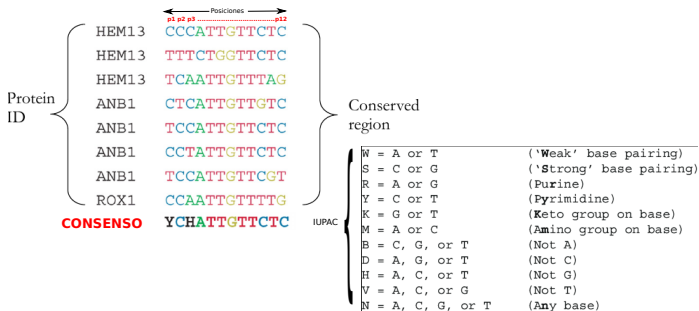
Después de que las regiones altamente conservadas son aisladas, se usan para construir **matrices de perfiles** para cada región.

Una **matrix de perfil** para un motivo dado contiene los conteos de frecuencias para cada letra en cada posición de la región conservada aislada

	←----- Posiciones -----→												
	p1	p2	p3	p12	
A	0	0	2	7	0	0	0	0	0	0	0	1	0
C	4	6	4	1	0	0	0	0	0	5	0	5	0
G	0	0	0	0	0	1	8	0	0	1	1	2	0
T	4	2	2	0	8	7	0	8	8	2	6	1	0

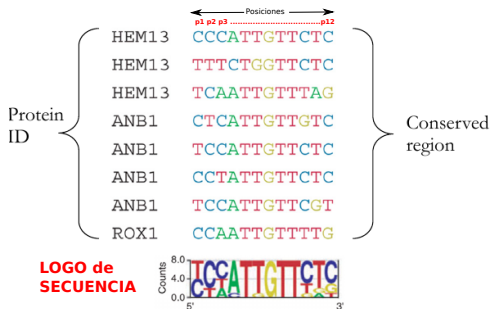
Representación de Motivos: Secuencia Consenso

Podemos extraer la **secuencia consenso**, es decir, la cadena de letras más frecuentes:



Representación de Motivos: Logo de Secuencia

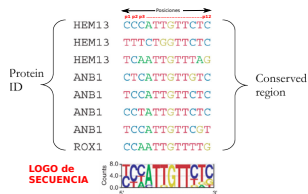
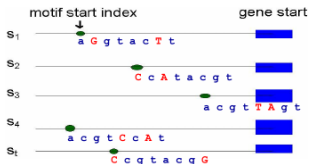
- Una representación gráfica de la secuencia consenso se conoce como el **logo de secuencia**



- La altura de las diferentes letras en una misma posición es proporcional a su frecuencia en los motivos.
- Entre más conservada sea la base en una posición, la altura de la letra será más alta.

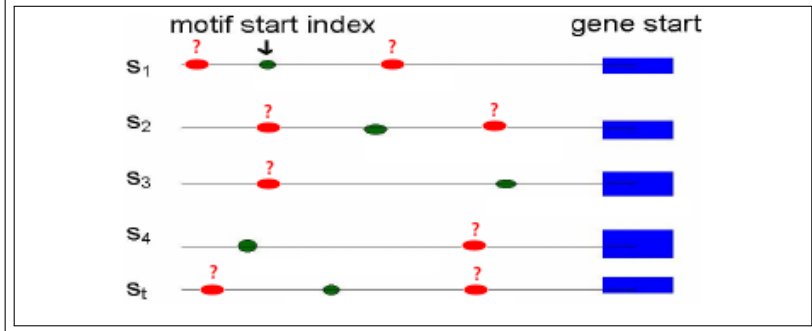
Problema de la Búsqueda de Motivos

- Primero, asumamos que *conocemos* donde comienza el motivo en el conjunto de secuencias de referencia.
- Las posiciones del motivo en las secuencias pueden representarse como el conjunto $S = (s_1, s_2, s_3, \dots, s_t)$ donde s_i es el índice de la posición:



Descubrimiento de Motivos sin Alineamientos

- El problema es encontrar las *posiciones iniciales* $s = (s_1, \dots, s_t)$ para maximizar la puntuación o *scores* (s) de la matrix de perfiles resultantes.



Puntuación de Motivos

- Dado $s = (s_1, s_2, s_3, \dots, s_t)$
alineamos t motivos de longitud
 l de todas las secuencias:

		l							
s_1	a	G	g	t	a	c	T	t	}
s_2	C	c	A	t	a	c	g	t	
.	a	c	g	t	T	A	g	t	
.	a	c	g	t	C	c	A	t	
s_t	C	c	g	t	a	c	g	G	

- Construimos una matrix de
perfiles:

A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

- Calculamos el puntaje del
motivos

$$\text{Score} = 3 + 4 + 4 + 5 + 3 + 4 + 3 + 4$$

$$\text{Score} = 30$$

$$\text{Score}(s) = \sum_{i=1}^l \max_{k \in \{A, C, G, T\}} [\text{count}(k, i)]$$

Tipos de Matrices de Perfiles

Los diferentes matrices de perfiles usadas para describir motivos son:

- Matriz de frecuencias de posición (conteo)
- Matriz de posiciones ponderada (probabilidades)
- Matriz de contenido de información (logaritmos)

Matrix de Frecuencias de Posiciones (PFM)

- Matrix de frecuencia de posición (PFM) registra la frecuencia dependiente de la posición f de cada letra, es decir, cuantas veces una letra ocurre en una posición dada en las N secuencias:

	I								
$s1$	a	G	g	t	a	c	T	t	} t
$s2$	C	c	A	t	a	c	g	t	
.	a	c	g	t	T	A	g	t	
.	a	c	g	t	C	c	A	t	
st	C	c	g	t	a	c	g	G	
	A	3	0	1	0	3	1	1	0
	C	2	4	0	0	1	4	0	0
	G	0	1	4	0	0	0	3	1
	T	0	0	0	5	1	0	1	4

Matrix de Ponderación de Posición (PWM)

- Cada posición de la matriz tiene un peso w o probabilidad.
- Cuando se normaliza a 1 estas frecuencias se vuelve una probabilidad, es decir, $P = f/N$

$$N=5, \text{ PWM } [\langle A \rangle, 1] = 3/5 = 0.6$$

	$\overbrace{\hspace{1.5cm}}^l \hspace{0.5cm} \overbrace{\hspace{1.5cm}}^l$							
s1	a	G	g	t	a	c	T	t
s2	C	c	A	t	a	c	g	t
.	a	c	g	t	T	A	g	t
.	a	c	g	t	C	c	A	t
st	C	c	g	t	a	c	g	G

	PWM							
	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

Probabilidad de una Secuencia dado una Matriz de Pesos: $P(S/M)$

$$P(S/M) = \text{Score}(s)$$

$$\text{Score}(s) = \prod_{i=1}^l p(X_{bi})$$

Donde:

$$b = A, C, G, T$$

$i = 1 \dots l$, l es la longitud de la secuencia

Ejemplo:

$P(\langle \text{ACGTACGT} \rangle / M)$

	PWM							
	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

$$P = 0,6 * 0,8 * 0,8 * 1,0 * 0,6 * 0,8 * 0,6 * 0,8$$

$$P = 0,088$$

Contenido de Información de una Posición

- Cada posición dentro de la matriz de un motivo tiene un contenido de información CI
 - ▶ teoría de información de Shannon, información se da en bits
- Entre más conservada la posición, CI más alto
- Entre más variada la posición, CI más bajo
- Unas posiciones de motivos pueden ser más importantes que otras.

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

Donde

I_i : Contenido de información de la posición i

$f_{b,i}$: Frecuencia de la base b en la posición i .

Contenido de Información de una Posición: Ejemplo

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

Donde

I_i : Contenido de información de la posición i

$f_{b,i}$: Frecuencia de la base b en la posición i .

PWM

CI=-0.32

	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

CI=2.0

$$CI(X_{T,3}) = 2 + \log_2(1,0) = 2 - 0 = 2$$

$$CI(X_{A,3}) = 2 + \log_2(0,2) = 2 - 2.32 = -0.32$$

Contenido de Información de una Secuencia S

Motivos de secuencias con CI bajo:

- Probablemente son aleatorias o
- Corresponden a motivos muy cortos
- O están muy degenerados (no conservados)
- Dado $s = (s_1, s_2, s_3, \dots, s_t)$,

$$l_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

$$CI(S) = \sum_{i=1}^l l_i$$

Pseudoconteos en la Matriz de Frecuencias

Ceros en la matriz de frecuencias pueden generar problemas en los calculos:

- Se presentan cuando hay pocas secuencias $s = (s_1, s_2, s_3, \dots, s_t)$
- Resulta en probabilidades iguales a 0
- Resulta en logaritmos inválidos (log de 0 es infinito)

Solución:

- Sumarle 1 valor que no afecte el conteo

	I								
s_1	a	G	g	t	a	c	T	t	} t
s_2	C	c	A	t	a	c	g	t	
.	a	c	g	t	T	A	g	t	
.	a	c	g	t	C	c	A	t	
s_t	C	c	g	t	a	c	g	G	
	A	3	0	1	0	3	1	1	0
	C	2	4	0	0	1	4	0	0
	G	0	1	4	0	0	0	3	1
	T	0	0	0	5	1	0	1	4

+ 1

- Tener en cuenta este cambio al calcular las otras matrices ($N = N+4$)

Puntuación de secuencias: logaritmos en vez de Probabilidades

$$P(S/M) = \text{Score}(s) \prod_{i=1}^l f_{b,i}$$

$$P(S/M) = \text{Score}(s) \sum_{i=1}^l \log_2 f_{b,i}$$

Donde:

$b = A, C, G, T$

$i = 1 \dots l$, l es la longitud de la secuencia

Ejemplo:

$P(\ll \text{ACGTACGT} \gg / M)$

PWM

	i1	i2	i3	i4	i5	i6	i7	i8
A	0.6	0.0	0.2	0.0	0.6	0.2	0.2	0.0
C	0.4	0.8	0.0	0.0	0.2	0.8	0.0	0.0
G	0.0	0.2	0.8	0.0	0.0	0.0	0.6	0.2
T	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.8

\log_2 $\left(\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \right)$

$$P = \log_2 0,6 + \log_2 0,8 + \log_2 0,8 + \log_2 1,0 + \log_2 0,6 + \log_2 0,8 + \log_2 0,6 + \log_2 0,8$$

Corrección en las Frecuencias Genómicas del ADN de Fondo

Se asume una frecuencia igual de ocurrencia de las bases:

- $f(A) = f(C) = f(G) = f(T) = 1/4 = 0.25$
- Pero el contenido G+C de los genomas cambia por especie:
 - ▶ E.coli (51 % GC), human (41 %) : **razonable**
 - ▶ S. cerevisiae (38 %), Caenorhabditis elegans (36 %) : **de cuidado**
 - ▶ Plasmodium falciparum (19 %), Streptomyces coelicolor (72 %) : **extremo**

Contenido de Información

$$l_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

$$l_i = - \sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

Donde

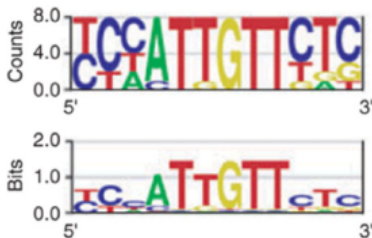
l_i : Contenido de información de la posición i

$f_{b,i}$: Frecuencia de la base b en la posición i .

p_b : Frecuencia de fondo de la base b en el genoms

Logo de Secuencia con Contenido de Información

El logo de secuencia se muestra generalmente usando la teoría de información (Shannon's information in bits):



$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

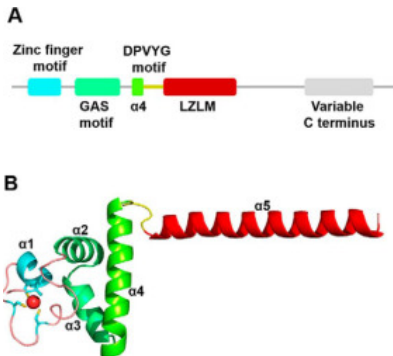
Donde

I_i : Contenido de información de la posición i

$f_{b,i}$: Frecuencia de la base b en la posición i .

PROSITE: base de datos de familias y dominios de proteínas.

- PROSITE es una base de datos de proteínas .
- Consiste en entradas que describen las familias de proteína, los dominios y los sitios funcionales , así como los patrones y perfiles de aminoácidos en ellos.





This form requires to have JavaScript enabled to work correctly.

This form allows you to scan proteins for matches against the **PROSITE** collection of motifs as well as against your own patterns.

- ☒ **Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.**
- ☐ Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.
- ☐ Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

[Reset](#)

STEP 1 - Submit PROTEIN sequences [\[help\]](#)

- ☒ Submit PROTEIN sequences (max. 10) [Examples](#)
- ☐ Submit a PROTEIN database (max. 16MB) for repeated scans (The data will be stored on our server for 1 month).

Enter UniProtKB accessions or identifiers or PDB identifiers or sequences in FASTA format

Supported input:

- UniProtKB accessions e.g. **P98073** or identifiers e.g. **ENTK_HUMAN**
- PDB identifiers e.g. **4DGJ**
- Sequences in **FASTA** format

STEP 2 - Select options [\[help\]](#)

- ☒ **Exclude motifs with a high probability of occurrence** from the scan
- ☐ Exclude profiles from the scan
- ☐ Run the scan at **high sensitivity** (show weak matches for profiles)

STEP 3 - Select output options and submit your job

Output format:

[Graphical view](#) ▼

Retrieve complete sequences: ☐ If you choose this option, not all output formats are available.

- ☐ Receive your results by email

[START THE SCAN](#) [Reset](#)

This view shows ScanProsite results together with rule-based predicted features inside (profile) matches.

exclude splice variants

Hits for all PROSITE (release 19.20) motifs on sequence ENTK_HUMAN [UniProtKB/Swiss-Prot (release 49.0)]:

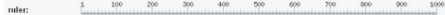
found: 13 hits in 1 sequence

P98073 ENTK_HUMAN (1019 aa)

Enteropeptidase precursor (EC 3.4.21.9) (Enterokinase) [Contains: Enteropeptidase non-catalytic heavy chain; Enteropeptidase catalytic light chain]. *Homo sapiens (Human)*

```

MGGRGSGSRSHSLSYEINFAALFAILVVLCAGLIAYSCLTIXESQGAALQSHKARATPRTIS
GYTTHMIGQELSYDPKYLAFLQIQKIDEIFLSKMLXETYSRVLQRENSITVPIIDFFAQVPS
DQVWEELIGLEAMSSQLVTFPHIDMSYDIDWLTTSHLATFMVSIKCLPSSSTDAALTCI
KAGLFDGEVHDSGSEIAGNCAATVCGRFLTSSSGSFQATHYKFSSTVYVQWILSPVQGLS
IKLSFDQPMITYTIDIDYEGVGSKILRASLWETPOTIRIFSBQVIATFLIEDSDYVGFNAT
VTAFESLHNTETNCHVEDGCFYVQDLDMSHEDIGSTPSPFQPMFHTVGRASGPTIST
PTFGQSGEVLILSLHPTLEPAFLPTHTYQENRHSINHEQUNKCTVFGKSEYVQSGE
MYGQVTLNCTVYKNAFNAFQMKILSDIALDDISLTVGICGSLVFEPTLVPTFPPELPTDCGGFF
ELAEFNTTSTNFPSTPBLAFVYVILMAQSGKTLQHFQEPDLENNVVEFDGEADSLILA
VYTGSPFQVYFTTBNITVLLTHVLARGQFQAMPTVTHLGIPEFCADGFPQKQDECVPLVN
LCQSHLCEVSDGEACVYSPGUTTNBQVHFVPSINNEACAEMUTTLSDWCVLGLGSGES
SKQIFSTSGSPFKMLTAFPGRLILTPQDCLQDLIRLQGNKSCGKGLAADIITHTVGGNAK
EGATVYVYVYLYSGRLLCASLVSQVLSAAKCTYGNLEPKRUTALGLHNSGHLTSPQVPEL
IDEIVINHTNDFPSSNDLMMHLEPKYNTLTIQPICLPEERQVYFPQHCSLAGSTVYVQDT
ANILQKADVPLISHEPCQQPFYHTIEMICAGTEEGGDSQSGSGGLNCGQNNHFLAGVTS
PQTKCALPBPFGYAPYSKPTETVIGSLR
    
```



hits by profiles: [8 hits (by 6 distinct profiles) on 1 sequence]



Enteropeptidase precursor (EC 3.4.21.9) (Enterokinase) [Contains: Enteropeptidase non-catalytic heavy chain; Enteropeptidase catalytic light chain]. *Homo sapiens (Human)*

PS50024 SEA SEA domain profile:

52 - 169: score = 24.822

```

LQSHREARATPKIT--SGVYVWPIQKLSVDPKYLAFLQIQKIDEIFLSKMLXETYSR
RYLQPEHNSII-VYFDLFFAQV--SDQVWEELIGLEAMSSQLVTFPHIDMSYDIDWL
LT
    
```

PS50068 LDLRA_2 LDL-receptor class A (LDLRA) domain profile:

183 - 222: score = 10.75

```

KCLPSSSTDAALTCIKADLPCEGVHDSGSEIAGNCA
    
```

Predicted features:

DISULFID	184	197	By similarity	[condition: C-x*-C]
DISULFID	191	210	By similarity	[condition: C-x*-C]
DISULFID	204	221	By similarity	[condition: C-x*-C]

PRATT: Herramienta para descubrir patrones conservados en secuencias de proteínas.

PRATT version 2.1

PRATT is a [tool to discover patterns conserved in a set of protein sequences.](#)

This tool can also be run from the [EBI server](#) with very similar modalities.

STEP 1 - Enter a set of PROTEIN sequences or an alignment

Examples

Enter sequences in FASTA or UniProtKB format or an alignment in FASTA format

Supported input:

- Sequences in [FASTA](#) or [UniProtKB/Swiss-Prot](#) format
- Alignment of sequences in FASTA format

Your input is ☒ a set of sequences ☐ an alignment

STEP 2 - Modify default parameters (optional)

- » Pattern parameters
- » Search parameters
- » Output parameters

If you're familiar with PRATT, you can directly enter parameters in a command line format:

e.g. -PL 25

STEP 3 - Submit your job

- ☒ Directly submit best pattern to [ScanProsite](#)
- ☐ View PRATT output file

PRATT: Resultado Mejores Patrones

Best Patterns (after refinement phase):

	fitness	hits(sequ)	Pattern
A 1:	187.5718	10(10)	D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T
B 2:	187.5718	10(10)	L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T
C 3:	186.5787	10(10)	R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A
D 4:	186.5787	10(10)	K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y
E 5:	183.4017	10(10)	L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P
F 6:	183.4017	10(10)	F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P
G 7:	182.8081	10(10)	L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q
H 8:	181.8293	10(10)	K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S
I 9:	181.7865	10(10)	L-[EQ]-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G
J 10:	179.2317	10(10)	D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P
K 11:	178.9369	13(10)	L-x(0,1)-L-x(0,1)-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q
L 12:	178.6380	10(10)	L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q
M 13:	178.5013	10(10)	E-L-x(0,1)-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S
N 14:	176.1961	10(10)	E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C-R-[EK]-E
O 15:	176.0141	10(10)	P-W-T-G-Y-G-L-L-L-[DEN]-G-x(2)-W-F-Q-H-R-R-M-L-T-P-A-F-H-Y-D-x-L-K-P-Y-V-x(2)-M-[AV]-D-S-[IV]-[QRS]-[ILV]-M-L-D-[KR]-x-E
P 16:	175.3700	10(10)	A-[KR]-[IM]-E-[DN]-G-x(0,1)-S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C
Q 17:	175.3605	15(10)	K-x(1,2)-R-x(0,1)-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]
R 18:	174.3313	12(10)	E-x(0,1)-L-x-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A
S 19:	173.8313	12(10)	E-x(0,1)-L-x(0,1)-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A
T 20:	171.9571	10(10)	Q-L-Q-x-[AE]-[EG]-E-L-[EQ]-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E
U 21:	171.8440	10(10)	W-T-G-Y-G-L-L-L-[DEN]-G-x(2)-W-F-Q-H-R-R-M-L-T-P-A-F-H-Y-D-x-L-K-P-Y-V-x(2)-M-[AV]-D-S-[IV]-[QRS]-[ILV]-M-L-D-[KR]-x-E
V 22:	171.6357	13(10)	L-x(0,2)-L-L-x(0,1)-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P
W 23:	170.9665	10(10)	I-G-Y-G-L-L-L-[DEN]-G-x(2)-W-F-Q-H-R-R-M-L-T-P-A-F-H-Y-D-x-L-K-P-Y-V-x(2)-M-[AV]-D-S-[IV]-[QRS]-[ILV]-M-L-D-[KR]-x-E-x-[IL]
X 24:	170.5792	10(10)	X-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C-R-[EK]-E-[IV]-x-[GS]-[ILV]-L
Y 25:	170.5792	10(10)	S-x(2)-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C-R-[EK]-E-[IV]-x-[GS]-[ILV]-L
Z 26:	169.4611	10(10)	Y-x(2)-[LMV]-[AT]-P-W-I-G-Y-G-L-L-L-L-[DEN]-G-x(2)-W-F-Q-H-R-R-M-L-T-P-A-F-H-Y-D-x-L-K-P-Y-V-x(2)-M-[AV]-D-S-[IV]-[QRS]-[ILV]-M
a 27:	169.6613	21(10)	E-x(0,2)-L-x-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T
b 28:	168.7117	13(10)	L-x(0,1)-A-x(2)-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q
c 29:	168.6732	10(10)	L-x(0,1)-A-x(2,3)-E-L-[EQ]-K-[AIV]-[KR]-x-K-R-[HR]-L-D-F-L-D-[IV]-L-L-[FL]-A-[KR]-[IM]-E-[DN]-G-x(2)-[LM]-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G
d 30:	168.4069	15(10)	S-x(1,2)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C-R-[EK]-E-[IV]
e 31:	168.4069	10(10)	G-x(2,3)-S-D-x-D-L-R-A-E-V-D-T-F-M-F-E-G-H-D-T-T-A-S-G-[IV]-S-W-[IV]-[FL]-Y-A-L-A-T-x-P-[DEK]-H-Q-x(2)-C-R-[EK]-E-[IV]