FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS
DEPARTAMENTO DE BIOTECNOLOGÍA Y TECNOLOGÍA ALIMENTARIA
UNIVERSIDAD ARGENTINA DE LA EMPRESA

Bioinformática

ANÁLISIS COMPUTACIONAL DE SECUENCIAS

Dr. Lucas L. Maldonado (PhD)

Lic. Biotechnologist and Molecular Biologist

Bioinformatics and genomics specialist

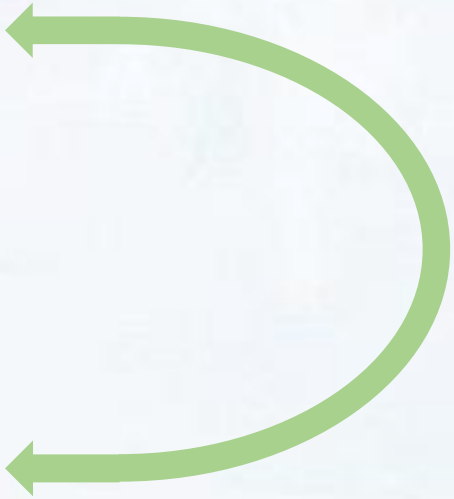
CONICET
Fac. de Medicina - UBA
Fac. de Ciencias Exactas y Naturales – UBA

lucamaldonado@uade.edu.ar
lmaldonado@fmed.uba.ar
lucas.l.maldonado@gmail.com.ar

Motivos, patrones y perfiles

Representación de alineamientos múltiples

1. REGULAR EXPRESSIONS
2. FINGERPRINTS
3. BLOCKS
4. PROFILES AND PSSMS
5. HMM



Bases de datos secundarias o de patrones

- PROSITE - a collection of patterns and profiles
- Pfam - A collection of Profiles generated using hidden Markov models
- PRINTS - provider of fingerprints (groups of aligned, un-weighted motifs)
- BLOCKS - a database of weighted profiles or blocks

Representación de los MSA

- Los MSA se utilizan para representar o caracterizar **familias** de secuencias relacionadas.
- No resulta práctico trabajar directamente con los MSA por lo que se han desarrollado diversas maneras de representarlos.
- Los distintos métodos de representación de AMS forman una *jerarquía de modelos*: cada método és un caso particular del que le sigue en complejidad.

Motivos, patrones y perfiles

Al alinear secuencias de proteínas, a menudo es evidente que ciertas regiones o aminoácidos específicos están más conservados que otros. Estas regiones conservadas se conservan a menudo porque codifican una parte de la proteína que es funcionalmente importante. **El término motivo se usa para referirse a una parte de una secuencia de proteína que está asociada con una función biológica particular.**

estas regiones se conservan y pueden ser reconocibles por la presencia de una secuencia particular de aminoácidos llamada **patrón**.

Un **patrón** es una descripción cualitativa de un **motivo** en términos de secuencia de aminoácidos.

El concepto de **perfil** (**Profiles**) amplía este concepto y otorga una descripción **cuantitativa de un motivo:**

- Se asignan probabilidades a la presencia de un aminoácido en particular en cada posición de un motivo.
- Se pueden utilizar para describir motivos muy divergentes.

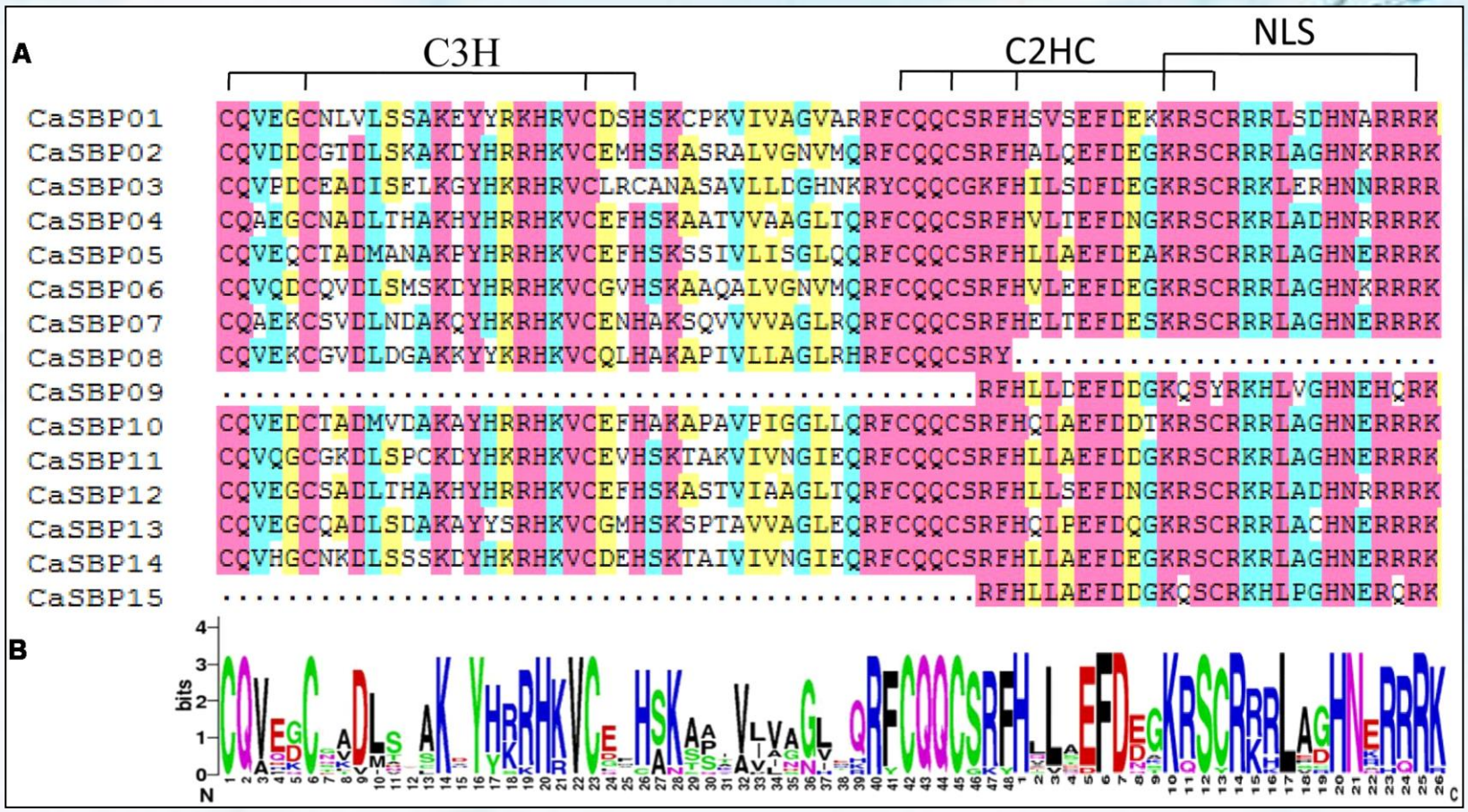
La presencia de un motivo particular dentro de una secuencia de proteína puede usarse para inferir funciones para proteínas no caracterizadas.

Aplicaciones de los Motivos

- Los motivos representan zonas conservadas entre las secuencias que suelen asociarse a características funcionales del grupo de secuencias.
- Una vez se ha construido un motivo o patrón de un grupo de secuencias puede utilizarse
  - Para asociar una nueva secuencia con la familia de secuencias que lo ha generado (si presenta el motivo es de la familia y puede que comparta sus funciones)
  - Para buscar secuencias que pertenezcan a aquella familia

MSA y logo que representa la secuencia consenso

En un alineamiento múltiple se colocan las secuencias de modo que el número de posiciones con residuos idénticos o parecidos sea máximo.



Cuanto más separadas estén las secuencias desde el punto de vista evolutivo, más diferencias encontraremos entre ellas

Motivos señales o patrones

- Consideremos un alfabeto como el del ADN o las proteínas.
- Un *motivo* (*patrón* o *señal*) es una forma de caracterizar un conjunto de secuencias de este alfabeto.
- Dada una secuencia, *S*, y un motivo *M* diremos que *M* está presente en *S* si cualquiera de las secuencias descritas por *M* ocurre en *S.*

- P.ej:

M="TATA", *S1="GATTACA"* *y* *S2="PATATA"*

M está presente en S2 pero no en S1

Modelos para MSA y motivos

- Una manera natural de representar un MSA es a través de los motivos o patrones que contiene.
- La jerarquía de modelos para AMS a la que hemos hecho referencia es, pues, también una jerarquía de modelos para motivos: El patrón característico del alineamiento es el "motivo" que lo caracteriza.

Representación de alineamientos múltiples

1. REGULAR EXPRESSIONS
2. FINGERPRINTS
3. BLOCKS
4. PROFILES AND PSSMS
5. HMM

Descripción de motivos (0)
Palabra exacta

- La manera más simple de describir un motivo contenido en un AMS es a través de la secuencia exacta de letras (la "palabra") que lo forman
- Muy preciso si se presenta pero no admite variaciones

G	A	T	T	A	C	A
G	A	T	T	A	C	T
G	A	T	T	A	C	T
G	A	T	T	A	C	A
G	A	T	T	A	C	C
	A	T	T	A	C	

G	A	T	T	A	C	A
G	A	C	T	A	C	T
T	A	T	T	A	C	T
C	A	T	T	G	C	A
A	A	T	T	A	C	C
	A	?	T	?	C	

Descripción de motivos (0)
La secuencia consenso

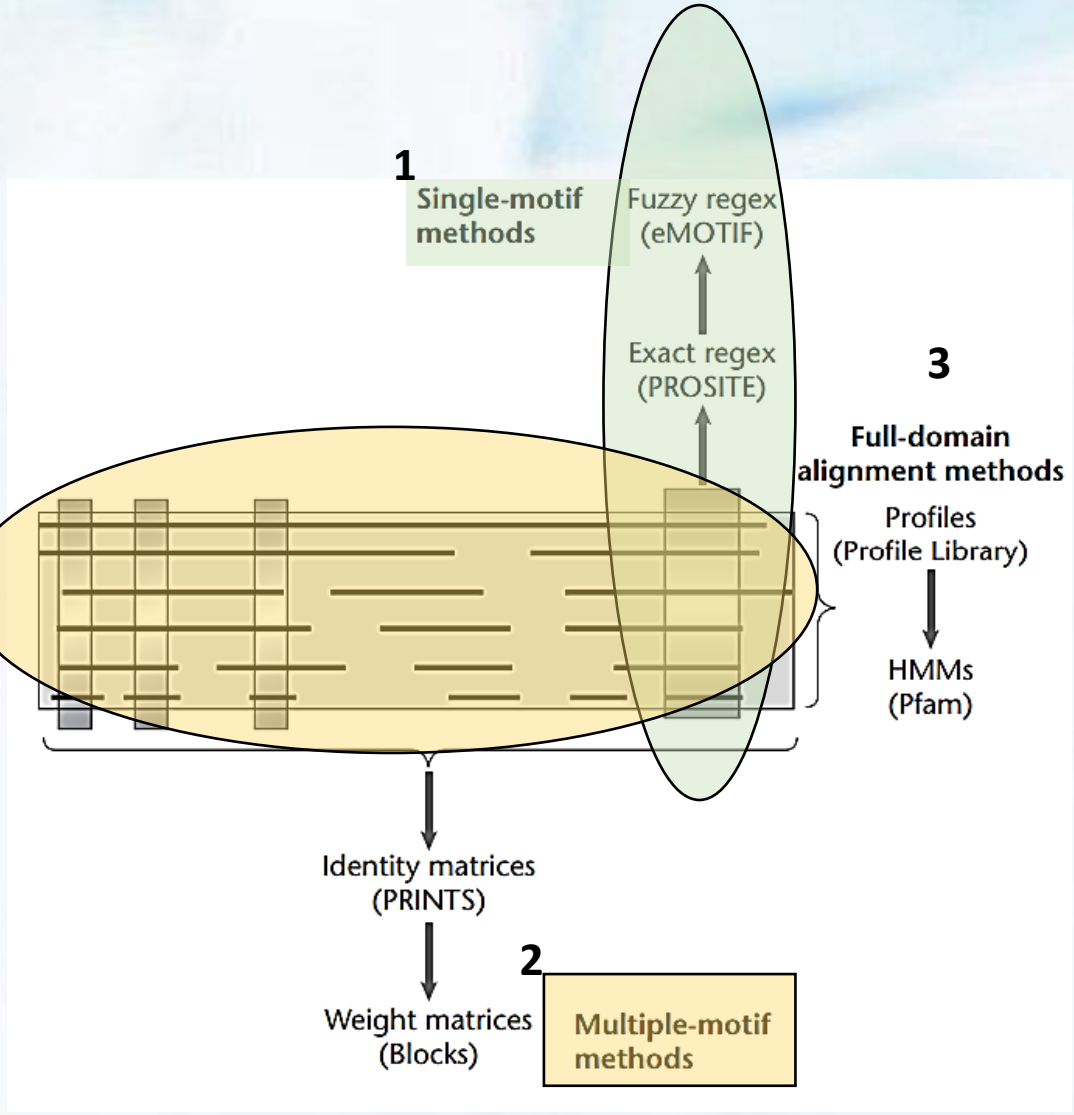
- Si en alguna posición aparecen cambios en la palabra exacta se pueden utilizar caracteres diversos para indicar estas variaciones.
- Por ejemplo
  - Si todas las secuencias tienen el mismo residuo en una posición dada se pone la letra mayúscula
  - Si la mayoría tiene la letra se pone minúscula
  - Si hay empate se ponen las letras empatadas

Secuencia consenso: útil si hay pocas variaciones.

Un ejemplo de secuencia consenso

	1	2	3	4	5	6	7	8	9	10
I	Y	D	G	G	A	V	-	E	A	L
II	Y	D	G	G	-	-	-	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	-	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
	y	d	G	G	A/I	V/L	V	e	A	l

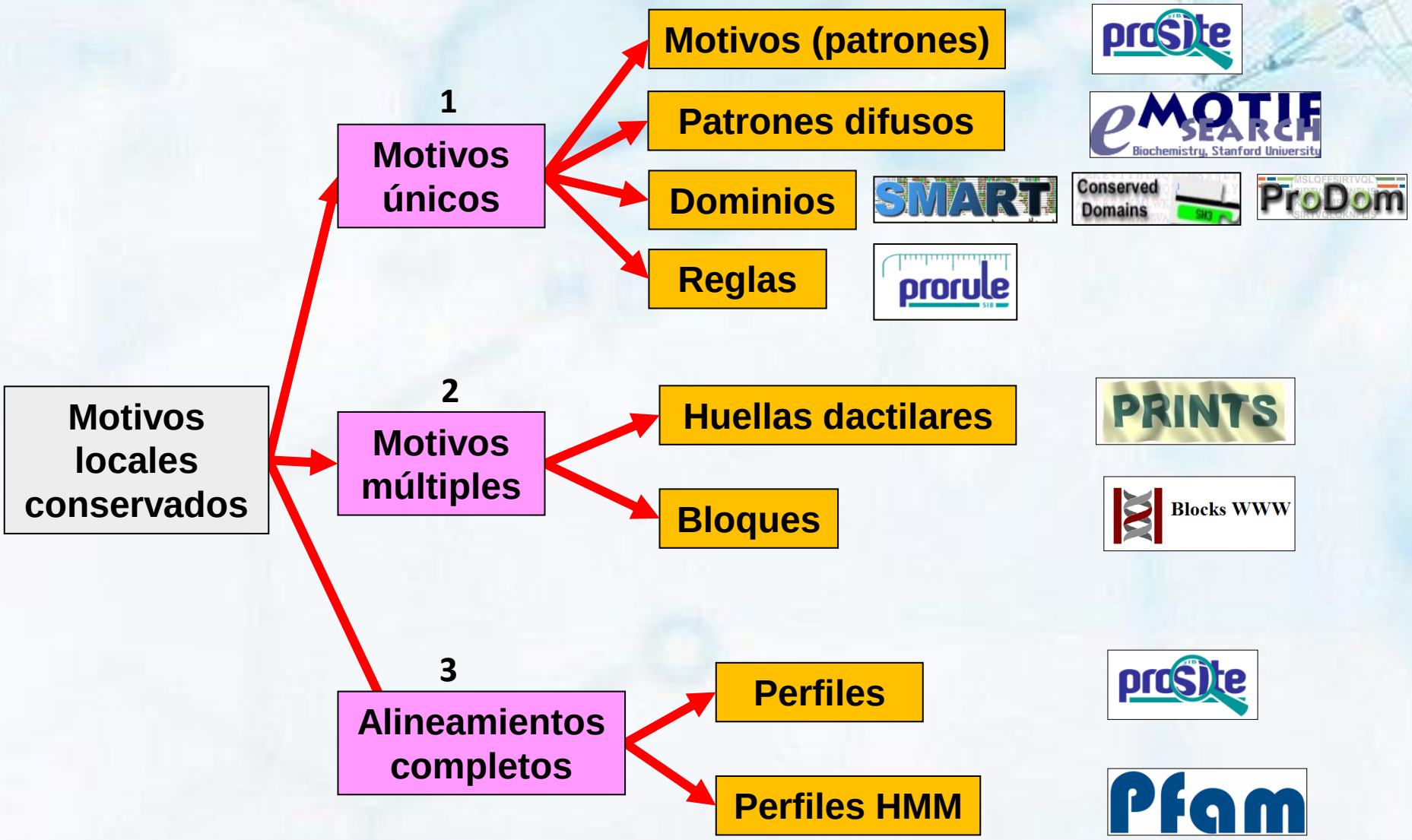
Métodos principales para construir bases de datos familiares, basados en el uso de motivos únicos, motivos múltiples y alineaciones de dominio completo.



Los alineamientos se han aprovechado de diversas formas para construir patrones o Prints (Huellas dactilares) de diagnóstico, o para discriminar entre diferentes familias de proteínas, *(Attwood 2000a, b, Attwood y Parry-Smith 1999*).

Una secuencia desconocida puede ser buscado en una base de datos de tales Prints para determinar si contiene alguna de las características predefinidas y, por lo tanto, si puede ser asignada a una familia en particular

Tipos de motivos conservados en un AMS y BD2



Los principales métodos utilizados para derivar tales firmas se dividen esencialmente en tres grupos, basados en el uso de motivos únicos, motivos múltiples o alineamientos de dominios completos.

1. REGULAR EXPRESSIONS

El método de reconocimiento de patrones más simple de entender es la **expresión regular**, o **regex**, a menudo simplemente (y de manera confusa) referida como un "**patrón (pattern)**".

- Buscando una mayor flexibilidad se propuso el uso de expresiones regulares
  - *Una expresión regular, a menudo llamada también **patrón**, es una expresión que describe un conjunto de cadenas sin enumerar sus elementos* (Wikipedia)
  - Son ampliamente utilizadas en informática, en entornos UNIX/Linux especialmente, para manipular cadenas de caracteres de manera muy flexible.

1. REGULAR EXPRESSIONS

Sintaxis de expresiones regulares
Caracteres comodín

- Si en una posición dada puede aparecer cualquier carácter se indica con el signo “comodín”
- Aunque en informática éste es a menudo un “*” aquí se utilizará una “x”

G	A	T	T	A	C	A
G	A	C	T	A	C	T
T	A	A	T	A	C	T
A	A	T	T	A	C	C
	A	x	T	A	C	

Patrón: A-x-T-A-C

1. REGULAR EXPRESSIONS

Sintaxis de expresiones regulares
Ambiguedades

- Si en una posición dada puede aparecer varios caracteres distintos podemos indicarlo de dos formas
  - Aquellos que pueden aparecer: entre “[“ y “]”
  - Aquellos que no se encuentran en la posición: entre “{“ y “}”
- Una misma secuencia se puede indicar de maneras distintas. *P.ej: [ATC] equivale a {G}*

G	A	T	T	A	C	A
G	A	C	T	T	C	T
T	T	A	T	C	C	T
A	T	T	T	A	C	C
	[AT]	x	T	{G}	C	

Patrón: [AT]-x-T-{G}-C={CG}-x-T-[ATC]-C= ...

1. REGULAR EXPRESSIONS

Sintaxis de expresiones regulares
Elementos repetidos

- La repetición de un elemento se indica con éste entre paréntesis: “(“y”)”
  - A(4) indica una “A” repetida 4 veces
  - x(3) indica un caracter cualquiera repetido 3 veces
- Si el elemento que se repite es uno cualquiera (“x”) puede asignarsele un número variable de repeticiones, incluso el cero
  - x(2-4): “x-x”, “x-x-x”, “x-x-x-x”
  - x(0-2): “”, “x”, “x-x”

1. REGULAR EXPRESSIONS

Cómo se obtiene una expresión regular a partir de un AMS

Patrón = Secuencia consenso = Expresión regular

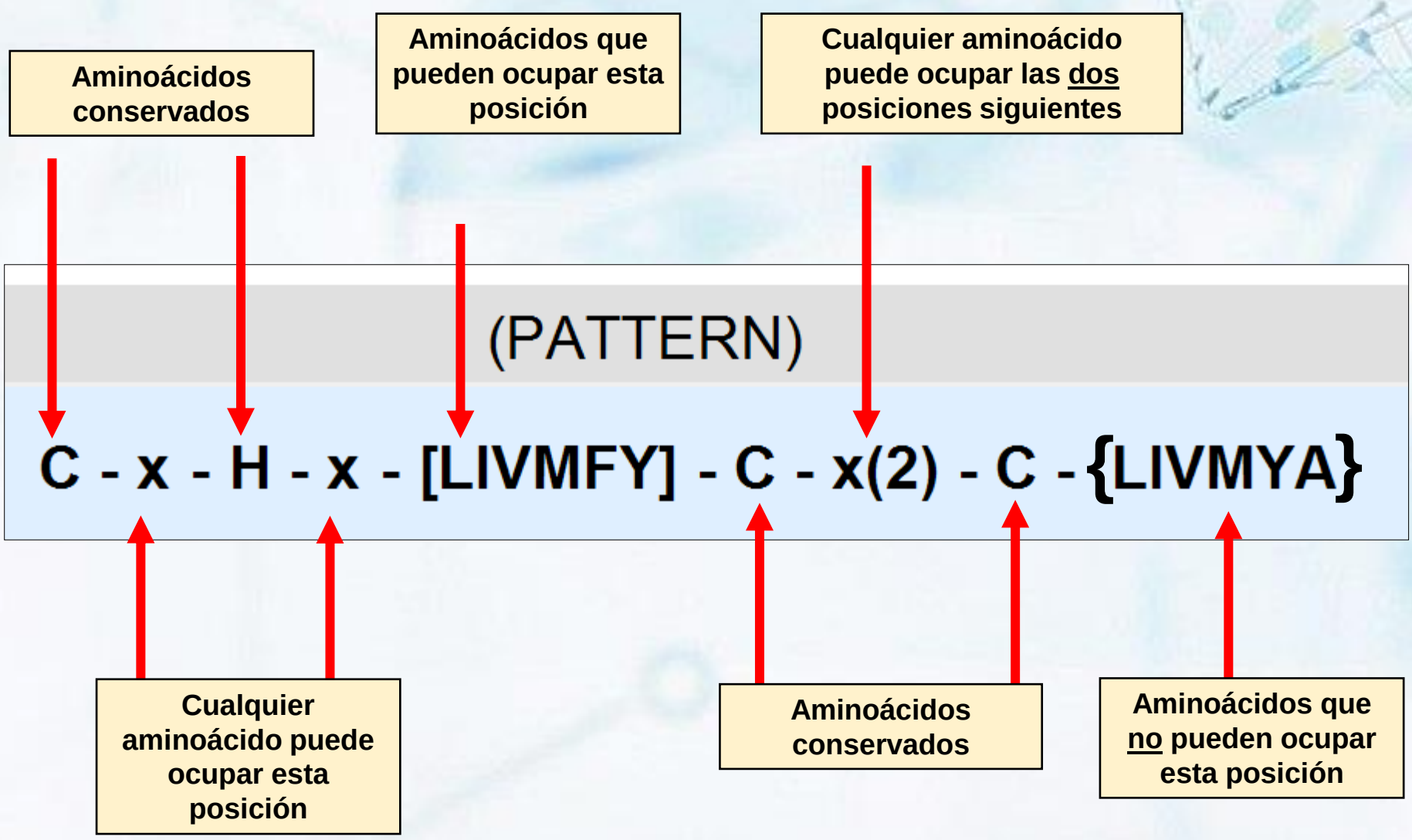
G	H	E	G	V	G	K	V	V	K	L	G	A	G	A
G	H	E	K	K	G	Y	F	E	D	R	G	P	S	A
G	H	E	G	Y	G	G	R	S	R	G	G	G	Y	S
G	H	E	F	E	G	P	K	G	C	G	A	L	Y	I
G	H	E	L	R	G	T	T	F	M	P	A	L	E	C

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	H	E	G K F L	V K Y E R	G	K Y G P T	V F R K T	V E S G F	K D R C M	L R G P	G A	A P G L	G S Y E	A S I C

Pattern: G–H–E–X(2)–G–X(5)–[GA]–X(3)

Search databases

1. REGULAR EXPRESSIONS



Sintaxis para generar una expresión regular

Características de los patrones

- **Patterns are derived from single conserved regions, which are reduced to consensus expressions for db searches**
  - they are minimal expressions, so sequence information is lost
  - the more divergent the sequences used, the more fuzzy & poorly discriminating the pattern becomes

Alignment	Pattern
GAVDFIALCDRYF	
GPIDFVCFCERFY	G-X-[IV]-[DE]-F-[IVL]-X2-C-[DE]-R-[FY]2
GRVEFLNRCDRYY	

- **Patterns do not tolerate similarity**
  - sequences either match or not, regardless of how similar they are
  - matching is a binary ‘on-off’ event & frequently misses true matches
  - single-motif methods are very hit-or-miss - how do you know if you've encoded the ‘best’ region?

1. REGULAR EXPRESSIONS

Ventajas e inconvenientes de los patrones

Patterns: Conclusion

- Patterns are appropriate to build models of short sequence signatures.
- Advantages:
  - Pattern matching is fast and easy to implement.
  - Models are easy to design for anyone with some training in biochemistry.
  - Models are easy to understand for anyone with some training in biochemistry.
- Limitations:
  - Poor model for insertions/deletions (indels).
  - Small patterns find a lot of false positives. Long patterns are very difficult to design.
  - Poor predictors that tend to recognize only the sequence of the training set.
  - No scoring system, only binary response (YES/NO).
- When I use patterns?
  - To search for small signatures or active sites.
  - To communicate with other biologists.

1. REGULAR EXPRESSIONS

Ventajas

1.- Son fáciles de entender y utilizar por el usuario

2.- Localizan las regiones más conservadas, que suelen estar asociadas a una función biológica

3.- La búsqueda de patrones en bases de datos de proteínas se hace en un periodo de tiempo razonable

Inconvenientes

1.- No localiza homólogos distantes, ya que ignora las secuencias que no coinciden por completo con el patrón

1. REGULAR EXPRESSIONS

La BD PROSITE

ExPASy Bioinformatics Resource Portal — PROSITE — Home | Contact

Home | ScanProsite | ProRule | Documents | Downloads | Links | Funding

Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule , a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2019_03 of 10-Apr-2019 contains 1829 documentation entries, 1310 patterns, 1240 profiles and 1264 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Search

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] Examples

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

Scan Clear

Exclude motifs with a high probability of occurrence from the scan

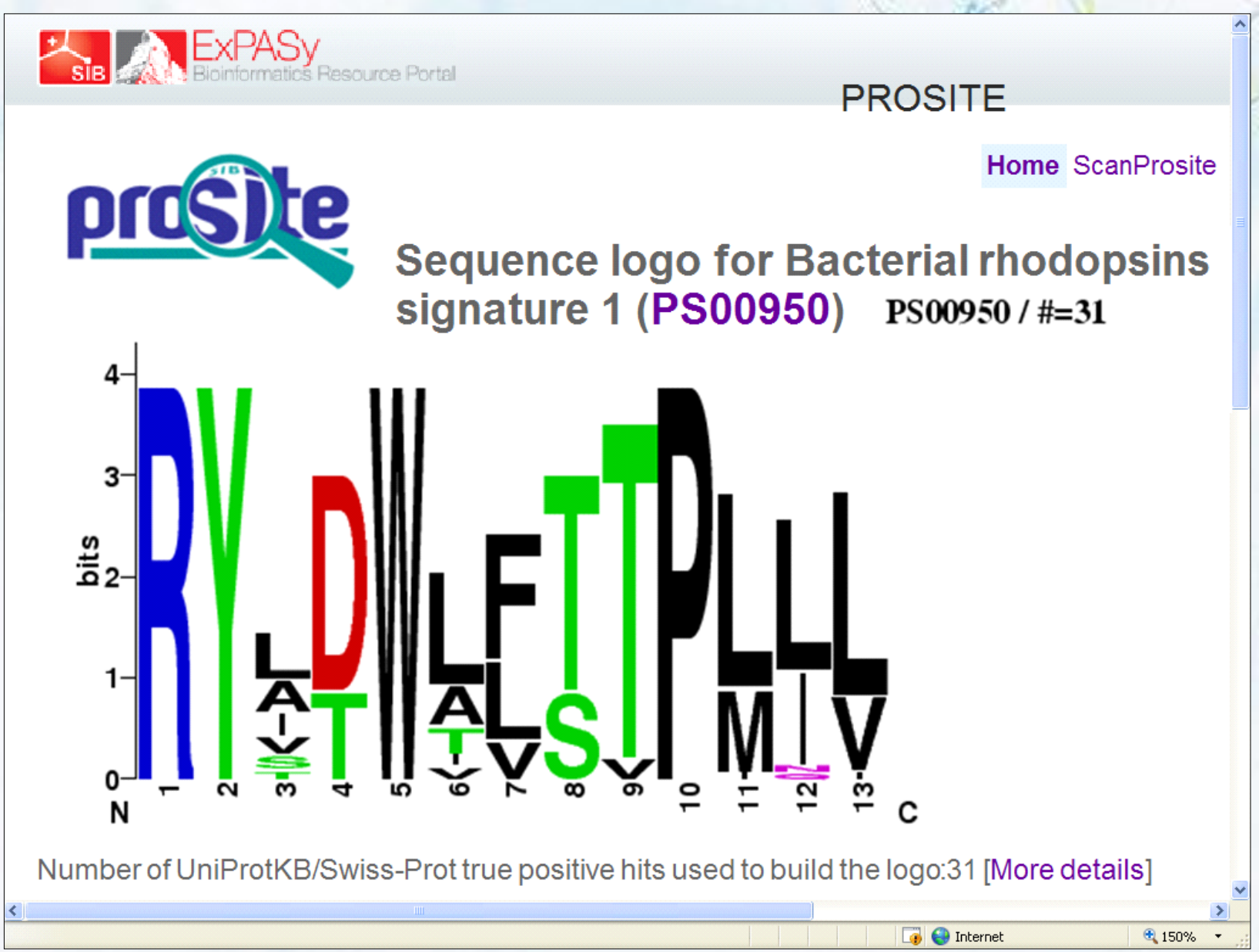
For more scanning options go to **ScanProsite**

Other tools

- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** - allows to generate custom domain figures.

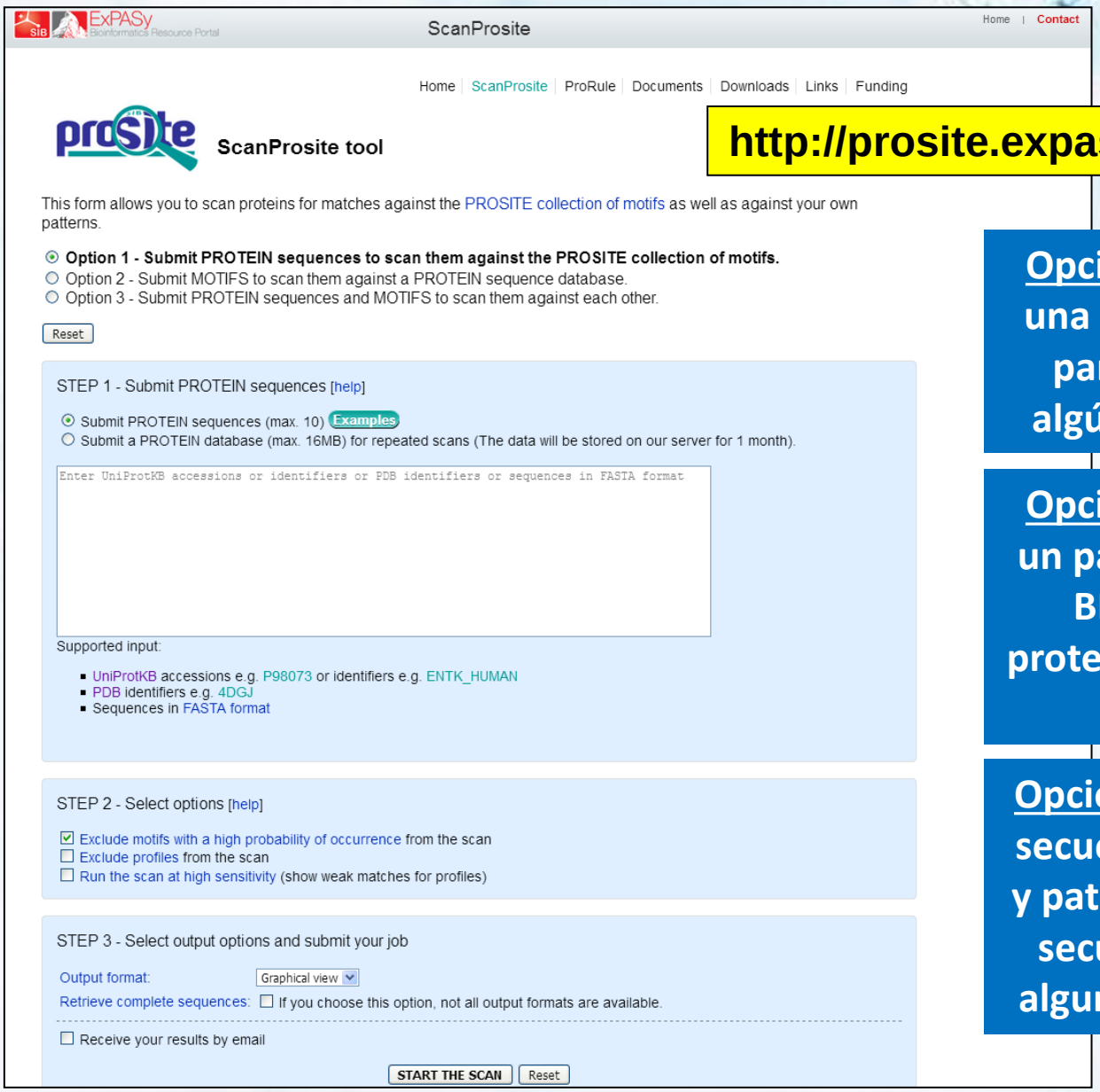
http://prosite.expasy.org/

1. REGULAR EXPRESSIONS



Logo de un patrón almacenado en la BD PROSITE

1. REGULAR EXPRESSIONS



http://prosite.expasy.org/scanprosite

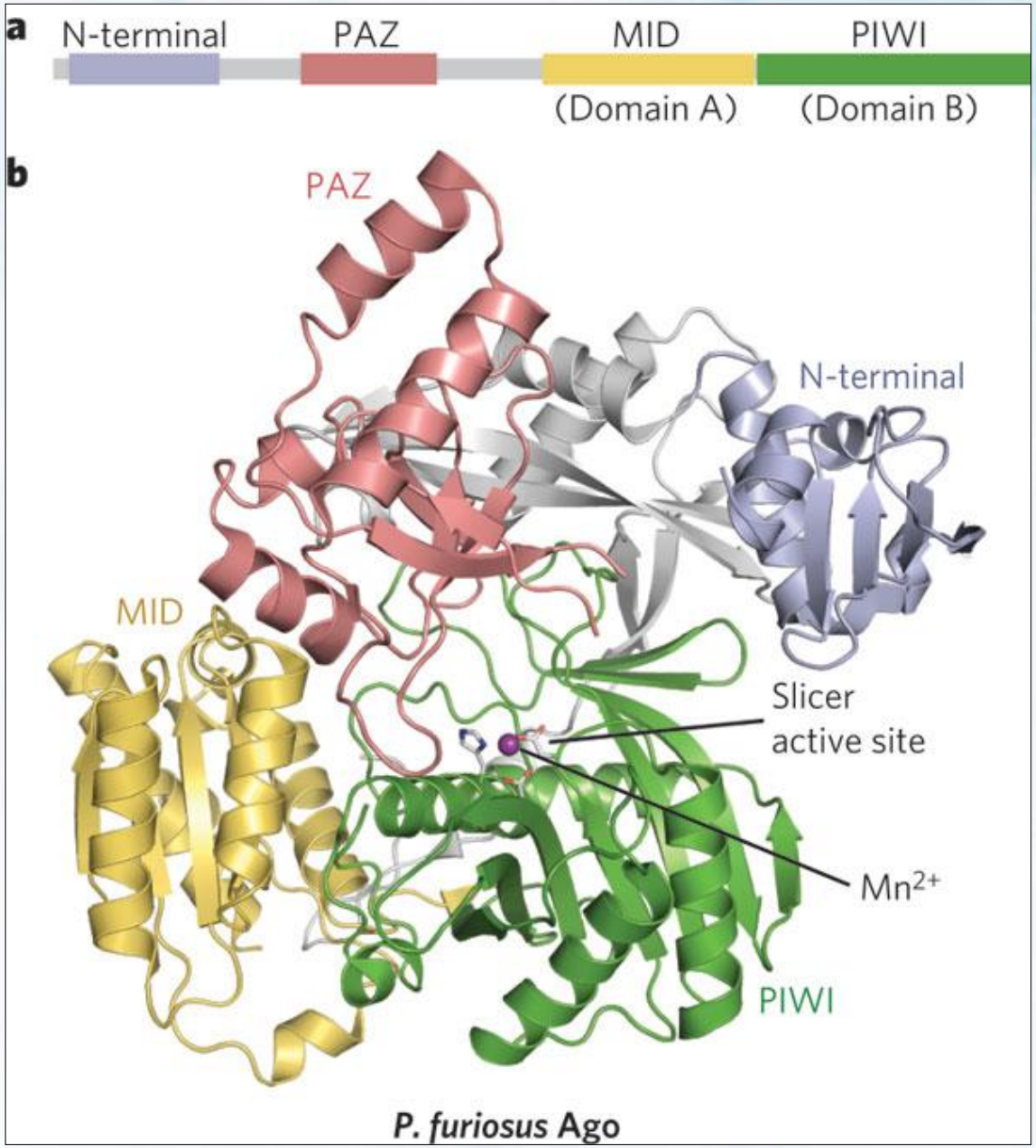
Opción 1: Se introduce una secuencia proteica para ver si contiene algún motivo de la BD

Opción 2: Se introduce un patrón y se busca en BD de secuencias proteicas aquéllas que lo contienen

Opción 3: Se introducen secuencias de proteínas y patrones para ver qué secuencias contienen alguno de esos motivos

La herramienta ScanProsite

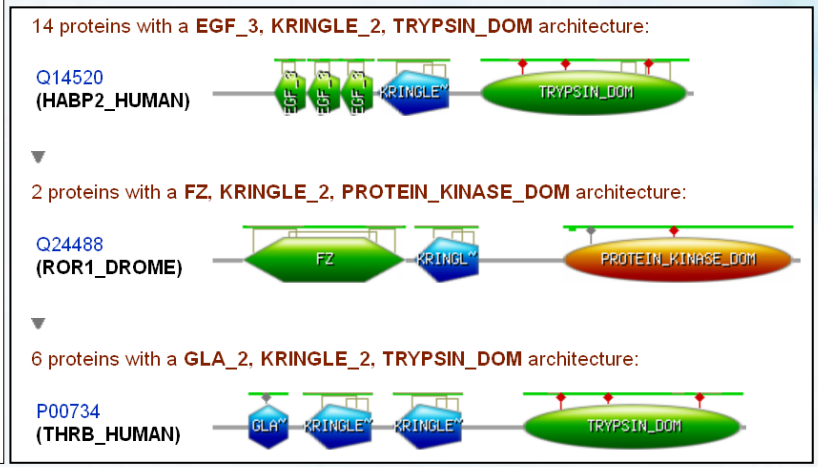
1. REGULAR EXPRESSIONS



Domains

Domains are compact, local, semi-independent folding units that need not be formed from contiguous regions of an amino-acid sequence: they may be discrete entities, joined only by a flexible linking region of polypeptide chain; they may have extensive interfaces, sharing many close contacts, and they may exchange chains with domain neighbours. In the simplest genetic terms, domains may be considered to have arisen via gene fusion and gene duplication events.

3D domains are compact structural units identified by purely geometric criteria.



PROSITE permite determinar Dominios completos

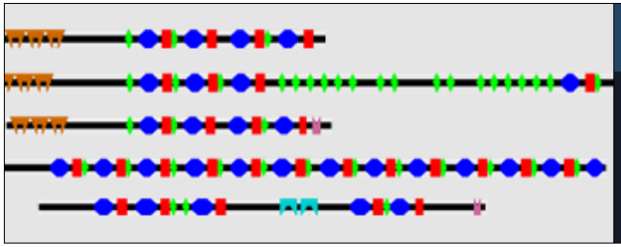
https://www.ncbi.nlm.nih.gov/cdd

NCBI Resources How To Sign in to NCBI

Conserved Domains

Conserved Domains Search

Advanced Help



CDD

The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships.

Using CDD

- Quick Start Guide
- How To Guides
- Help
- FTP
- News
- Publications

CDD Tools

- Overview of CDD Resources
- CD-Search
- Batch CD-Search
- CDART (Domain Architectures)
- CDTree (classification and research tool)
- BLAST

Other Resources

- Structure Group Home Page
- Entrez Structure (Molecular Modeling Database)
- Entrez Gene
- Entrez Protein
- BioSystems
- FLink

You are here: NCBI > Domains & Structures > Conserved Domain Database (CDD) Support Center

CDD: una BD de dominios conservados (NCBI)

Home | ScanProsite | ProRule | Documents | Downloads | Links | Funding

Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2018_03 of 28-Mar-2018 contains 1806 documentation entries, 1309 patterns, 1214 profiles and 1234 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Search

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] Examples

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

Scan | Clear

Exclude motifs with a high probability of occurrence from the scan

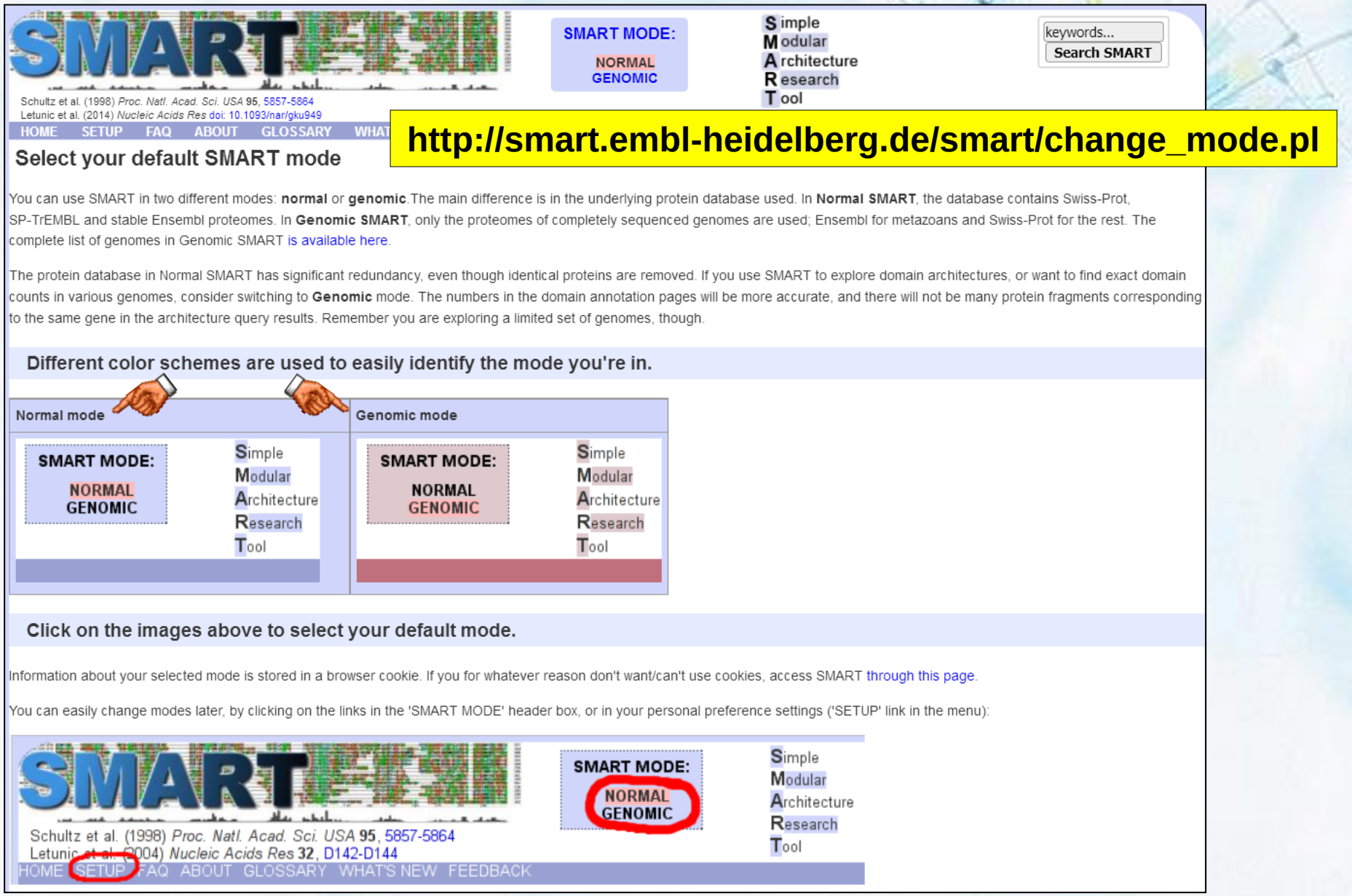
For more scanning options go to **ScanProsite**

Other tools

- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** - allows to generate custom domain figures.

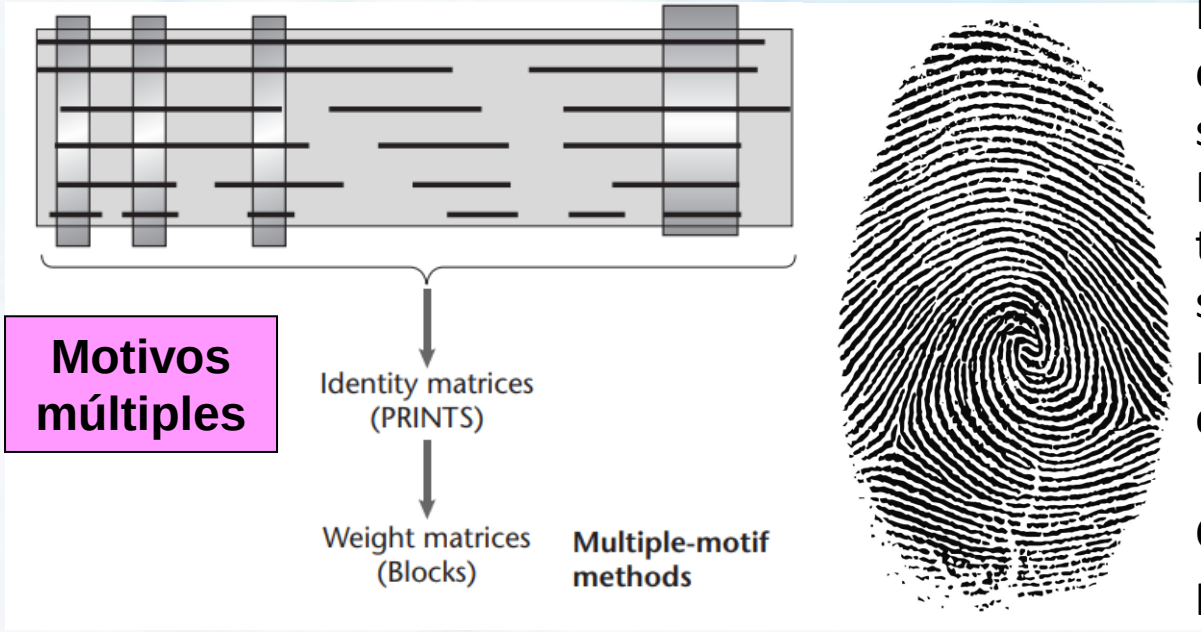
https://prosite.expasy.org/

La BD PROSITE


SMART
SMART MODE:
NORMAL
GENOMIC
Simple
Modular
Architecture
Research
Tool
keywords...
Search SMART
Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Letunic et al. (2014) Nucleic Acids Res doi: 10.1093/nar/gku949
HOME SETUP FAQ ABOUT GLOSSARY WHAT
http://smart.embl-heidelberg.de/smart/change_mode.pl
Select your default SMART mode
You can use SMART in two different modes: normal or genomic.The main difference is in the underlying protein database used. In Normal SMART, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes. In Genomic SMART, only the proteomes of completely sequenced genomes are used; Ensembl for metazoans and Swiss-Prot for the rest. The complete list of genomes in Genomic SMART is available here.
The protein database in Normal SMART has significant redundancy, even though identical proteins are removed. If you use SMART to explore domain architectures, or want to find exact domain counts in various genomes, consider switching to Genomic mode. The numbers in the domain annotation pages will be more accurate, and there will not be many protein fragments corresponding to the same gene in the architecture query results. Remember you are exploring a limited set of genomes, though.
Different color schemes are used to easily identify the mode you're in.
Normal mode
Genomic mode
Click on the images above to select your default mode.
Information about your selected mode is stored in a browser cookie. If you for whatever reason don't want/can't use cookies, access SMART through this page.
You can easily change modes later, by clicking on the links in the 'SMART MODE' header box, or in your personal preference settings ('SETUP' link in the menu):
Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Letunic et al. (2004) Nucleic Acids Res 32, D142-D144
HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

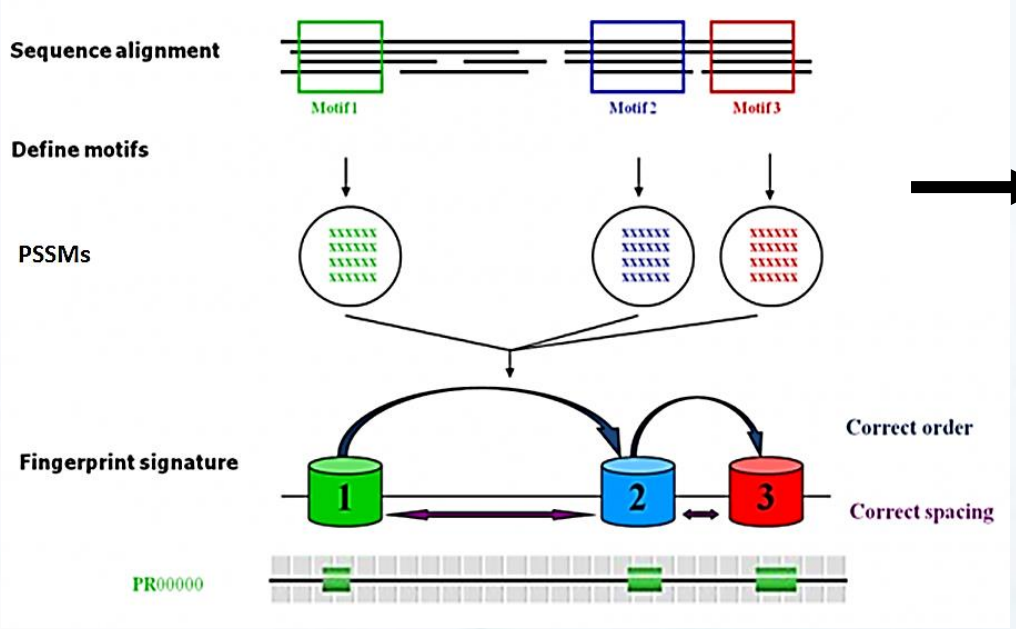
La BD SMART

2. FINGERPRINTS



Normalmente, los alineamientos de secuencias contienen no una, sino varias regiones conservadas. Por lo cual tiene sentido utilizar tantas de estas regiones como sea posible para crear una firma para un alineamiento de familias de proteínas.

Conduce a una mayor probabilidad de identificar a un pariente lejano, ya sea que todas las partes de la firma o estén o no emparejados.



La clave es la puntuación recibida al emparejar cualquier aminoácido dado en una posición determinada. La puntuación depende del conjunto de aminoácidos que ocurrieron en esa posición en la familia de proteínas que estamos estudiando.

2. FINGERPRINTS

Con matrices de frecuencias de residuos NO ponderados

YVTVQH**K**K**L**RT**P**L
YVTVQH**K**K**L**RT**P**L
YVTVQH**K**K**L**RT**P**L
AATMKF**K**K**L**RH**P**L
AATMKF**K**K**L**RH**P**L
YIFATT**K**S**L**RT**P**A
VATLRY**K**K**L**RQ**P**L
YIFGGT**K**S**L**RT**P**A
WVFSAA**K**S**L**RT**P**S
WIFSTS**K**S**L**RT**P**S
YLFSKT**K**S**L**QT**P**A
YLFTKT**K**S**L**QT**P**A

El primer método para explotar este enfoque fue la FingerPrintings *(Attwood, Eliopoulos y Findlay 1991, Parry-Smith y Attwood 1992, Attwood y Findlay 1993).*

La información de secuencia que contienen se convierte en matrices construidas solo por las frecuencias de residuos observados en cada posición de los motivos

Se dice que este tipo de sistema de puntuación no está ponderado ya que no utilizan sistemas de puntuaciones adicionales (por ejemplo, de matrices de mutación o sustitución).

Se cuentan las veces que aparece cada uno de los 20 aminoácidos en cada columna del alineamiento y se elabora una matriz de frecuencias (PSSM). Se trata de un sistema de puntuación específico de la posición que no está ponderado.

T	C	A	G	N	S	**P**	F	**L**	Y	H	Q	V	**K**	D	E	I	W	R	M	B	X	Z
0	0	2	0	0	0	0	0	0	7	0	0	1	0	0	0	0	2	0	0	0	0	0
0	0	3	0	0	0	0	0	2	0	0	0	4	0	0	0	3	0	0	0	0	0	0
0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	3	0	0	1	0	0	0	3	0	0	0	0	0	0	2	0	0	0
0	0	1	1	0	0	0	0	0	0	0	3	0	4	0	0	0	0	1	0	0	0	0
0	0	1	0	0	1	0	2	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0	0	0	0	0
0	0	0	0	0	6	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	**10**	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	4	0	0	2	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2. FINGERPRINTS

¿Como se construye a DB?

- **La mayoría de las familias de proteínas se caracterizan por > 1 motivo**
  - Se utilizan todos para construir una firma de diagnóstico
- **Este es el principio de las FINGERPRINTS (huellas dactilares)**
  - Estos ofrecen una fiabilidad diagnóstica mejorada en virtud del contexto biológico proporcionado por los motivos vecinos
- **Los motivos se extraen de los alineamientos a mano y se recodifican como alineamientos locales no ponderadas y sin espacios.**
  - la información de residuos se aumenta mediante búsquedas iterativas
  - las secuencias que coinciden con todos los motivos que no estaban en el alineamiento original se agregan a los motivos, y la base de datos se reconstruye
- **El proceso se repite hasta la convergencia**
  - los resultados se anotan manualmente antes de su inclusión en la base de datos

2. FINGERPRINTS

Home | Databases | Services and Tools | EU Projects | Education | Research Group | Videos | Societies

PRINTS

http://130.88.97.239/PRINTS/index.php

PRINTS is a compendium of protein **fingerprints**. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. References

Direct PRINTS access:

- By accession number
- By PRINTS code
- By database code
- By text
- By sequence
- By title
- By number of motifs
- By author
- By query language

PRINTS search:

- FPScan - search PRINTS with a query sequence/ID
- GRAPHScan - search a sequence with a named fingerprint
- FingerPRINTScan source is available: contact attwood@bioinf.man.ac.uk

PRINTS BLAST search

- Run a BLAST search of sequences in PRINTS

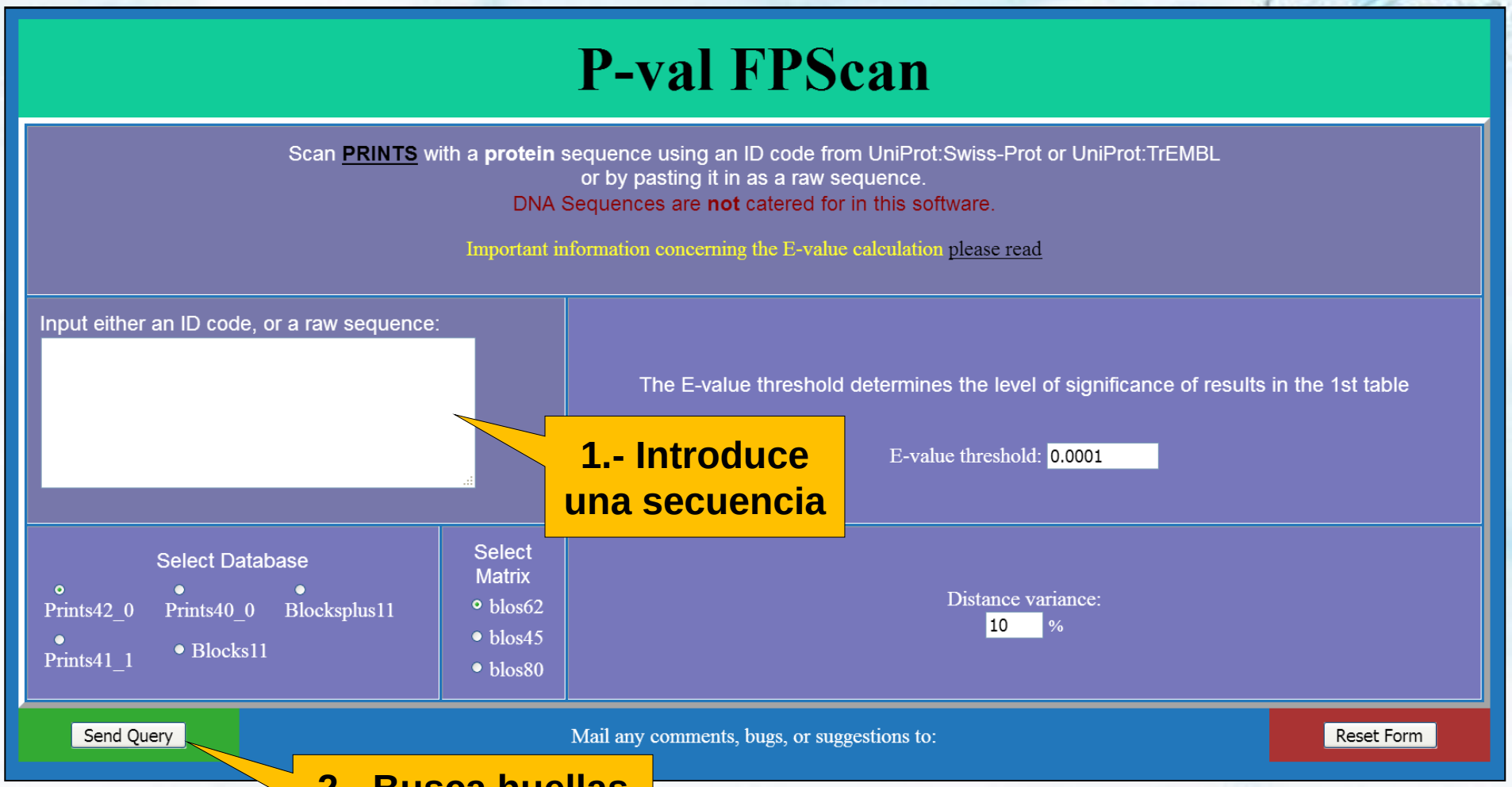
BLOCKS/PRINTS Search:

- Search by **user** query sequence
- About BLOCKS

La BD PRINTS

2. FINGERPRINTS

http://130.88.97.239/cgi-bin/dbbrowser/fingerPRINTScan/FPScan_fam.cgi



La herramienta FPScan

3. BLOCKS

Con matrices de residuos ponderados con PAM

YVTVQH**K**K**L**RT**P**L
YVTVQH**K**K**L**RT**P**L
YVTVQH**K**K**L**RT**P**L
AATMKF**K**K**L**RH**P**L
AATMKF**K**K**L**RH**P**L
YIFATT**K**S**L**RT**P**A
VATLRY**K**K**L**RQ**P**L
YIFGGT**K**S**L**RT**P**A
WVFSAA**K**S**L**RT**P**S
WIFSTS**K**S**L**RT**P**S
YLFSKT**K**S**L**QT**P**A
YLFTKT**K**S**L**QT**P**A

Otra versión de FINGERPRINTS son los BLOCKS que surgen de ponderar por BLOSSUM la matriz. En consecuencia, las búsquedas de bases de datos usando dicha matriz función con altos niveles de ruido y una especificidad relativamente baja.

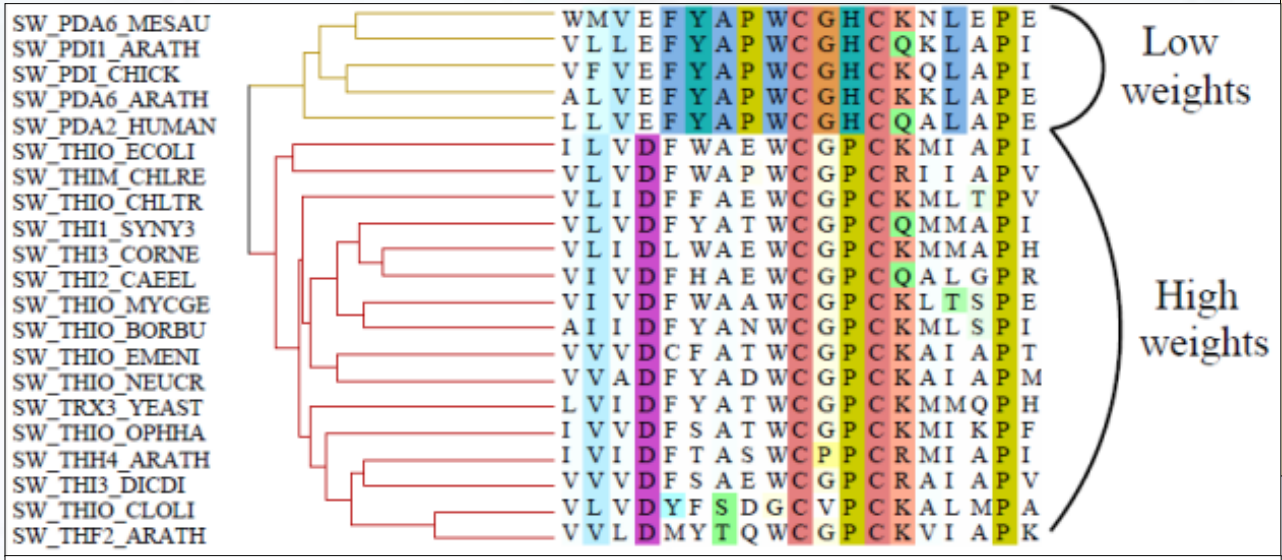
Permite una puntuación más alta en parientes lejanos pero inevitablemente también genera coincidencias aleatorias.

T	C	A	G	N	S	P	**F**	**L**	Y	H	Q	**V**	K	D	E	**I**	W	R	**M**
–29	–22	–29	–48	–24	–24	–46	40	–13	62	–10	–40	–22	–38	–44	–44	–15	16	–30	–22
–1	–32	–1	–18	–20	–10	–13	–9	20	–22	–21	–18	32	–23	–22	–20	32	–61	–26	19
0	–36	–18	–30	–24	–12	–30	36	0	24	–18	–36	–6	–30	–36	–30	6	–30	–30	–6
3	–29	3	–4	–10	–1	–7	–22	3	–31	–19	–15	14	–12	–15	–13	11	–52	–15	11
3	–48	–1	–8	7	1	–4	–54	–31	–46	6	14	–17	23	6	5	–20	–48	14	–9
2	–27	–7	–19	–3	–5	–13	0	–16	6	8	–10	–11	–15	–13	–11	–7	–37	–12	–15
0	–60	–12	–24	12	0	–12	–60	–36	–48	0	12	–24	60	0	0	–24	–36	36	0
6	–30	0	–6	12	12	0	–48	–36	–42	–6	0	–18	30	0	0	–18	–30	18	–12
–24	–72	–24	–48	–36	–36	–36	**24**	**72**	–12	–24	–24	**24**	–36	–48	–36	**24**	–24	–36	**48**
–12	–50	–20	–32	2	–2	0	–50	–34	–48	26	18	–24	32	–6	–6	–24	10	62	–2
24	–29	7	–5	5	6	0	–36	–24	–31	6	1	–6	1	4	4	–6	–56	–4	–14
0	–36	12	–12	–12	12	72	–60	–36	–60	0	0	–12	–12	–12	–12	–24	–72	0	–24
–6	–44	–2	–18	–16	–10	–12	–10	22	–24	–18	–14	10	–22	–24	–18	6	–40	–26	16

3. BLOCKS

- En situaciones en las que hay pocas secuencias disponibles, esto puede comprometer el rendimiento del diagnóstico.
- Es posible construir representaciones de motivos alternativos aplicando diferentes

Counting all sequences equally can lead to a loss of information when a sequence is copied multiple times, because it can dilute independent information from other sequences. Identical or nearly identical copies of the same sequence provide little new information.



It may be possible to mitigate this problem by giving each sequence a *weight*, with nearly identical sequences downweighted, and unusual sequence upweighted.

En un MSA, cuando hay secuencias muy parecidas que pertenecen a organismos muy próximos desde el punto de vista evolutivo, la información está sesgada. Para compensar este efecto, se aplica un sistema de ponderación que reduce el peso relativo de las secuencias muy parecidas y aumenta el de las menos representadas.

3. BLOCKS ¿Como se construye a DB?

1. Se utilizan alineamientos múltiples derivados de las regiones más conservadas sin espacios de secuencias de proteínas homólogas.
2. Los alineamientos se generan automáticamente utilizando los mismos conjuntos de datos utilizados para derivar las matrices BLOSUM
3. Los alineamientos sin huecos derivadas se denominan bloques.
4. Los bloques, que suelen ser más largos que los motivos, se convierten posteriormente en PSSM.
5. Los motivos conservados en los bloques, se localizan buscando tripletes de residuos espaciados (por ejemplo, Asn-x-x-Glu-x-x-x-x-x-x-x-x-x-Glu, donde x representa cualquier aminoácido).
6. Todos los bloques redundantes generados por este proceso se eliminan y solo se retienen los bloques con la puntuación más alta
7. Posteriormente, se aplica un esquema de ponderación
8. las puntuaciones de los bloques se calculan utilizando la matriz de sustitución BLOSUM62.
9. Los bloques encontrados por ambos métodos se consideran confiables y se almacenan en la base de datos como alineaciones locales sin espacios (Henikoff et al. 2000, Henikoff y Henikoff 1994a). .

El contenido de información de los bloques puede visualizarse examinando sus denominados logotipos de secuencia. La altura de la letra aumenta al aumentar la frecuencia del residuo, de modo que las posiciones más conservadas tienen las letras más altas.

3. BLOCKS

Blocks WWW Server

A service for biological sequence analysis at the Fred Hutchinson Cancer Research Center in Seattle, Washington, USA.

NOTICE: The Blocks Database is no longer updated. We suggest you use InterPro to annotate your sequences.
SIFT has moved and requests are being re-directed, see link below.
The other pages are no longer supported, use them at your own risk.

Blocks-Based Tools

- About Blocks
- Final Blocks Release
- View Blocks
  - Get Blocks by key word
  - Get Blocks by number
- Search a Sequence vs the Blocks Database
  - Block Searcher to search a sequence vs Blocks Help
  - Reverse PSI-BLAST Searcher to search a sequence vs Blocks using NCBI's RPS-BLAST program Help
  - Impala Searcher to search a sequence vs Blocks using NCBI's IMPALA progra
- Make Blocks
  - Block Maker to create Blocks Help
  - Multiple Alignment Processor to excise Blocks from multiple alignments
    NOTICE: The Multiple Alignment Processor has been de-activated due to abu
- Search Blocks vs Blocks or Sequences
  - LAMA Searcher to search Blocks vs Blocks Help
  - COBBLER to search embedded Blocks vs sequence databases Help
- Biassed Block Checker

La BD BLOCKS ya no se actualiza

http://blocks.fhcrc.org/

Parece que el servidor ya no existe

Servidor no encontrado

Firefox no puede encontrar el servidor en www.blocks.fhcrc.org.

- Compruebe que la dirección no tiene errores de escritura del tipo **ww**.example.com en lugar de **www**.example.com
- Si no puede cargar ninguna página, compruebe la conexión de red de su equipo.
- Si su equipo o red están protegidos por un cortafuegos o proxy, asegúrese de que Firefox tiene permiso para acceder a la web.

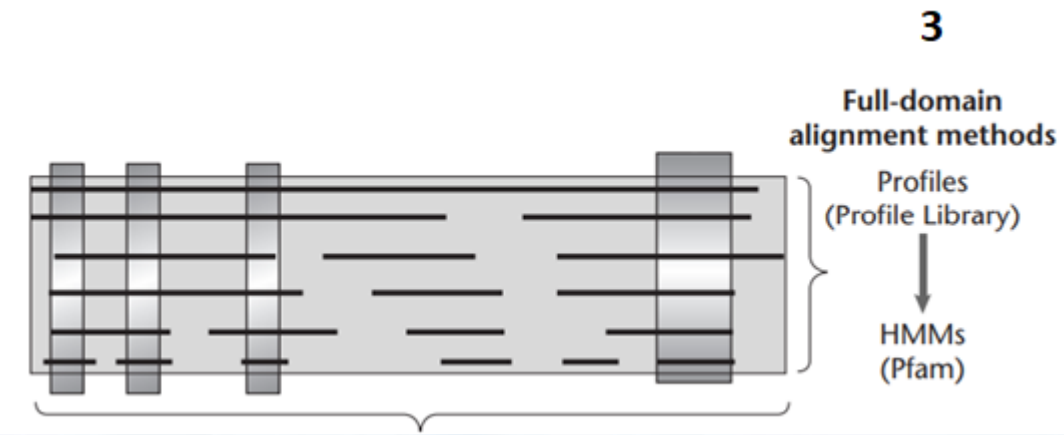
Reintentar

La BD BLOCKS

4. PROFILES

Alineamientos completos

Perfiles = Regiones conservadas + no conservadas



Se puede pensar en un perfil como una secuencia alterna de posiciones de "coincidencia" e "inserción" que contienen puntuaciones que reflejan el grado de conservación en cada posición del alineamiento.

- Se vuelca la información de las secuencias dentro de los ALINEAMIENTOS completos en tablas de puntuación o perfiles (Gribskov, McLachlan y Eisenberg 1987, Bucher y Bairoch 1994).
- El sistema de puntuación es complejo: además de ponderaciones evolutivas (por ejemplo, puntuaciones PAM), incluye penalizaciones variables para ponderar las inserciones y deleciones que ocurren dentro de los elementos centrales de la estructura secundaria; penalizaciones por extensión de coincidencia, inserción y eliminación; y puntajes de estados de transición.

4. PROFILES

El perfil tiene 23 columnas: 20 AA + z (aa desconocido) + Go + Ge

El perfil tiene tantas filas como columnas tiene el AMS local

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len
I	8	-2	5	4	5	5	-4	24	0	15	13	1	1	1	-7	2	22	21	-18	-6	4	100	100
T	13	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	29	3	-28	-14	15	100	100
L	5	-5	3	4	13	4	2	8	-4	14	12	8	-5	0	-10	0	10	10	-1	5	2	22	22
S	17	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	34	19	1	-8	-15		100	100
T	15	22	0	-1	-5	12	-2	7	-3	-8	-6	5	7	-8	-7	16	29	9	-22	6	-4	100	100
T	8	12	-2	0	5	6	-4	19	-4	8	5	-1	2	-8	-8	7	22	19	-15	4	-3	100	100
C	17	24	-1	-3	11	8	-1	7	-10	1	-2	1	-3	-8	-14	8	5	9	-5	14	-7	100	100
V	11	18	-1	-2	2	14	-10	26	-4	9	7	-3	7	-7	-7	21	10	31	-19	-5	-5	100	100
C	10	15	-11	-11	6	8	-7	11	-10	4	3	-7	0	-11	-4	11	5	15	-22	14	-11	100	100
V	7	-3	8	8	-3	11	1	20	-1	14	10	4	2	8	-5	0	5	26	-24	-6	8	100	100

El perfil es un fiel reflejo de las secuencias de partida

El problema de las pseudocuentas: si algún AA no aparece en el AMS inicial no quiere decir que no haya alguna secuencia relacionada que sí lo tenga.

El perfil que define el AMS

4. PROFILES

Name and characterization of the entry

Description: G-protein coupled receptors family 1 profile.

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=259;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=254;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.9359; R2=0.02006056; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=327; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=227; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-100; E1=-100; MI=-105; MD=-105; IM=-105; DM=-105; MM=1; MO=-10;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:         B1=0; BI=-105; BD=-105;
/M: SY='G'; M=  1,-11,-24,-13,-15,-19, 30,-19,-20,-18,-15,-11, -5,-19,-16,-18, -2,-11,-15,-22,-20,-16;
/M: SY='N'; M= -9, 33,-19, 15, -2,-18, -2,  8,-18, -2,-26,-18, 51,-20, -1, -2,  9,  1,-26,-38,-17, -2;
/M: SY='I'; M= -1,-21,-16,-26,-21,  0,-16,-22, 10,-22,  8,  4,-17,-22,-19,-20, -8, -3, 10,-21, -7,-20;
/M: SY='L'; M= -6,-24,-17,-28,-21,  8,-25,-20, 14,-24, 23, 12,-21,-25,-20,-19,-17, -5, 11,-17,  0,-20;
/M: SY='V'; M= -1,-19,-16,-23,-21, -2,-24,-14, 15,-18,  6,  7,-16,-24,-18,-17, -7, -1, 21,-26, -5,-20;
/M: SY='I'; M= -8,-28,-16,-33,-25,  6,-31,-24, 26,-26, 24, 15,-24,-26,-21,-23,-20, -8, 20,-20, -1,-24;
/M: SY='I'; M= -6,-23,-19,-27,-21,  5,-24,-16,  8,-19,  8,  5,-20,-23,-17,-17,-15, -6,  5, -2,  7,-19;
/M: SY='V'; M=  4,-22,-14,-26,-21, -5,-23,-23, 16,-18,  7,  6,-19,-22,-18,-19, -7, -1, 20,-24, -8,-21;
/M: SY='I'; M= -8,-25,-19,-30,-24, 14,-29,-20, 19,-24, 19, 11,-21,-25,-21,-20,-17, -4, 16,-16,  5,-23;
/M: SY='F'; M= -3,-18,-12,-23,-18,  4,-20,-16,  2,-17,  3,  1,-14,-22,-16,-14, -7,  0,  3,-16,  0,-17;
/M: SY='R'; M= -9,-10,-22,-12, -6,-10,-18, -7,-15,  7,-11, -5, -5,-17, -1, 14, -6, -4,-11,-18, -5, -5;
/M: SY='K'; M= -8,  1,-21, -1,  0,-14,-16, -1,-18,  4,-17,-10,  3,-12, -1,  3, -1, -1,-15,-23, -5, -1;
/M: SY='R'; M= -8, -7,-25, -7,  0,-19,-16, -6,-19, 11,-18, -7, -3, -5,  2, 13, -3, -4,-15,-23,-10,  0;
/M: SY='R'; M= -8, -7,-21, -9, -3,-16,-16, -4,-14,  7,-12, -1, -3,-14,  3, 11, -4, -4,-11,-23, -8, -1;
/I:         I=-4; MD=-23;
/M: SY='L'; M= -7,-17,-20,-19,-11, -2,-20,-11,  1, -8, 11,  8,-14,-19, -8, -1,-14, -7,  0,-18, -3,-10; D=-4;
/I:         I=-4; MI=0; MD=-23; IM=0; DM=-23;
/M: SY='R'; M=-11, -5,-24, -7, -1,-18,-17,  7,-20,  8,-16, -6,  1,-15,  6, 17, -5, -6,-17,-22, -6,  0;
/M: SY='T'; M= -2,  3,-16, -3, -3,-17,-11, -7,-17, -1,-20,-13,  9,-13, -1,  0, 15, 16,-11,-31,-13, -2;
/M: SY='P'; M= -1,-13,-21,-13, -8,-17,-15,-16, -8, -8,-13, -7,-10,  8,-10,-10,  1,  2, -4,-29,-17,-11;
/M: SY='T'; M= -3,-14,-14,-20,-15, -2,-20,-15,  2,-14, -1,  3,-11,-14,-13,-13, -1,  7,  4,-23, -5,-15;
/M: SY='N'; M= -9, 11,-22,  4, -4,-10, -9,  5,-15, -5,-16,-11, 19,-20, -4, -4,  1, -2,-18,-23, -1, -5;
/M: SY='I'; M= -9,-24,-20,-29,-23, 12,-29,-16, 17,-21, 14, 10,-20,-24,-19,-18,-16, -6, 12,-11, 11,-22;
/M: SY='F'; M=-15,-26,-22,-31,-23, 34,-30,-11,  9,-24, 18,  7,-22,-27,-23,-18,-21, -9,  3,  1, 26,-23;
/M: SY='L'; M= -9,-27,-20,-31,-23,  4,-30,-20, 25,-24, 26, 21,-23,-24,-17,-20,-21, -8, 17,-21, -1,-21;
/M: SY='V'; M=  1,-19, -8,-24,-19, -2,-18,-20,  5,-19,  7,  4,-17,-22,-17,-18, -7, -1,  8,-23, -8,-18;
/M: SY='N'; M=  1,  8,-16, -1, -6,-15, -2,  0,-15, -9,-19,-13, 19,-18, -5, -8, 10,  2,-15,-30,-14, -6;
/M: SY='L'; M= -9,-27,-19,-29,-19,  7,-29,-18, 18,-25, 36, 18,-25,-27,-17,-17,-24, -8, 10,-20, -1,-19;
/M: SY='A'; M= 29, -9,  0,-16,-11,-17, -5,-18,-12,-12,-14,-12, -6,-14,-11,-18, 13,  5, -2,-26,-18,-11;
/M: SY='I'; M= -4,-26,-13,-31,-25, 10,-28,-24, 20,-24, 17, 10,-22,-26,-23,-21,-15, -5, 20,-20, -1,-25;
/M: SY='A'; M= 17,-11, -3,-17,-12,-10, -7,-18,-10,-15,-11,-10, -6,-16,-12,-17, 10,  5, -2,-25,-14,-12;
/M: SY='D'; M=-17, 42,-28, 57, 17,-37, -9,  2,-36, -1,-28,-26, 20,-11,  2, -9,  1, -9,-29,-38,-18,  9;
/M: SY='L'; M= -8,-27,-19,-31,-22, 13,-28,-21, 19,-26, 27, 14,-23,-26,-21,-21,-20, -7, 13,-17,  1,-22;
/M: SY='L'; M= -7,-25,-10,-30,-23,  9,-24,-21, 11,-25, 20, 10,-22,-27,-22,-20,-18, -8,  9,-17, -2,-22;
/M: SY='F'; M= -5,-21,-13,-25,-20,  9,-22,-14,  7,-19,  8,  7,-17,-24,-18,-16,-11, -3,  9,-17,  4,-19;
/M: SY='A'; M=  2,-16, -7,-20,-17, -8,-10,-18, -1,-19,  2,  1,-12,-21,-15,-18, -3, -1,  2,-26,-12,-16;
/M: SY='L'; M= -4,-23,-10,-27,-21,  4,-25,-22, 13,-24, 16,  7,-19,-24,-20,-20,-11,  0, 13,-23, -5,-21;
```

Un perfil de la BD PROSITE

4. PROFILES

- Profiles are scoring tables derived from full alignments
  - these define which residues are allowed at given positions
  - which positions are conserved & which degenerate
  - which positions, or regions, can tolerate insertions
  - the scoring system is intricate, & may include evolutionary weights, results from structural studies, & data implicit in the alignment
  - variable penalties are specified to weight against INDELs occurring in core 2' structure elements
- Within a profile, the I & M fields contain position-specific scores for insert & match positions
  - in conserved regions, INDELs aren't totally forbidden, but are strongly impeded by large penalties defined in the DEFAULT field
  - these are superseded by more permissive values in gapped regions
  - the inherent complexity of profiles renders them highly potent discriminators, but they are time-consuming to derive

Características de un perfil

4. PROFILES

Ventajas

1.- Localiza homólogos distantes (con poca conservación de la secuencia

2.- Caracteriza la secuencia completa del dominio, no sólo la región más conservada

3.- Son más adecuados para predecir características estructurales de las proteínas

Inconvenientes

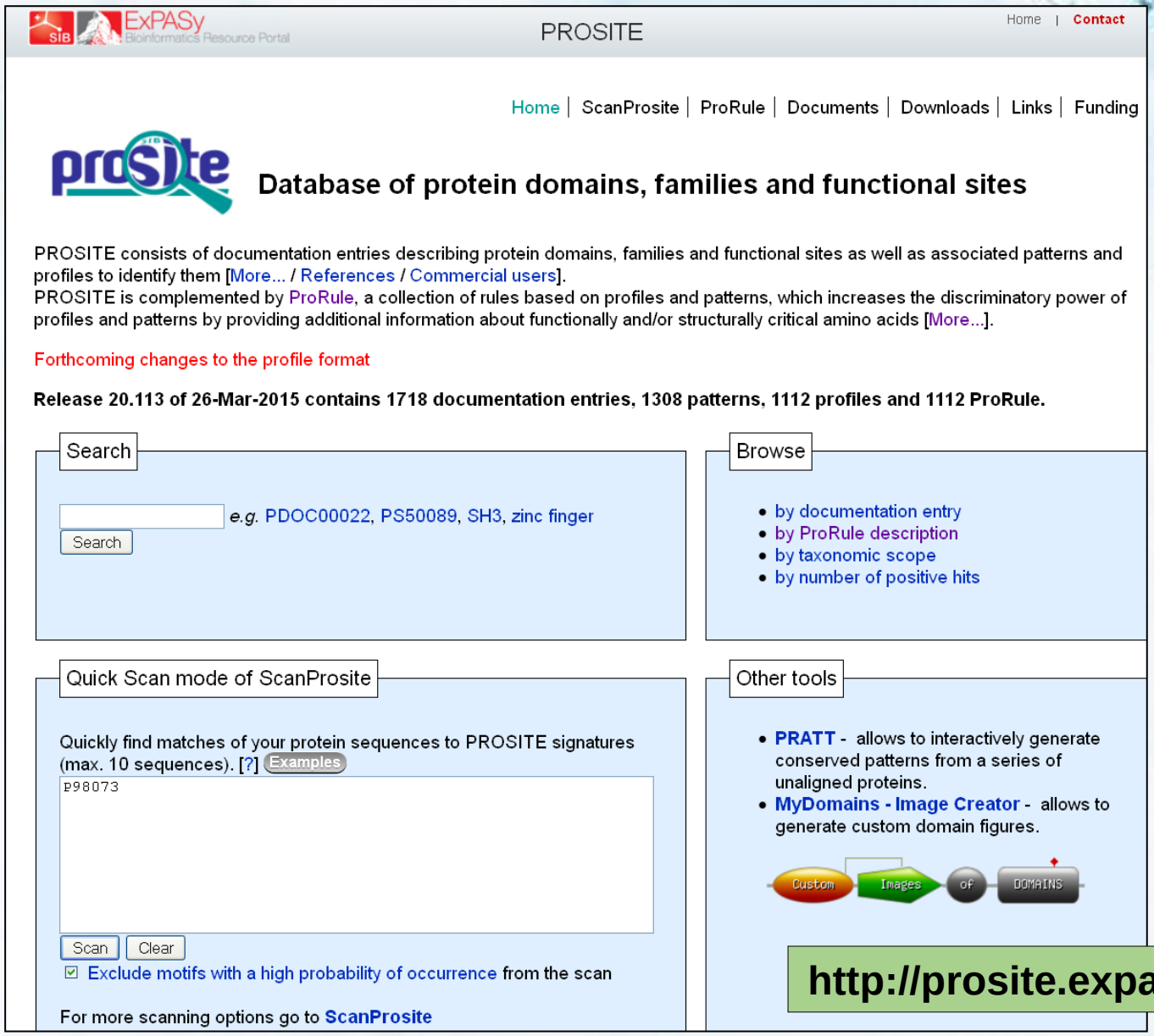
1.- Son difíciles de construir

2.- Son menos adecuados para la detección de una función biológica determinada

La complejidad inherente de los perfiles los convierte en discriminadores muy potentes. Se utilizan para complementar algunas de las expresiones regulares más pobres en PROSITE y / o para proporcionar una alternativa de diagnóstico donde la divergencia extrema de secuencia hace que el uso de expresiones regulares sea inapropiado.

Ventajas e inconvenientes de los perfiles

4. PROFILES



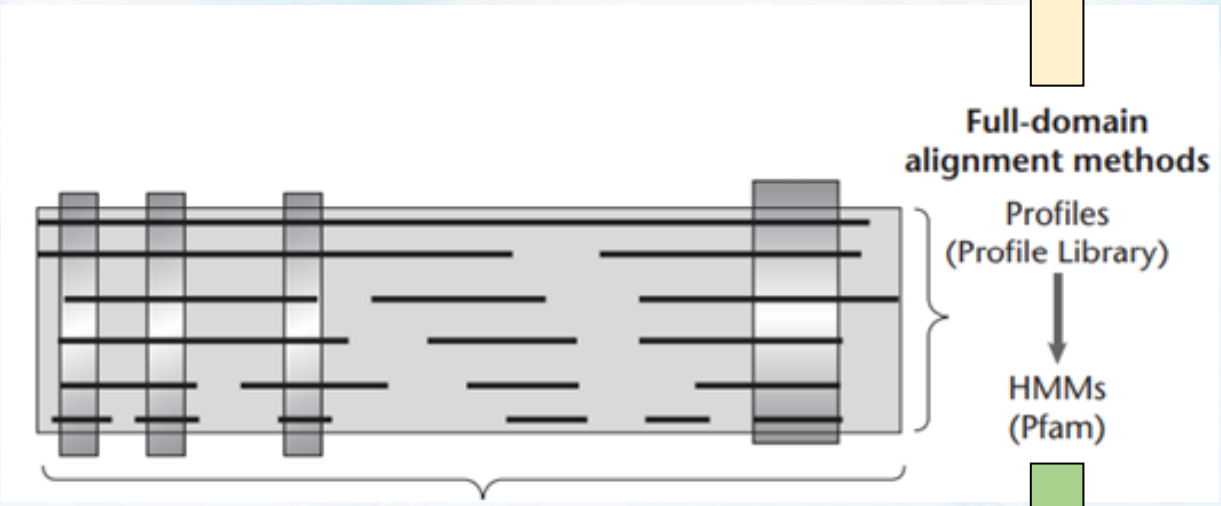
http://prosite.expasy.org/prosite.html

La BD PROSITE

Alineamientos completos

Los perfiles utilizan la información presente en las regiones conservadas, en los huecos y en las regiones no conservadas.

Por matrices ponderadas



Por Machine Learning

Los modelos de Markov ocultos son modelos probabilísticos que se obtienen a partir de los alineamientos completos

Perfiles y modelos de Markov ocultos

5. Modelos de Markov Ocultos (HMM)

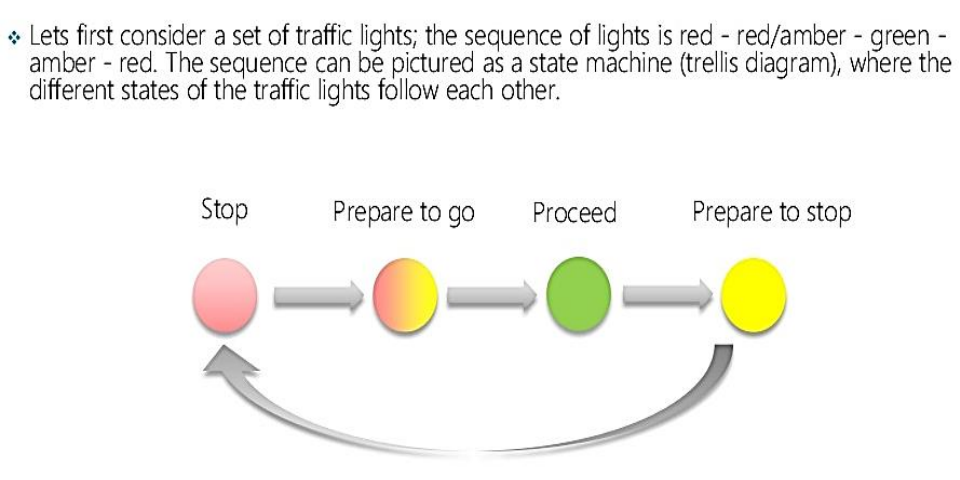
USING MACHINE LEARNING FOR PATTERN RECOGNITION IN BIOINFORMATICS

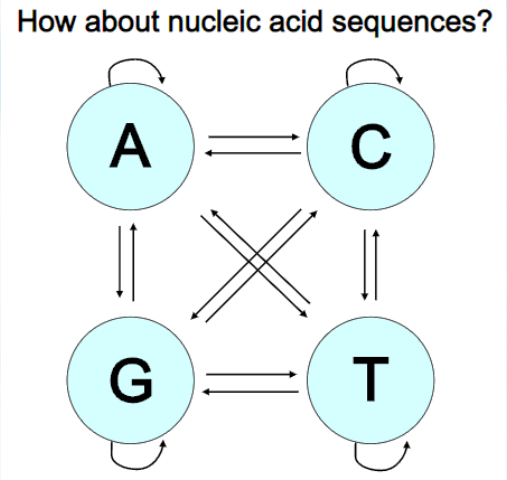
HMM, un método poderoso utilizado para describir las propiedades estadísticas de secuencias individuales y familias de secuencias

Un modelo de Markov, también conocido como cadena de Markov, **describe una secuencia de eventos que ocurren uno tras otro en una cadena.** Cada evento determina la probabilidad del próximo evento. Una cadena de Markov se puede considerar como un proceso que se mueve en una dirección de un estado al siguiente con una cierta probabilidad, lo que se conoce como probabilidad de transición.

5. Modelos de Markov Ocultos (HMM)

USING MACHINE LEARNING FOR PATTERN RECOGNITION IN BIOINFORMATICS





Un buen ejemplo de un modelo de Markov es el cambio de señal de los semáforos en el que el estado de la señal actual depende del estado de la señal anterior (por ejemplo, la luz verde se enciende después de la luz roja, que se enciende después de la luz amarilla).

Las secuencias biológicas escritas como cadenas de letras también pueden describirse mediante cadenas de Markov: cada letra que representa un estado está vinculada con valores de probabilidad de transición.

La descripción de secuencias biológicas usando cadenas de Markov permite el cálculo de valores de probabilidad para un residuo dado de acuerdo con las frecuencias de distribución únicas de nucleótidos o aminoácidos.

5. Modelos de Markov Ocultos (HMM)

Por ejemplo:

Supongamos que estamos interesados en proteínas de membrana y deseamos predecir qué partes de las secuencias son hélices transmembrana (TM) y qué partes son bucles a cada lado de la membrana.
Usemos la información que tenemos a nuestro favor

Una característica importante de las hélices es que tienen un alto contenido de residuos de aminoácidos hidrófobos. Esto debería proporcionar alguna información con la que distinguir posibles regiones helicoidales en una secuencia de estructura desconocida.

Sean P_{helix} y P_{coil} las frecuencias del aminoácido a en las hélices y los bucles, respectivamente. Estas frecuencias se pueden medir en un conjunto de proteínas de estructura conocida.

Ahora considere una sección de amino ácidos tomados de una proteína desconocida, y sea A_s el aminoácido en la posición S en esta secuencia:

Las probabilidades **L1 y L0** de que la secuencia ocurra en una región helicoidal o no helicoidal son solo el producto de las frecuencias de los aminoácidos en la secuencia:
Es decir que depende de las frecuencias de los amino ácidos en cada posición

La relación L1 / L0 es más informativa: nos dice si es más o menos probable que la secuencia esté en una región helicoidal o no helicoidal

5. Modelos de Markov Ocultos (HMM)

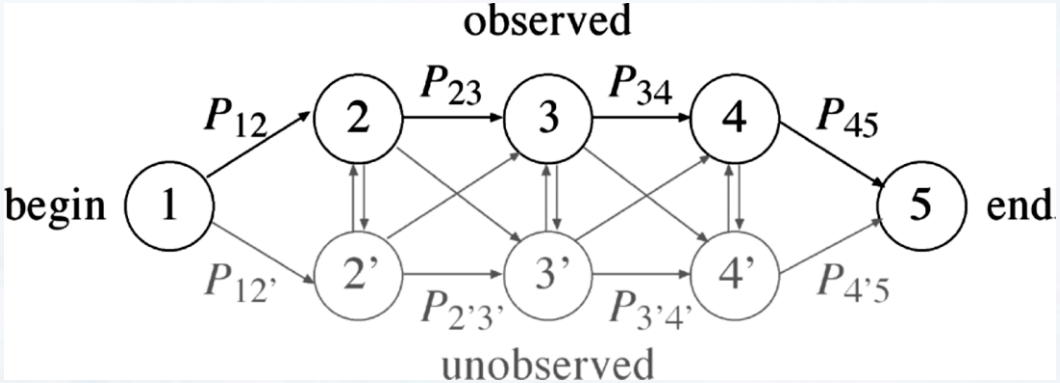
En el modelo de HMM, todos los estados en una secuencia lineal son directamente observables. En algunas situaciones, algunos factores no observados influyen en los cálculos de transición de estado.

En un **HMM** la **probabilidad de pasar de un estado a otro es la probabilidad de transición**. Cada estado puede estar compuesto por varios elementos o símbolos:

- Para las secuencias de nucleótidos, hay cuatro símbolos posibles, A, T, G y C, en cada estado.
- Para las secuencias de aminoácidos, hay veinte símbolos.

El valor de probabilidad asociado con cada símbolo en cada estado se llama probabilidad de emisión.

Para calcular la probabilidad total de una trayectoria particular del modelo, se deben tener en cuenta tanto las **probabilidades de transición como las de emisión que vinculan todos los estados "ocultos" y observados**.



5. Modelos de Markov Ocultos (HMM)

- Para construir un HMM funcional que pueda usarse para representar mejor un alineamiento de secuencia, el modelo estadístico tiene que ser "entrenado“: es un proceso para obtener los parámetros estadísticos óptimos en el HMM.
- El proceso de entrenamiento implica el cálculo de las frecuencias de los residuos en cada columna en el alineamiento múltiple construida a partir de un conjunto de secuencias relacionadas.
- Los valores de frecuencia se utilizan para completar los valores de probabilidad de emisión y transición en el modelo.

Para construir el perfil HMM hay que determinar las probabilidades de emisión de cada estado y las probabilidades de transición de un estado a otro. Estos hay que estimarlos a partir del AMS.

Lo ideal es partir de un MSA que incluya entre 20 y 100 secuencias homólogas.

5. Modelos de Markov Ocultos (HMM)

From Profile to HMM

	1	2	3	4	5	6	7	8
Alignment	A	C	D	E	F	A C	A	D	F
	A	F	D	A	-	- -	C	C	F
	A	-	-	E	F	D -	F	D	C
	A	C	A	E	F	- -	A	-	C
	A	D	D	E	F	A A	A	D	F

Se parte de un alineamiento de secuencias de proteínas relacionadas

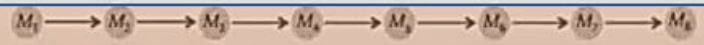
	1	2	3	4	5	6	7	8
*Alignment**	A	C	D	E	F	A	D	F
	A	F	D	A	-	C	C	F
	A	-	-	E	F	F	D	C
	A	C	A	E	F	A	-	C
	A	D	D	E	F	A	D	F

Se eliminan las columnas si los INDEL sobrepasan un umbral Ø

PROFILE(*Alignment**)	1	2	3	4	5	6	7	8
A	1	0	0	1/5	0	3/5	0	0
C	0	2/4	0	0	0	1/5	1/4	2/5
D	0	1/4	3/4	0	0	0	3/4	0
E	0	0	0	4/5	0	0	0	0
F	0	1/4	0	0	1	1/5	0	3/5

Se construye un PROFILE de dicho alineamiento

HMM diagram



A	D	D	A	F	F	D	F
1	* .25	* .75	* .20	* 1	* .20	* .75	* .60

Se construye un HMM correspondiente a dicho PROFILE

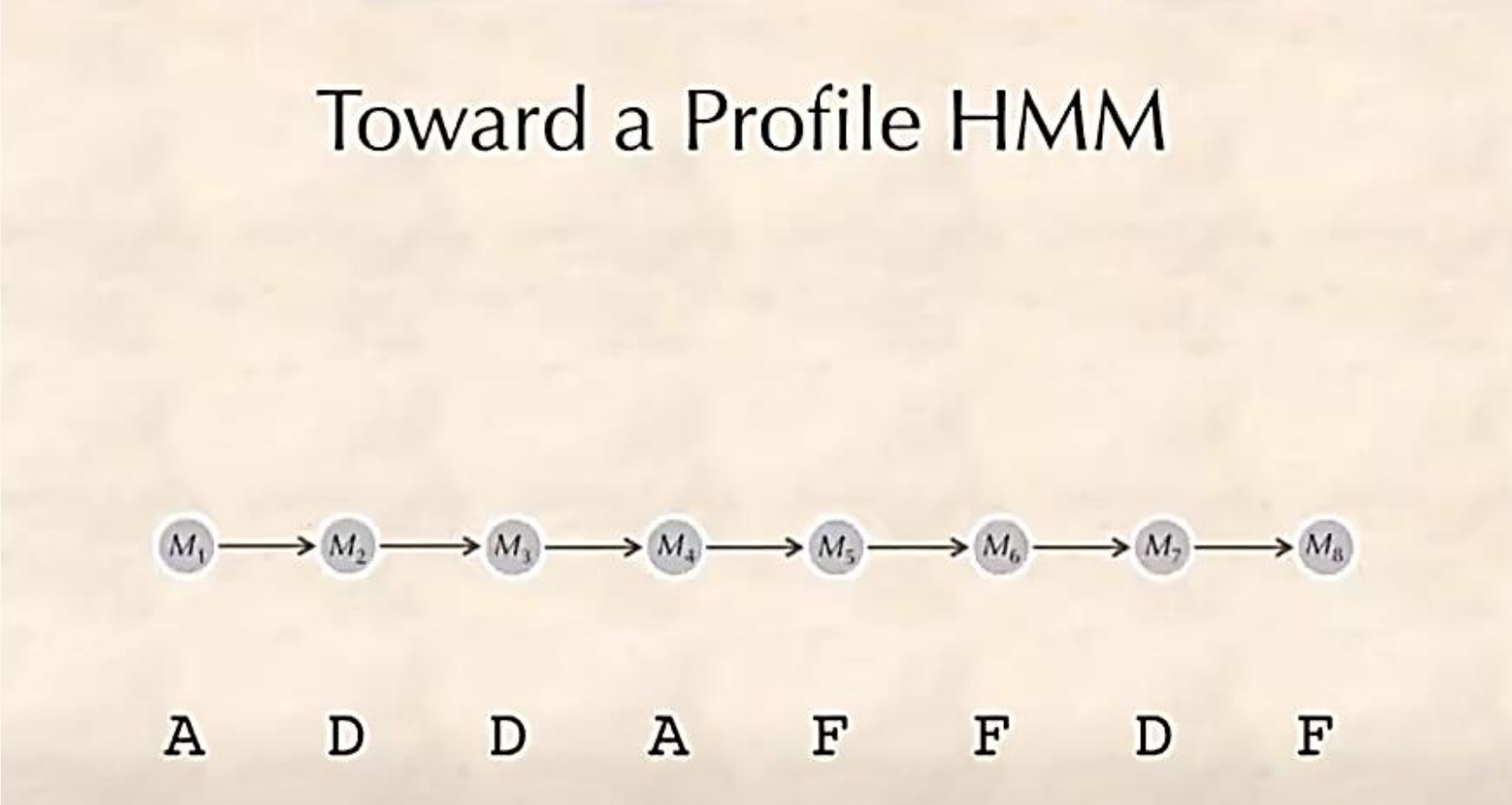
Solo tengo valores para las **emisiones**: Los valores de la emisiones est{a dado por las frecuencias

la probabilidad de transición es 1: Solo tengo una dirección

¡Que me falta modelar? → **Inserciones y Deleciones**

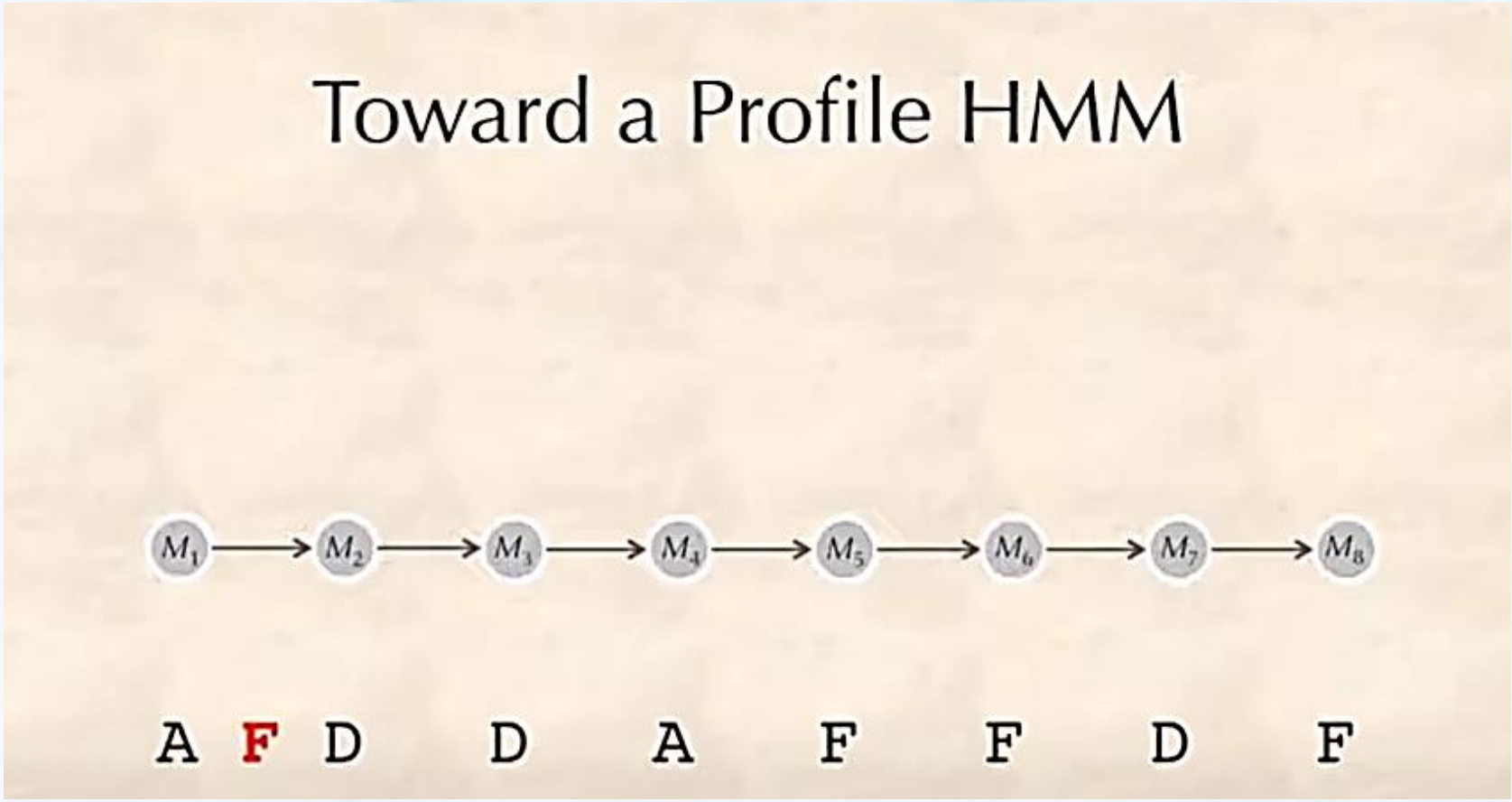
5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la inserciones?



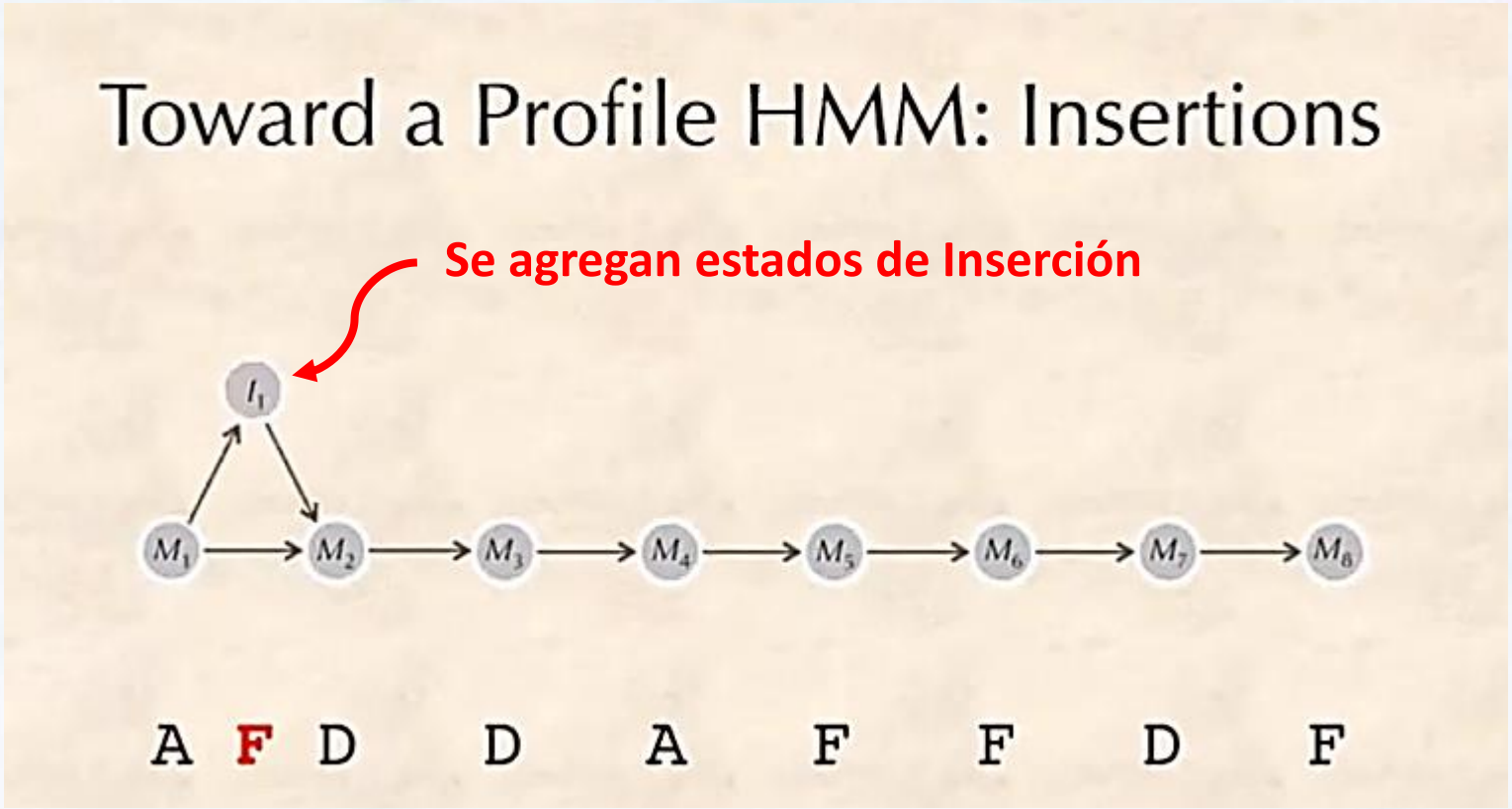
5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la inserciones?



5. Modelos de Markov Ocultos (HMM)

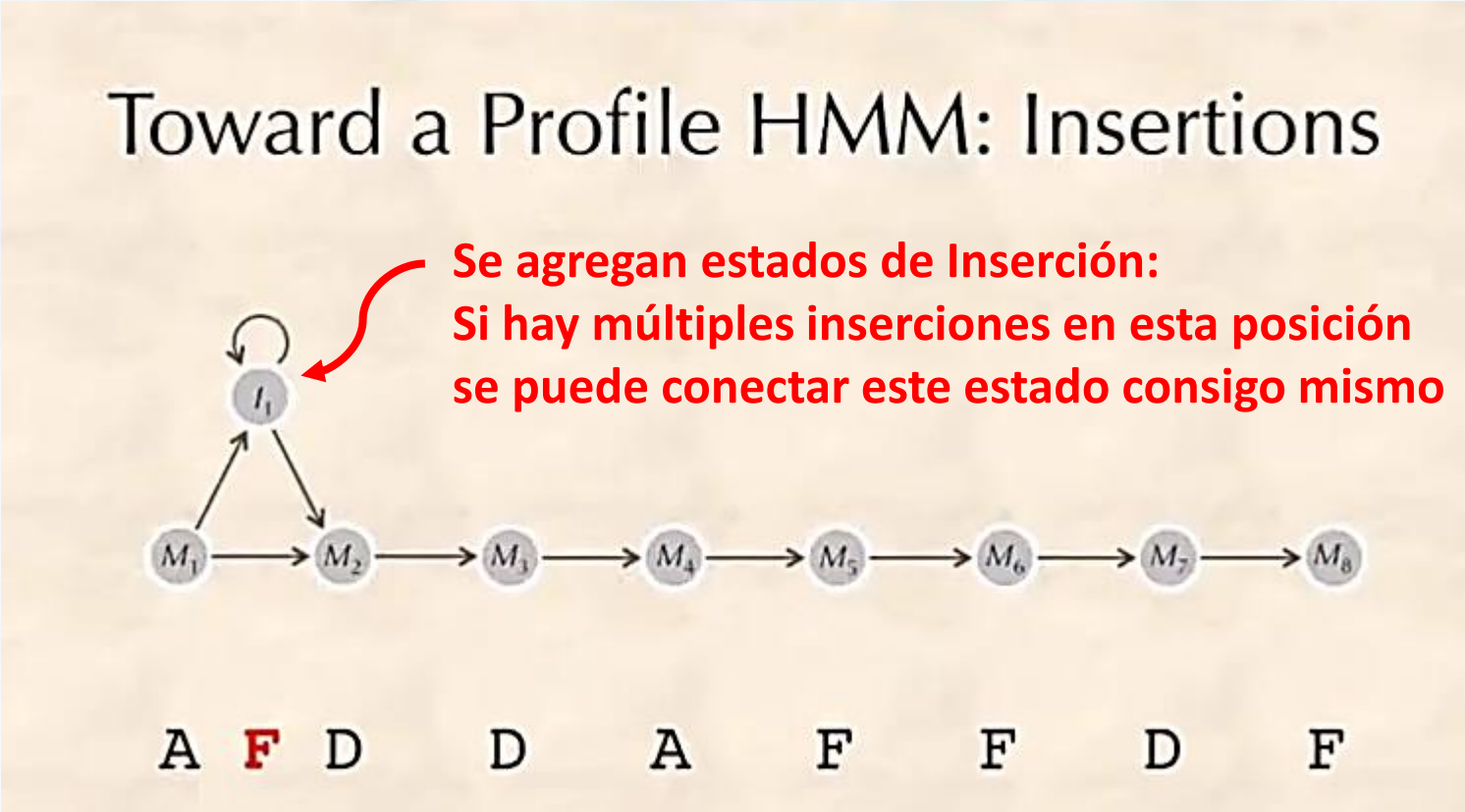
¿Como modelamos la inserciones?



Esto cambia las probabilidades de transición ya que M1 ahora tiene 2 posibilidades: pasar al estado I1 o M2

5. Modelos de Markov Ocultos (HMM)

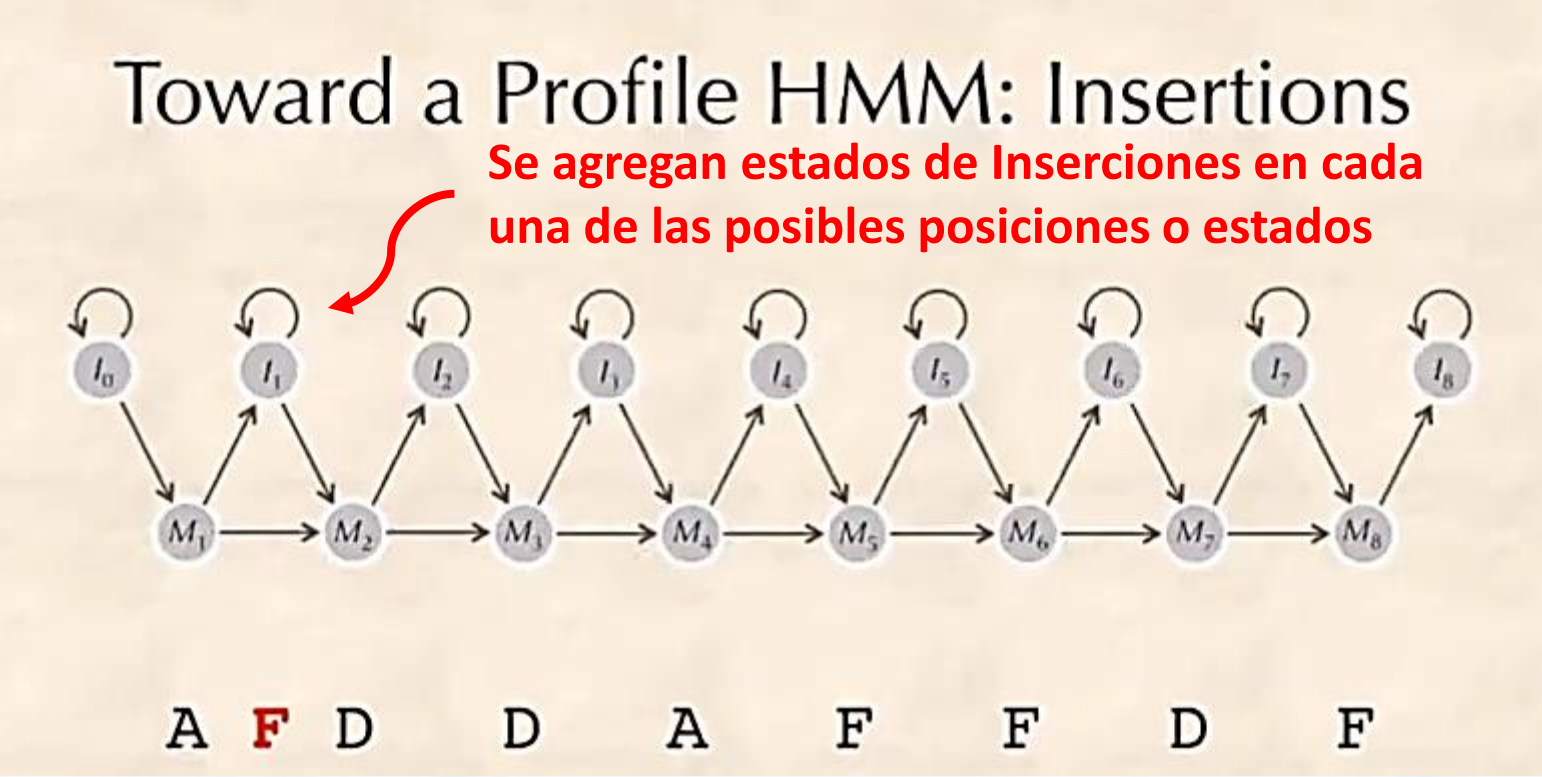
¿Como modelamos la inserciones?



Esto cambia las probabilidades de transición ya que M1 ahora tiene 2 posibilidades: pasar al estado I1 o M2

5. Modelos de Markov Ocultos (HMM)

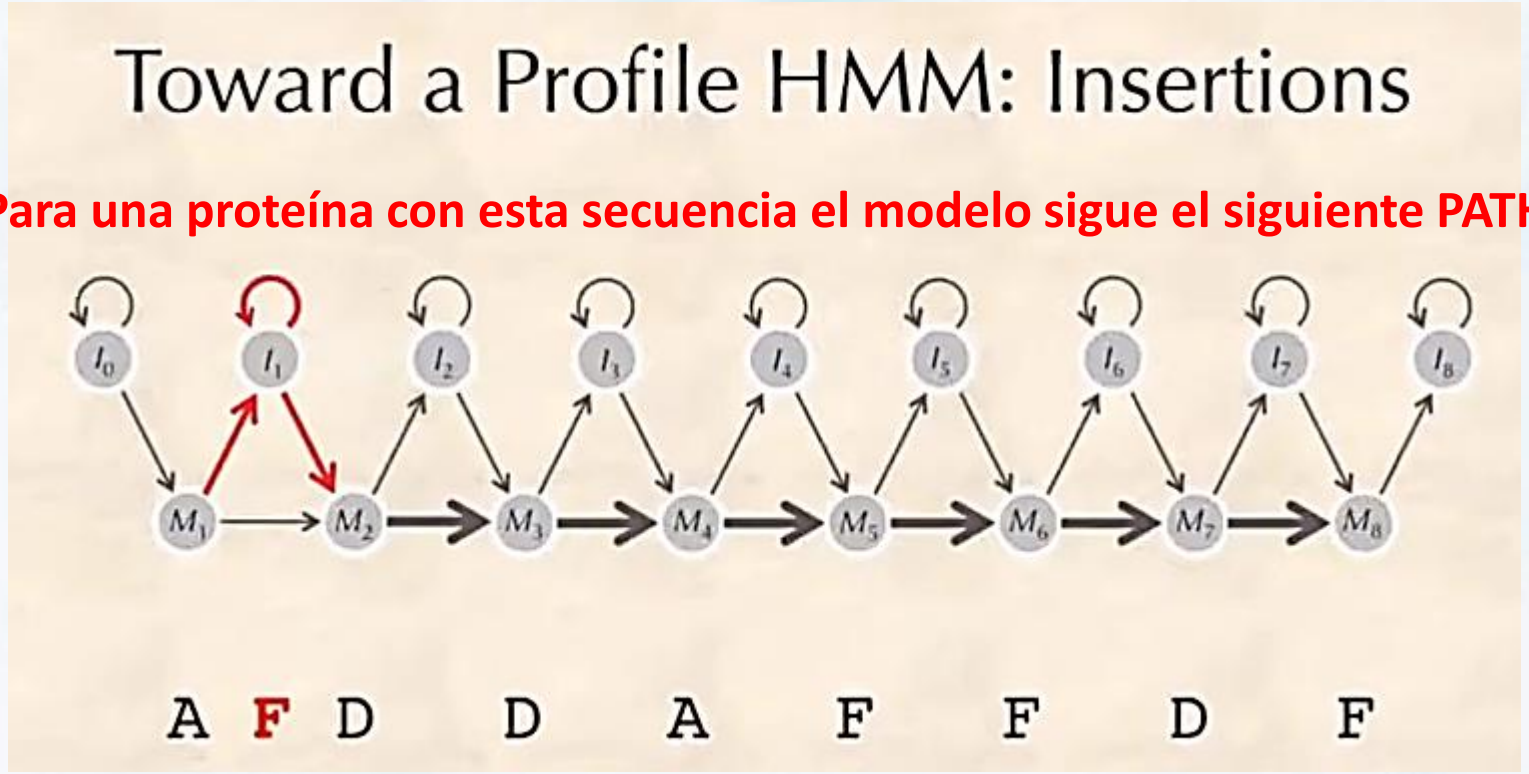
¿Como modelamos la inserciones?



Esto cambia las probabilidades de transición ya que M1 ahora tiene 2 posibilidades: pasar al estado I1 o M2

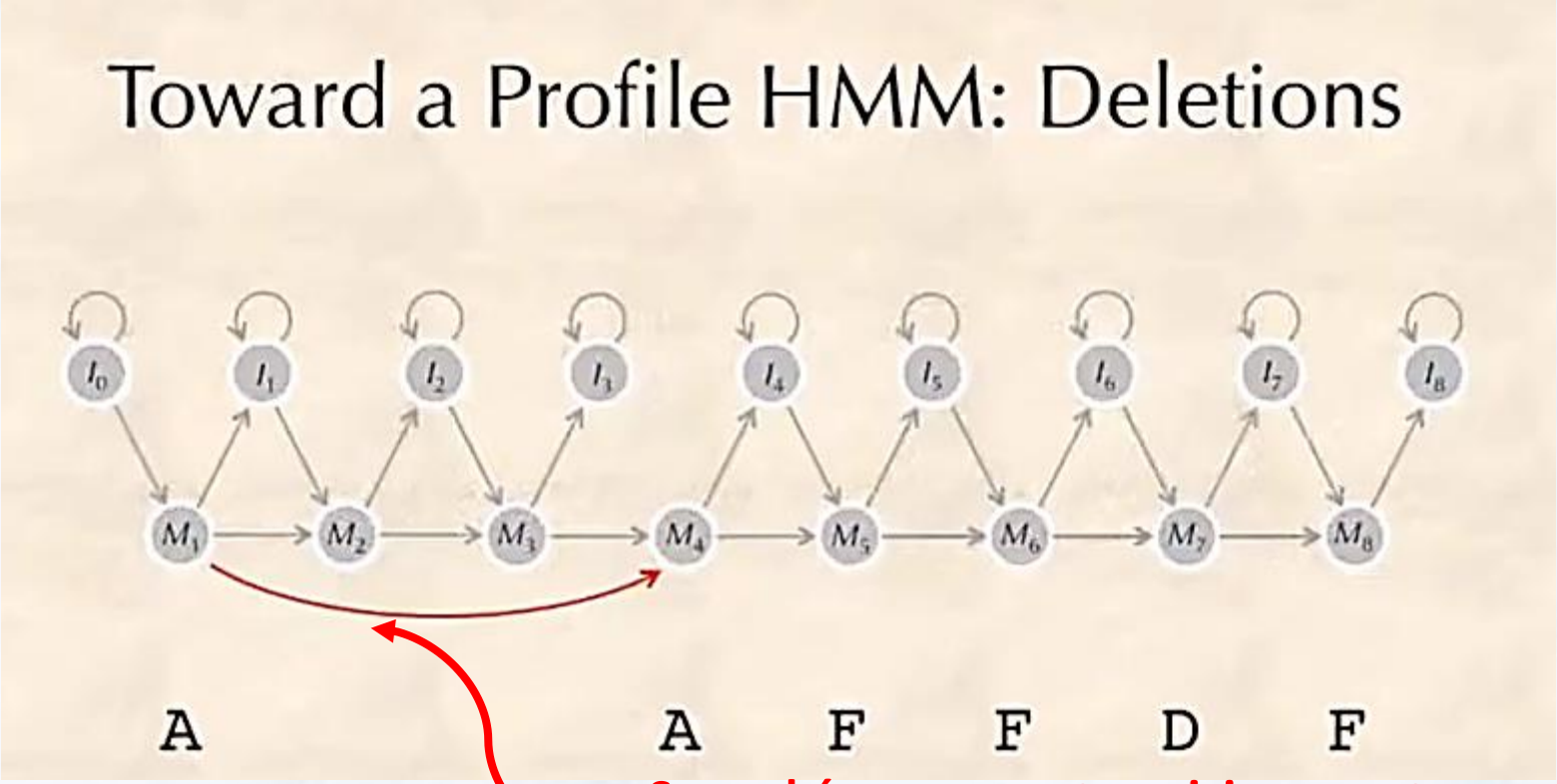
5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la inserciones?



5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la Deleciones?

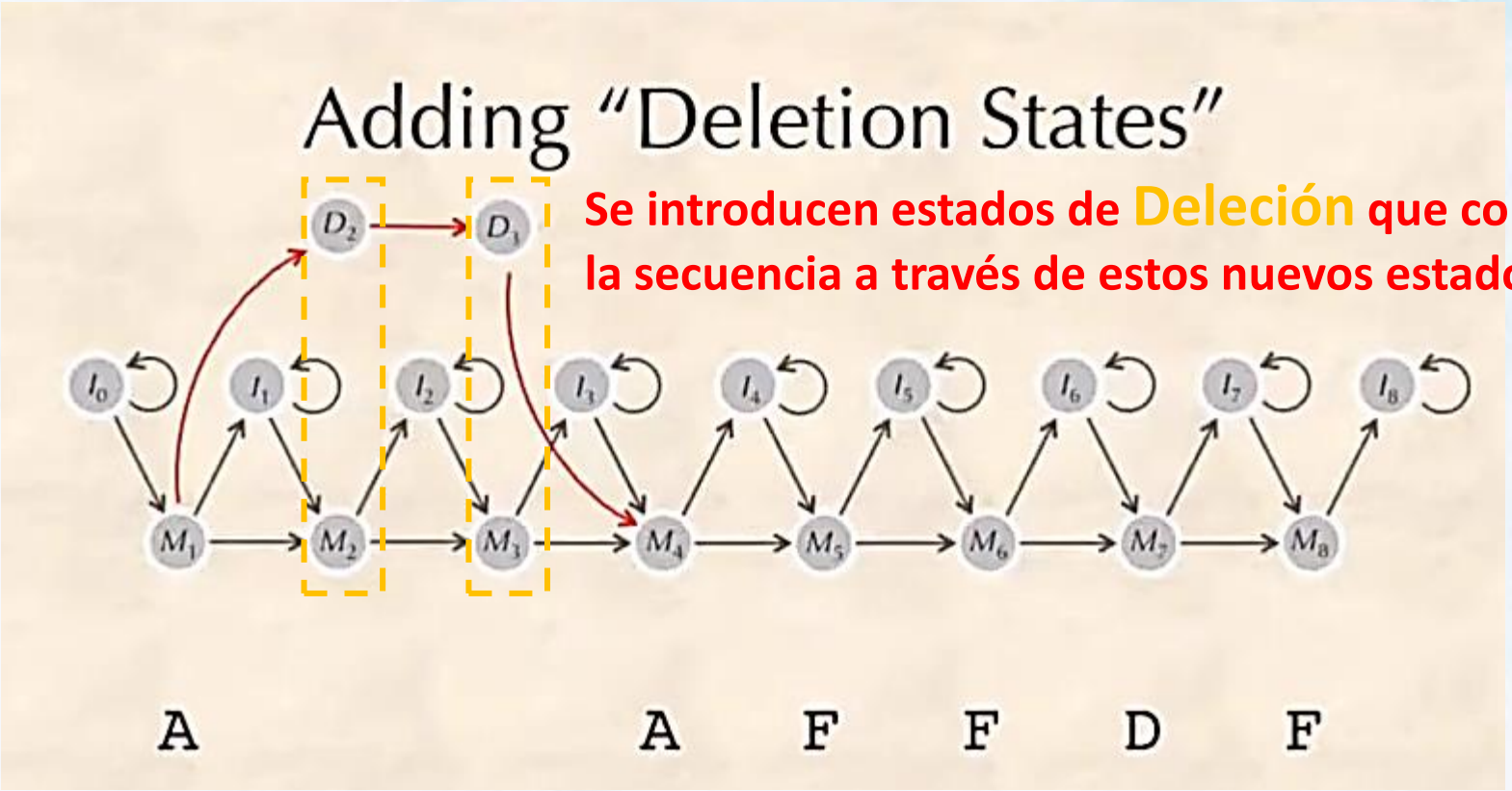


Se podrían agregar transiciones entre cualquiera de los estados siguientes al estado actual

Esto cambia las probabilidades de transición ya que M1 ahora tiene 3 posibilidades: pasar al estado I1, M2 o M4 (según el ejemplo)

5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la Deleciones?

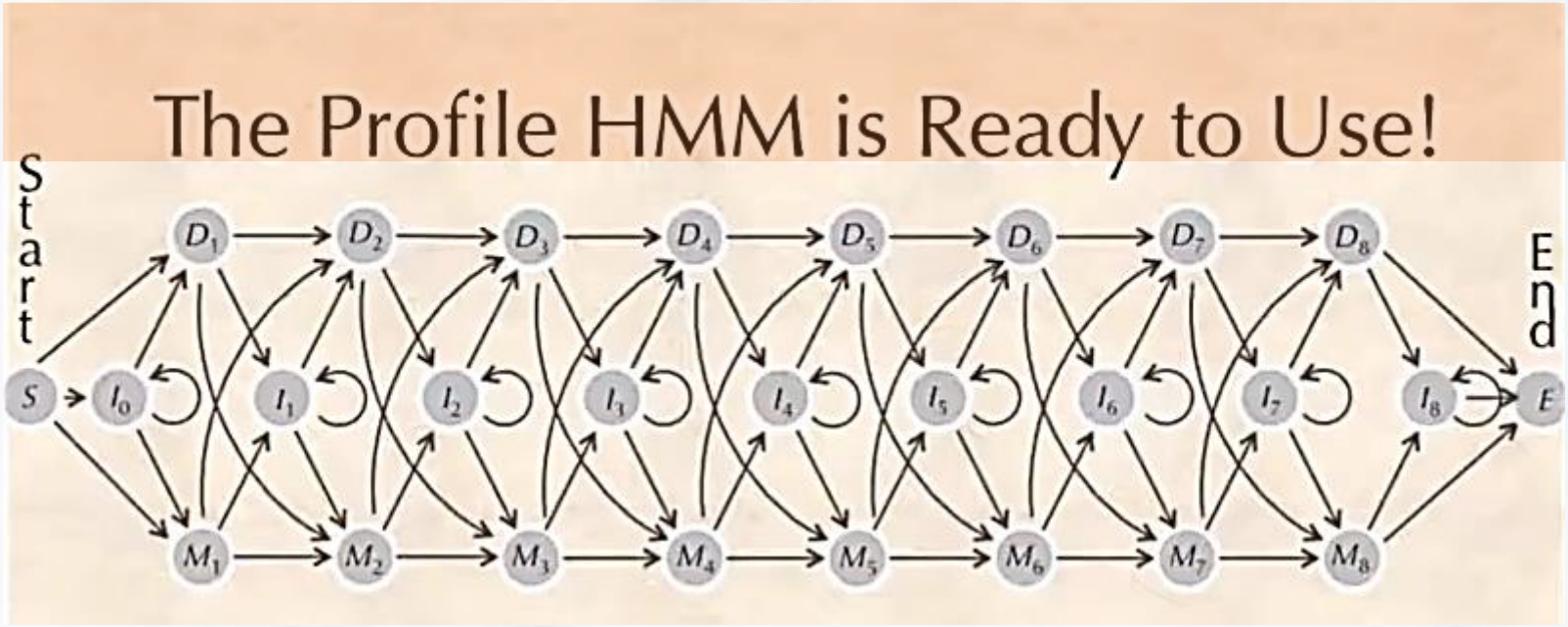


Esto cambia las probabilidades de transición ya que M1 ahora tiene 3 posibilidades: pasar al estado I1, M2 o M4 (según el ejemplo)

5. Modelos de Markov Ocultos (HMM)

¿Como modelamos la Deleciones?

Se introducen estados de Deleción que conducen la secuencia a través de estos nuevos estados



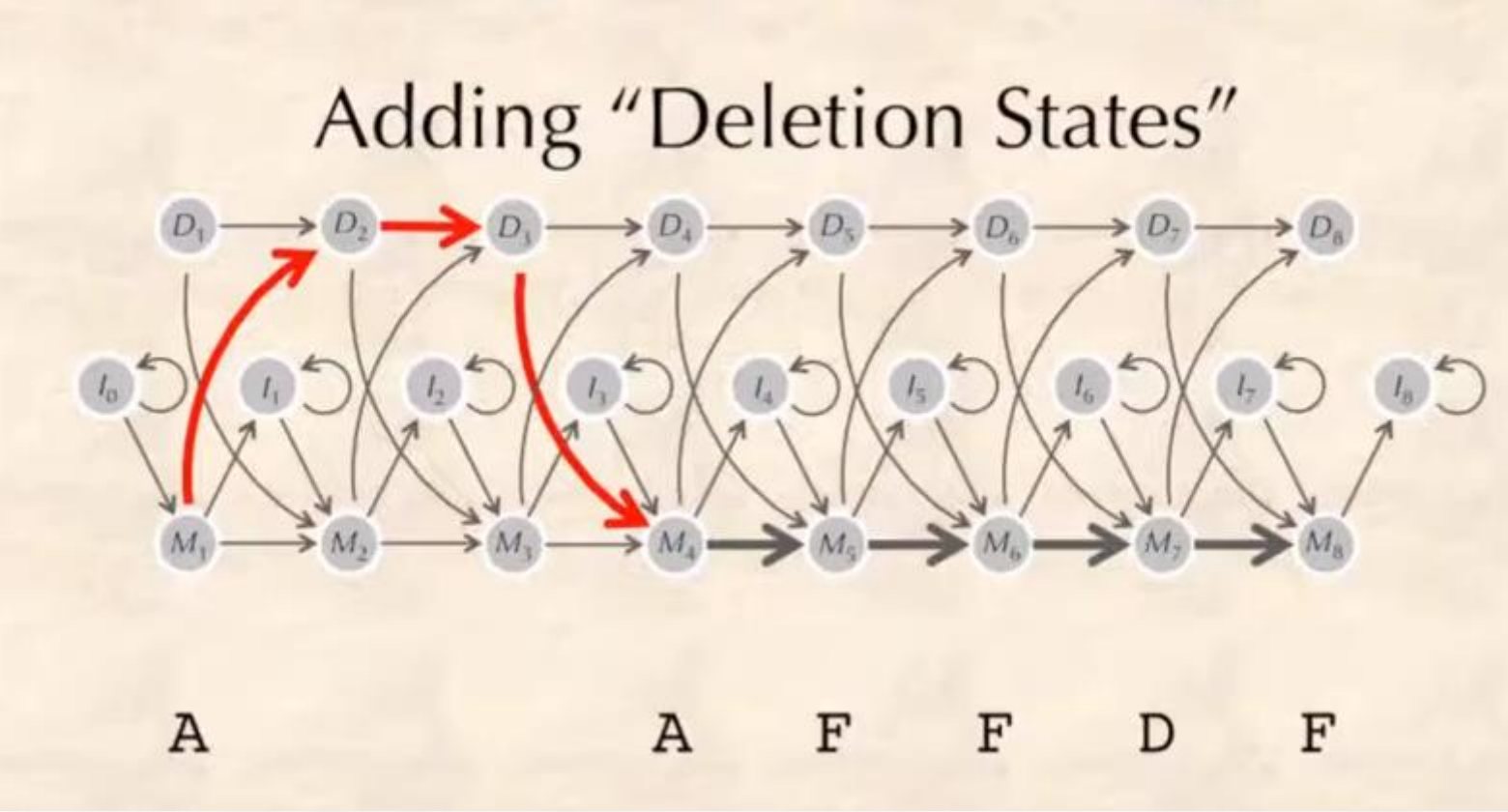
Se agregan estados de Deleciones en cada una de las posibles posiciones o estados

Esto cambia las probabilidades de transición ya que M1 ahora tiene 3 posibilidades: pasar al estado I1, M2 o M4 (según el ejemplo)

5. Modelos de Markov Ocultos (HMM)

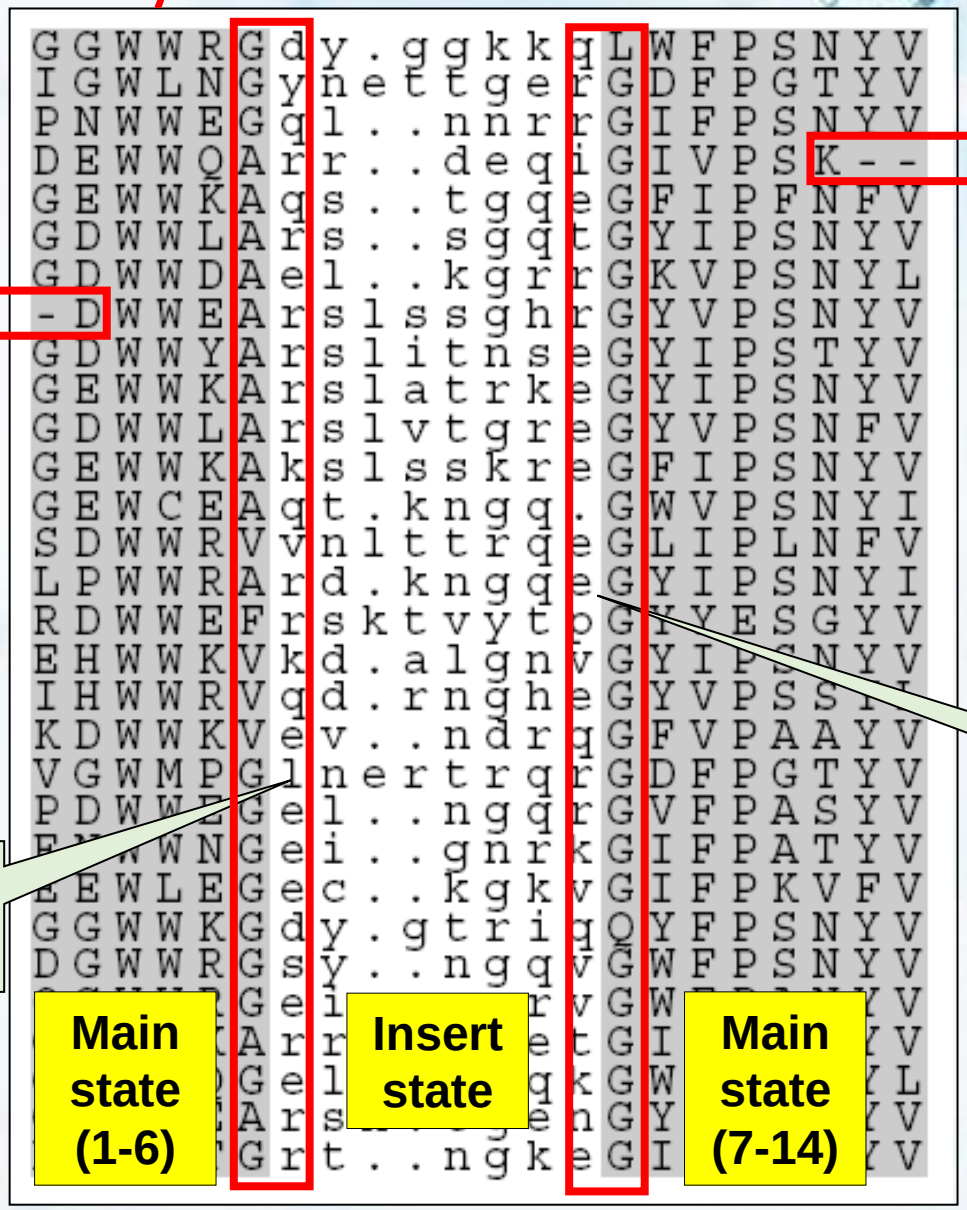
¿Como modelamos la inserciones?

Para una proteína con esta secuencia el modelo sigue el siguiente PATH



5. Modelos de Markov Ocultos (HMM)

¿Como veo y calculo las P en el alineamiento multiple?



Begin to delete transition

Main to delete transition

Main to insert transition

Insert to main transition

Main state (1-6)

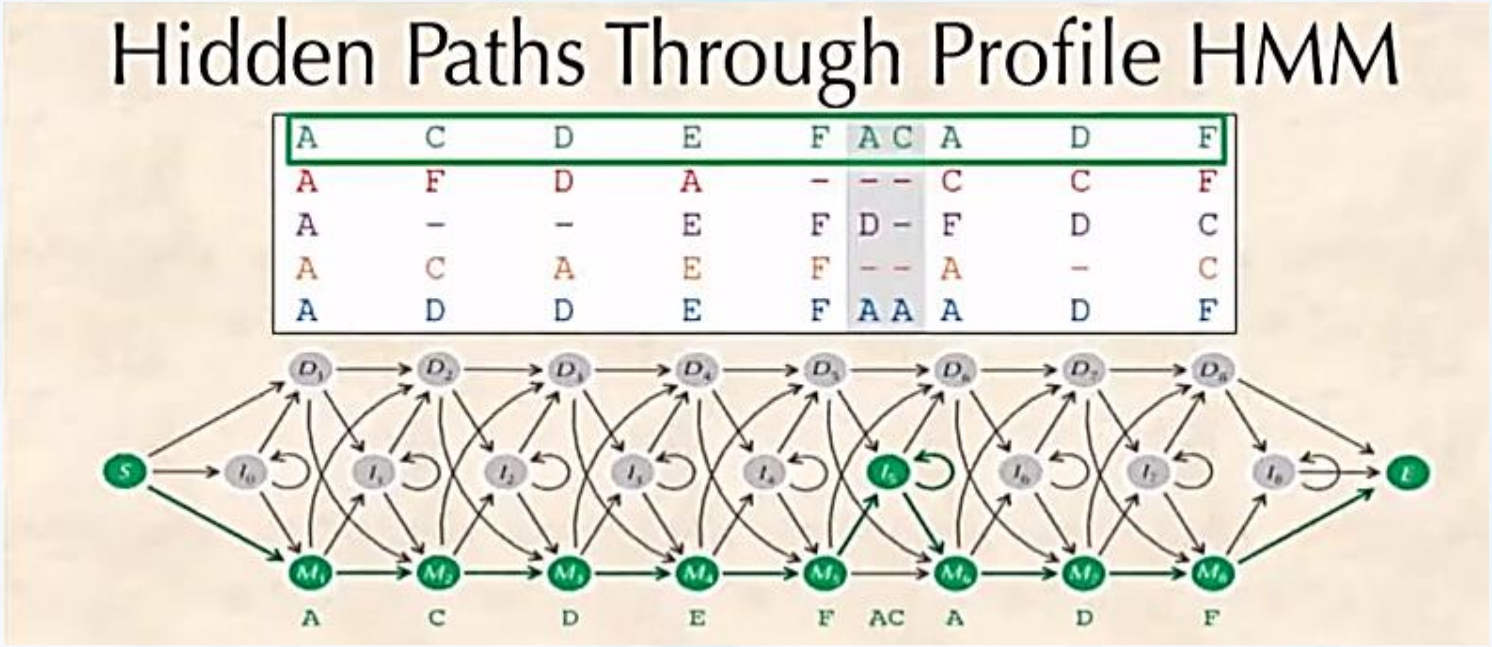
Insert state

Main state (7-14)

Cálculo de las probabilidades de emisión y de transición

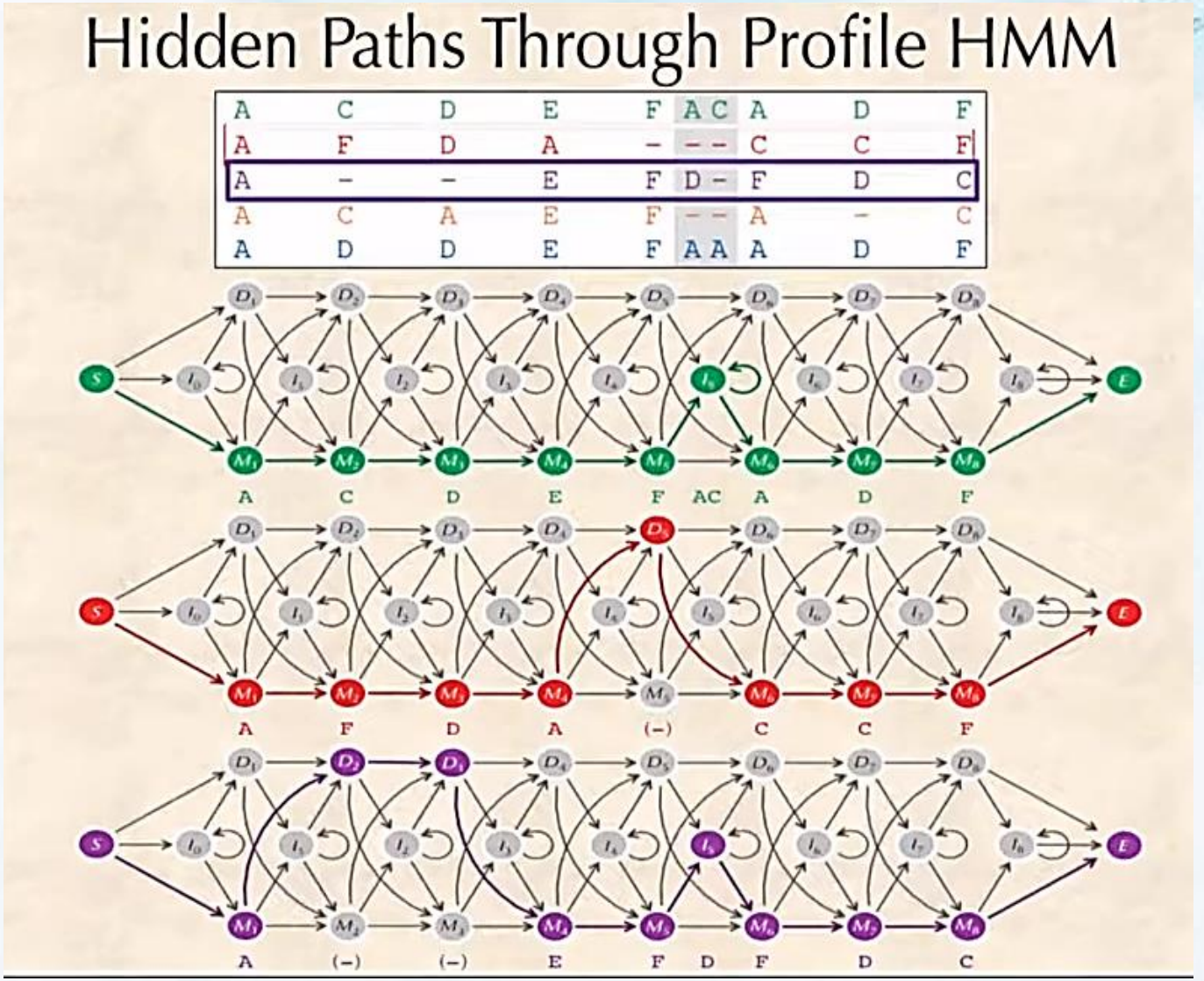
5. Modelos de Markov Ocultos (HMM)

Analicemos los Paths para el alineamiento del ejemplo anterior



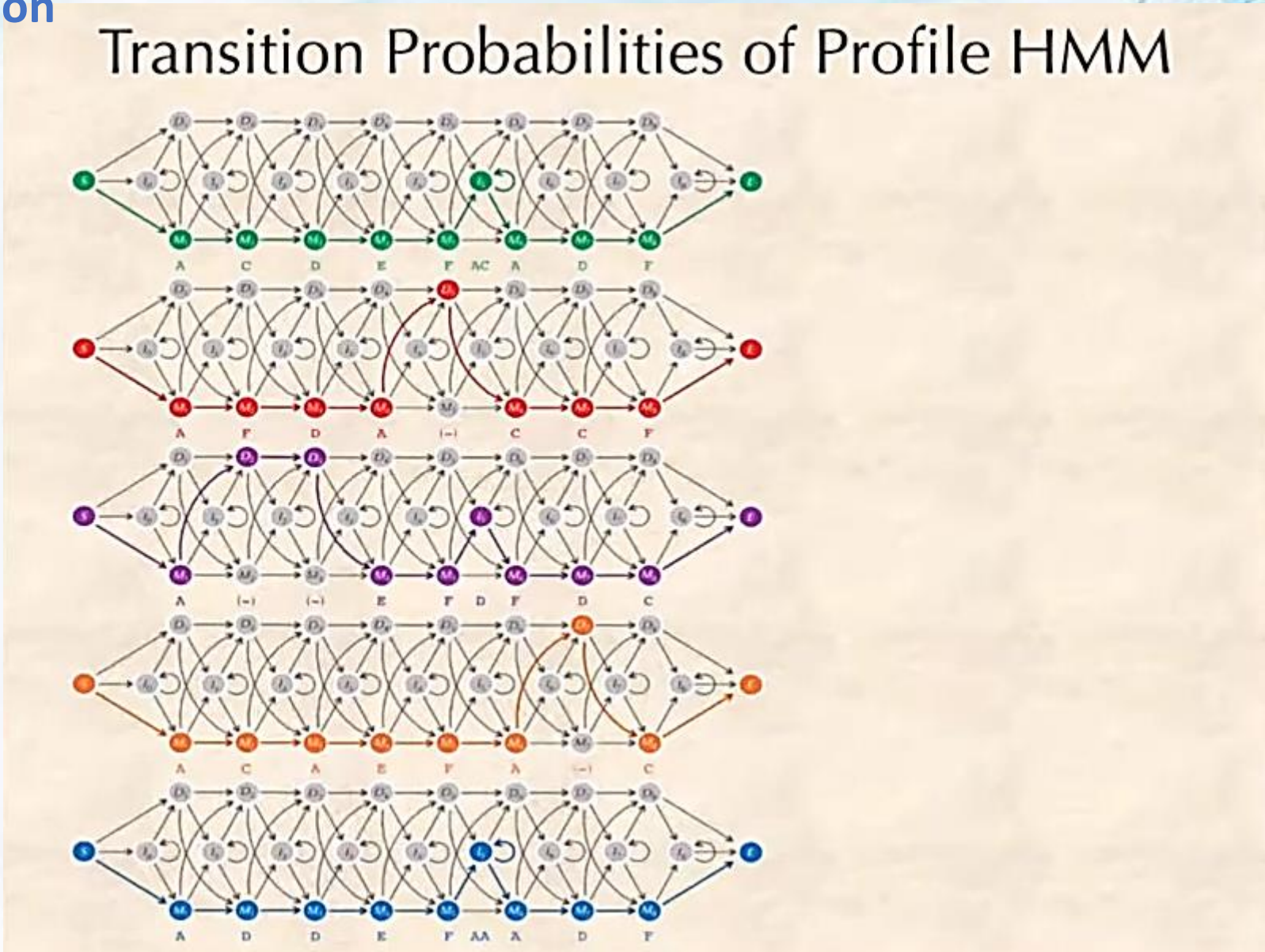
5. Modelos de Markov Ocultos (HMM)

Analicemos los Paths para el alineamiento del ejemplo anterior



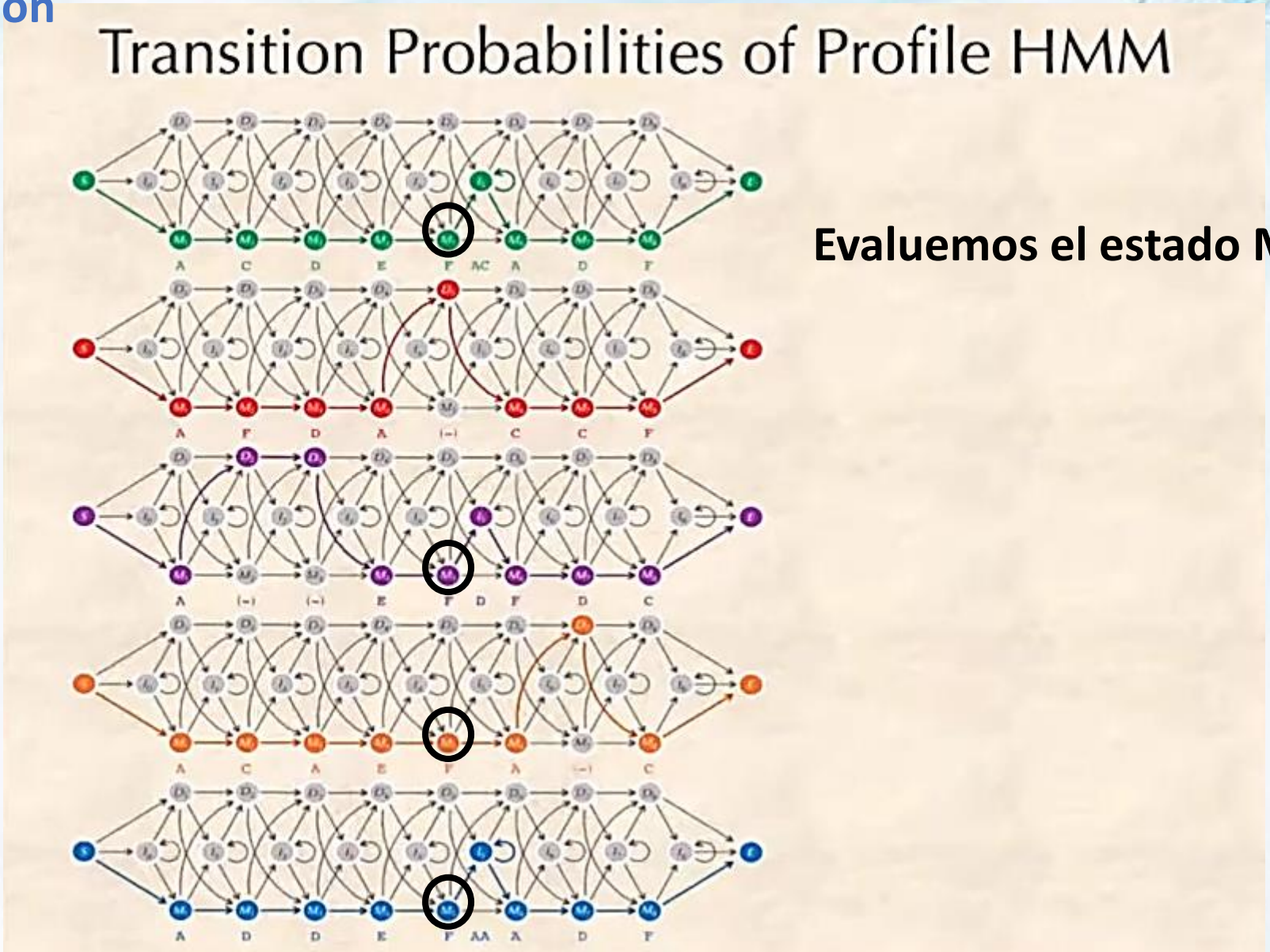
5. Modelos de Markov Ocultos (HMM)

Sabiendo los Paths Analicemos las probabilidades de transición y emisión



5. Modelos de Markov Ocultos (HMM)

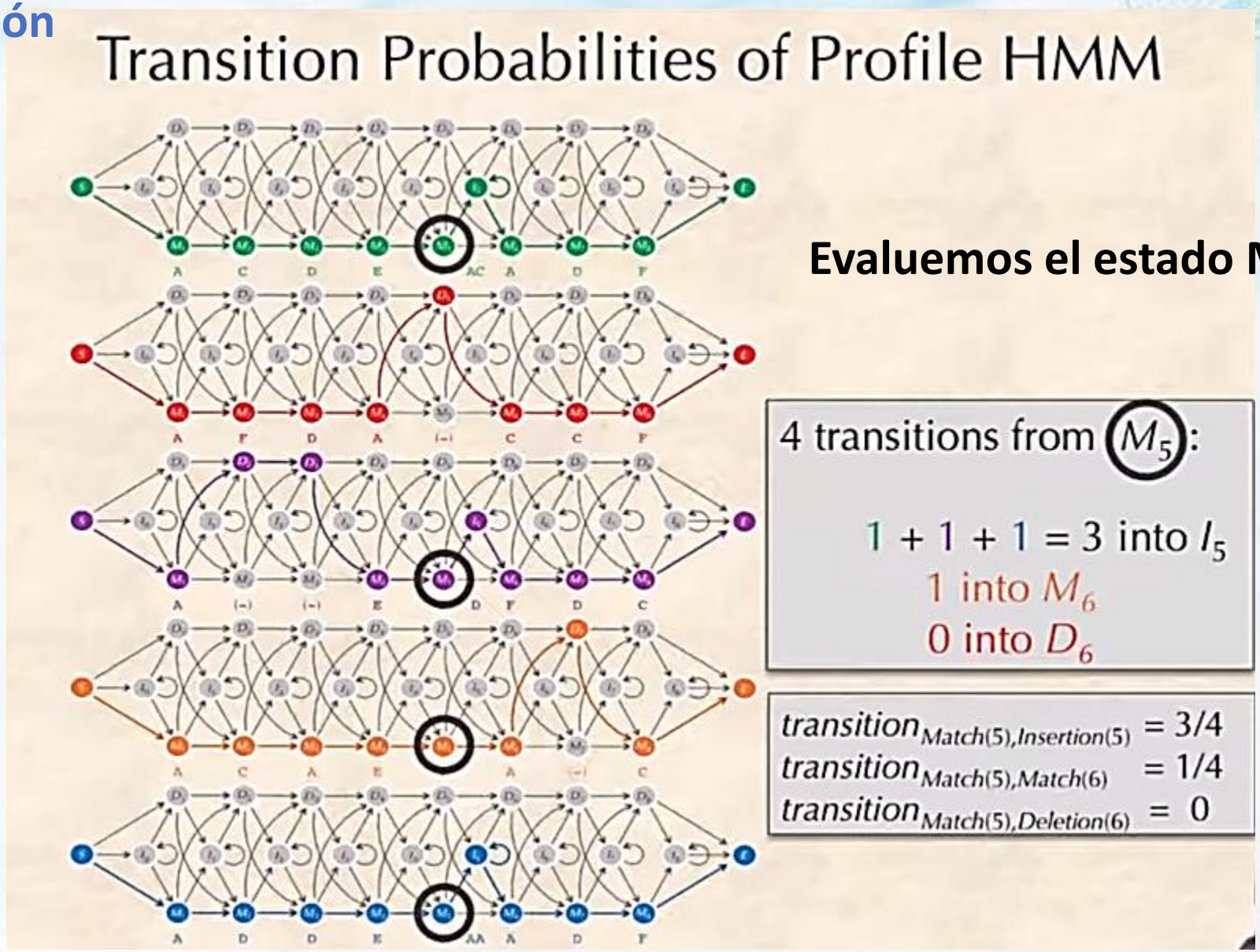
Sabiendo los Paths Analicemos las probabilidades de transición y emisión



Evaluemos el estado M5

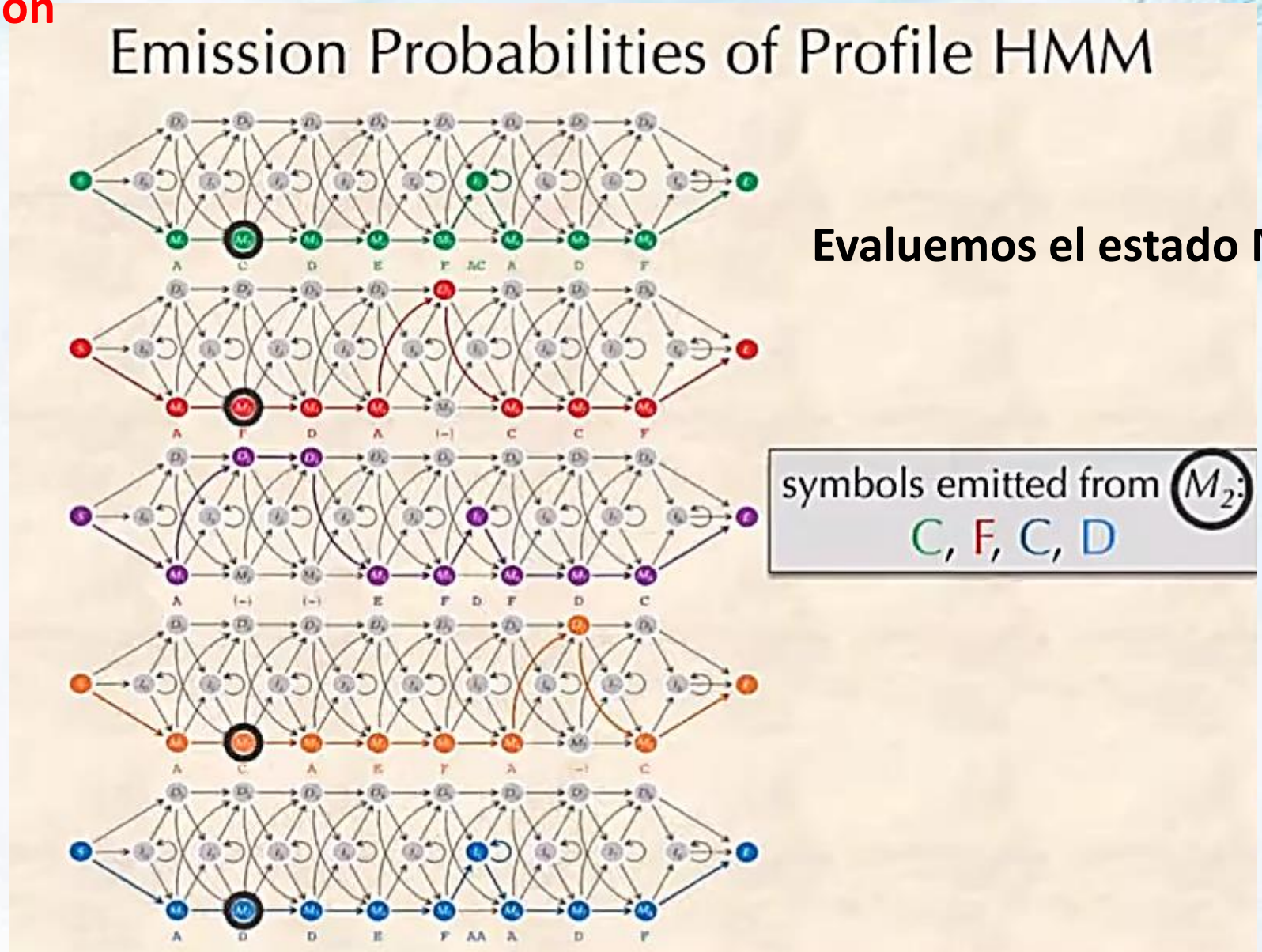
5. Modelos de Markov Ocultos (HMM)

Sabiendo los Paths Analicemos las probabilidades de transición y emisión



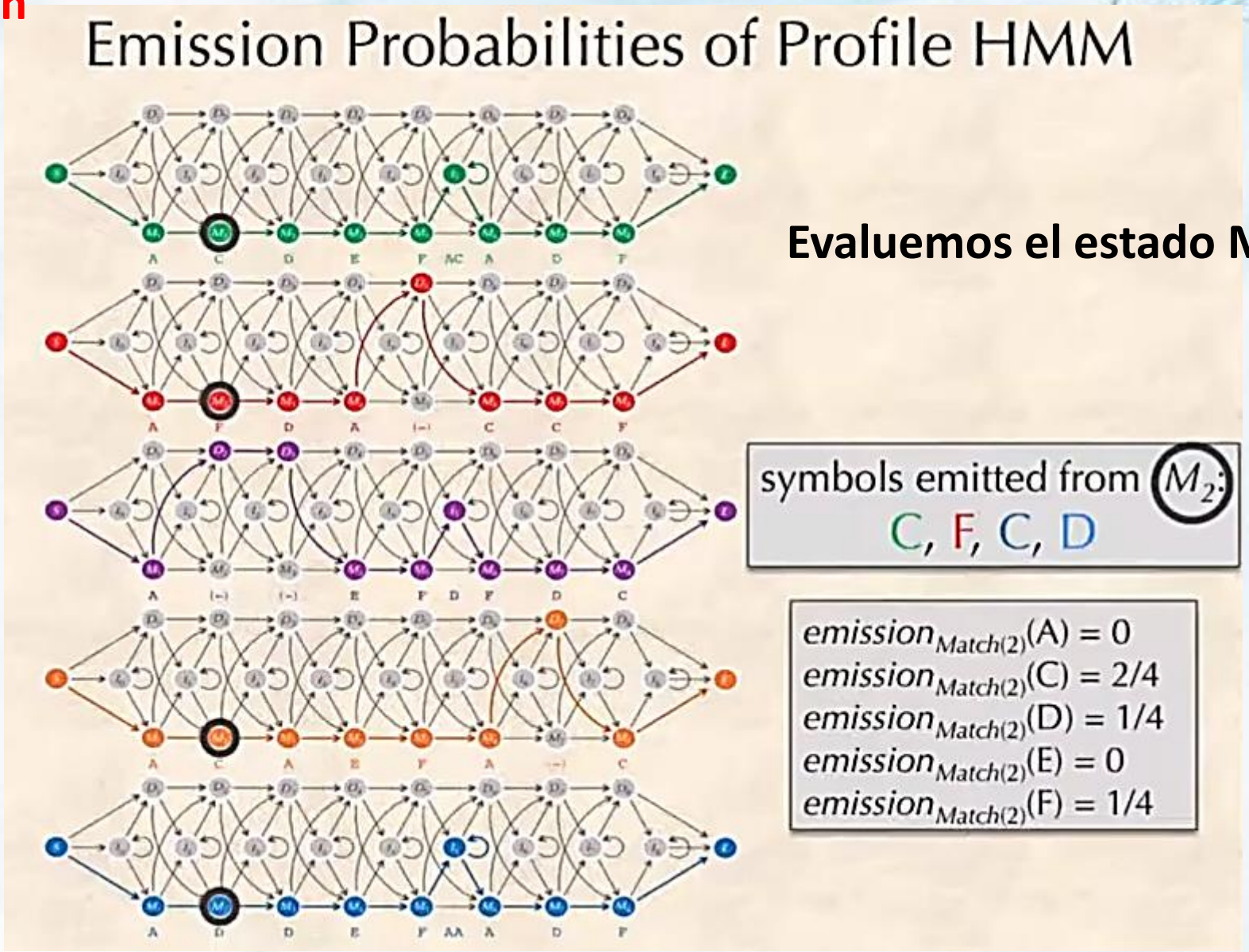
5. Modelos de Markov Ocultos (HMM)

Sabiendo los Paths Analicemos las probabilidades de transición y emisión



5. Modelos de Markov Ocultos (HMM)

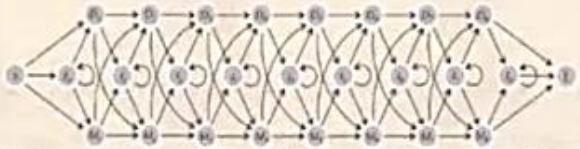
Sabiendo los Paths Analicemos las probabilidades de transición y emisión



5. Modelos de Markov Ocultos (HMM)

Sabiendo los Paths Analicemos las probabilidades de transición y emisión

Forbidden Transitions

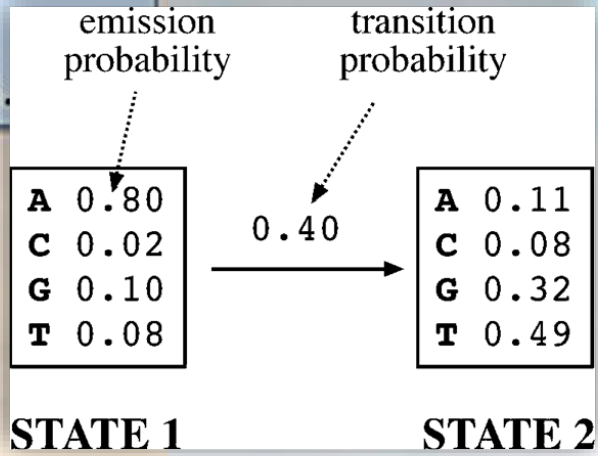


	S	I_0	M_1	D_1	I_1	M_2	D_2	I_2	M_3	D_3	I_3	M_4	D_4	I_4	M_5	D_5	I_5	M_6	D_6	I_6	M_7	D_7	I_7	M_8	D_8	I_8	E
S			1																								
I_0																											
M_1						.8	.2																				
D_1																											
I_1																											
M_2									1																		
D_2																											
I_2																											
M_3												1															
D_3																											
I_3												1															
M_4															.8	.2											
D_4																											
I_4																											
M_5																	.25	.75									
D_5																	.33	.67									
I_5																		1									
M_6																			.8	.2							
D_6																											
I_6																											
M_7																							1				
D_7																											
I_7																							1				
M_8																											1
D_8																											
I_8																											
E																											

Se construye una matriz de transiciones y emisiones conteniendo todas las probabilidades

Gray cells: edges in the HMM diagram.

Clear cells: *forbidden* transitions.



5. Modelos de Markov Ocultos (HMM)

¿Como funciona con una secuencia Query?

1. Para encontrar una ruta óptima dentro de un HMM que coincida con una secuencia Query con la probabilidad más alta, se usa la matriz de valores de probabilidad para cada estado en cada posición de residuo.

2. Se determina la ruta más probable para esta matriz.

1. Construye una matriz con los valores máximos de probabilidad de emisión de todos los símbolos en un estado multiplicado por la probabilidad de transición para ese estado

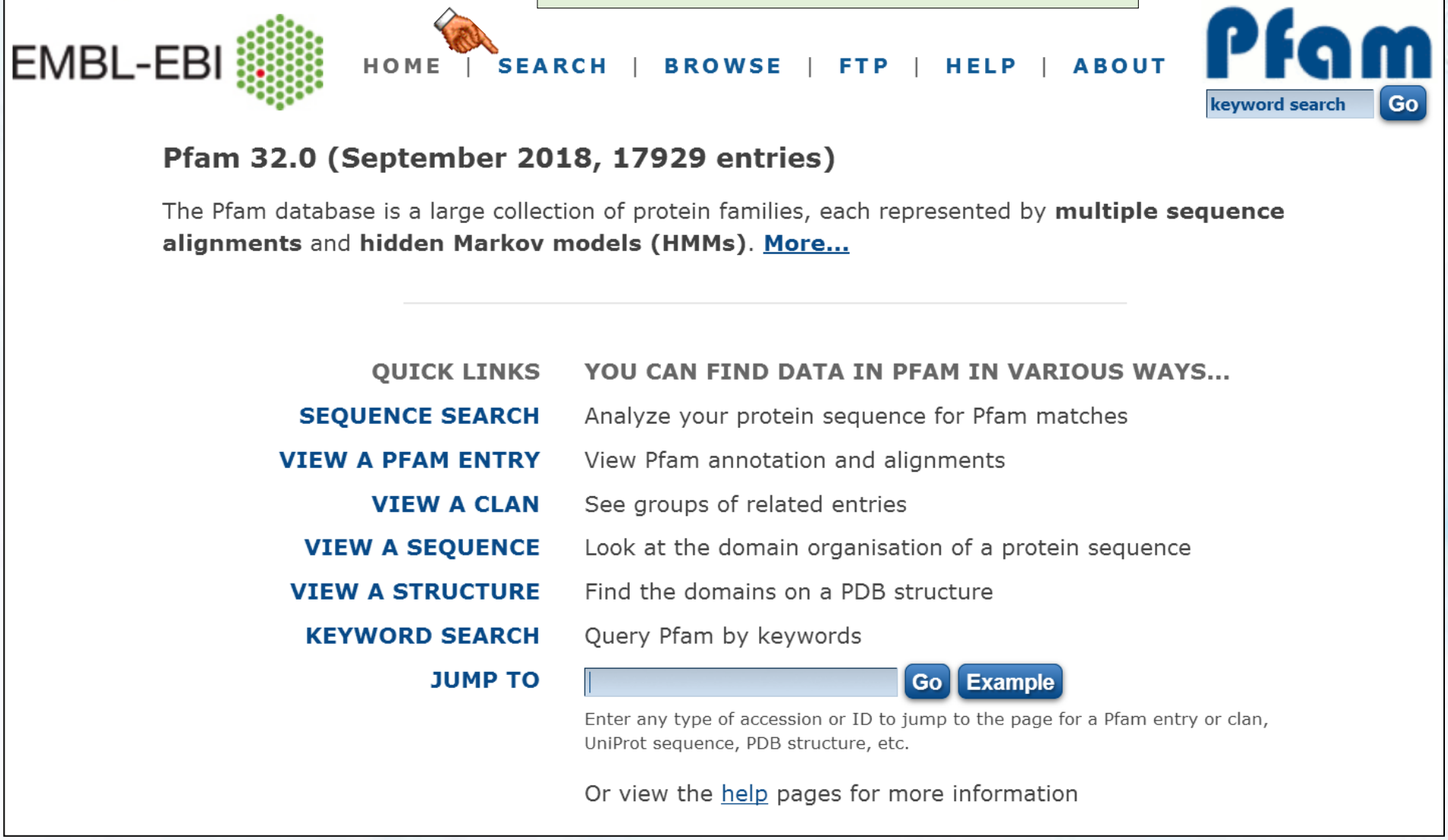
2. Luego utiliza un procedimiento de rastreo que va desde la esquina inferior derecha a la esquina superior izquierda para encontrar la ruta con los valores más altos en la matriz.

	M_0	I_1	D_1	M_1	I_2	D_2	M_2	I_3	D_3	M_3
S_0	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S_1	0.00	0.00	0.00	0.19	0.03	0.00	0.00	0.00	0.00	0.00
S_2	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.01	0.00	0.00
S_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42

5. Modelos de Markov Ocultos (HMM)

5. Modelos de Markov Ocultos (HMM)

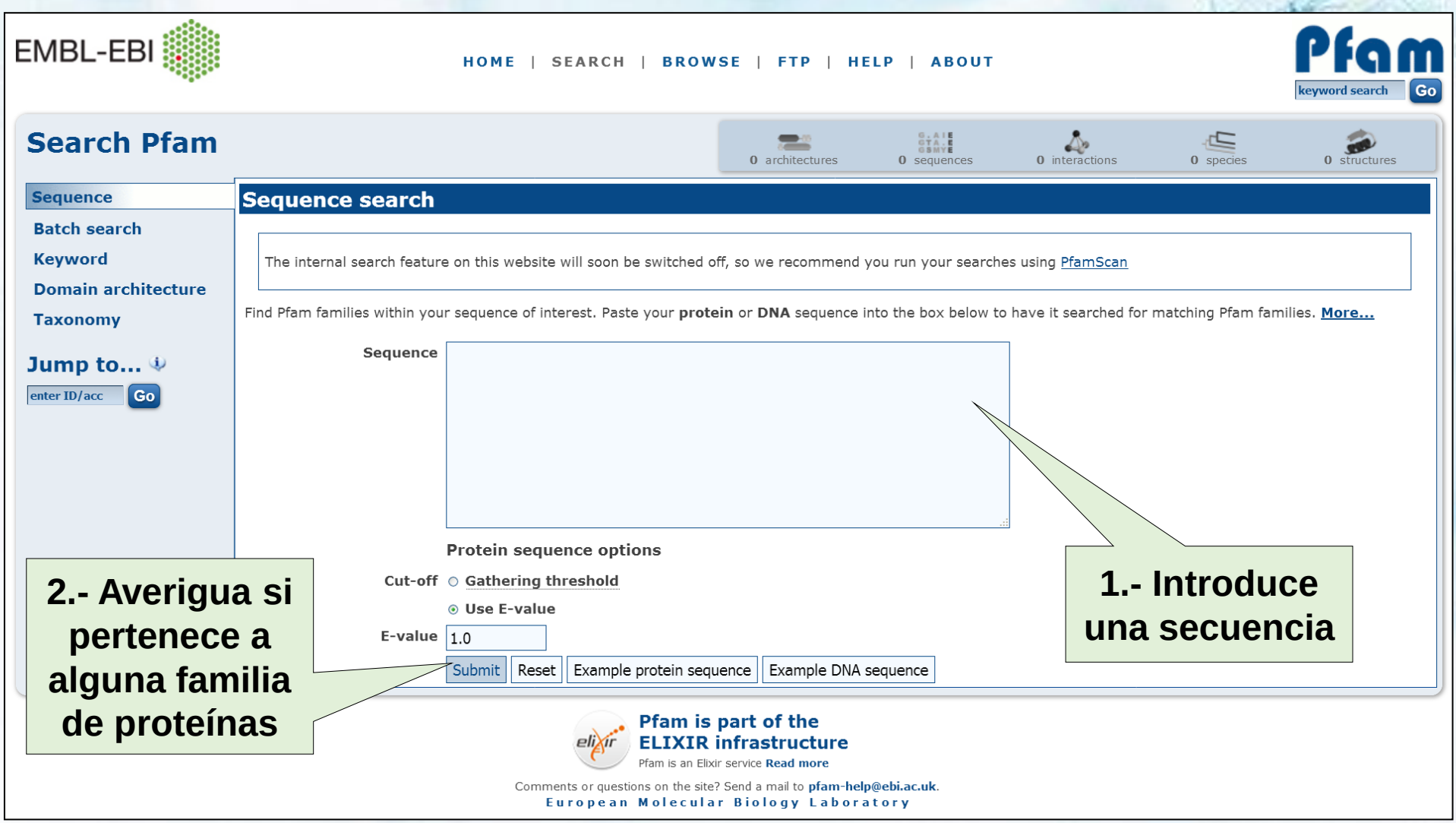
http://pfam.xfam.org/



La BD Pfam

5. Modelos de Markov Ocultos (HMM)

http://pfam.xfam.org/search#tabview=tab0



2.- Averigua si pertenece a alguna familia de proteínas

1.- Introduce una secuencia

Una búsqueda en la BD Pfam

5. Modelos de Markov Ocultos (HMM)

http://pfam.xfam.org/help

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Pfam Help

0 architectures 0 sequences 0 interactions 0 species 0 structures

- Summary
- Changes
- Getting Started
- Training
- FAQ
- Glossary
- Scores
- Citing Pfam
- Pfam and Wikipedia
- Linking to Pfam
- Guide to Graphics
- RESTful interface
- Pfam database
- FTP site
- Privacy
- Team Members
- Contact Us

Jump to...

Help Summary

Pfam 31.0 (Mar 2017 , 16712 families)

Proteins are generally comprised of one or more functional regions, commonly termed domains. The presence of different domains in varying combinations in different proteins gives rise to the diverse repertoire of proteins found in nature. Identifying the domains present in a protein can provide insights into the function of that protein.

The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and a hidden Markov model (HMMs).

Each Pfam family, often referred to as a Pfam-A entry, consists of a curated seed alignment containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases.

Pfam entries are classified in one of six ways:

Family:
A collection of related protein regions

Domain:
A structural unit

Repeat:
A short unit which is unstable in isolation but forms a stable structure when multiple copies are present

Motifs:
A short unit found outside globular domains

Coiled-Coil:
Regions that predominantly contain coiled-coil motifs, regions that typically contain alpha-helices that are coiled together in bundles of 2-7.

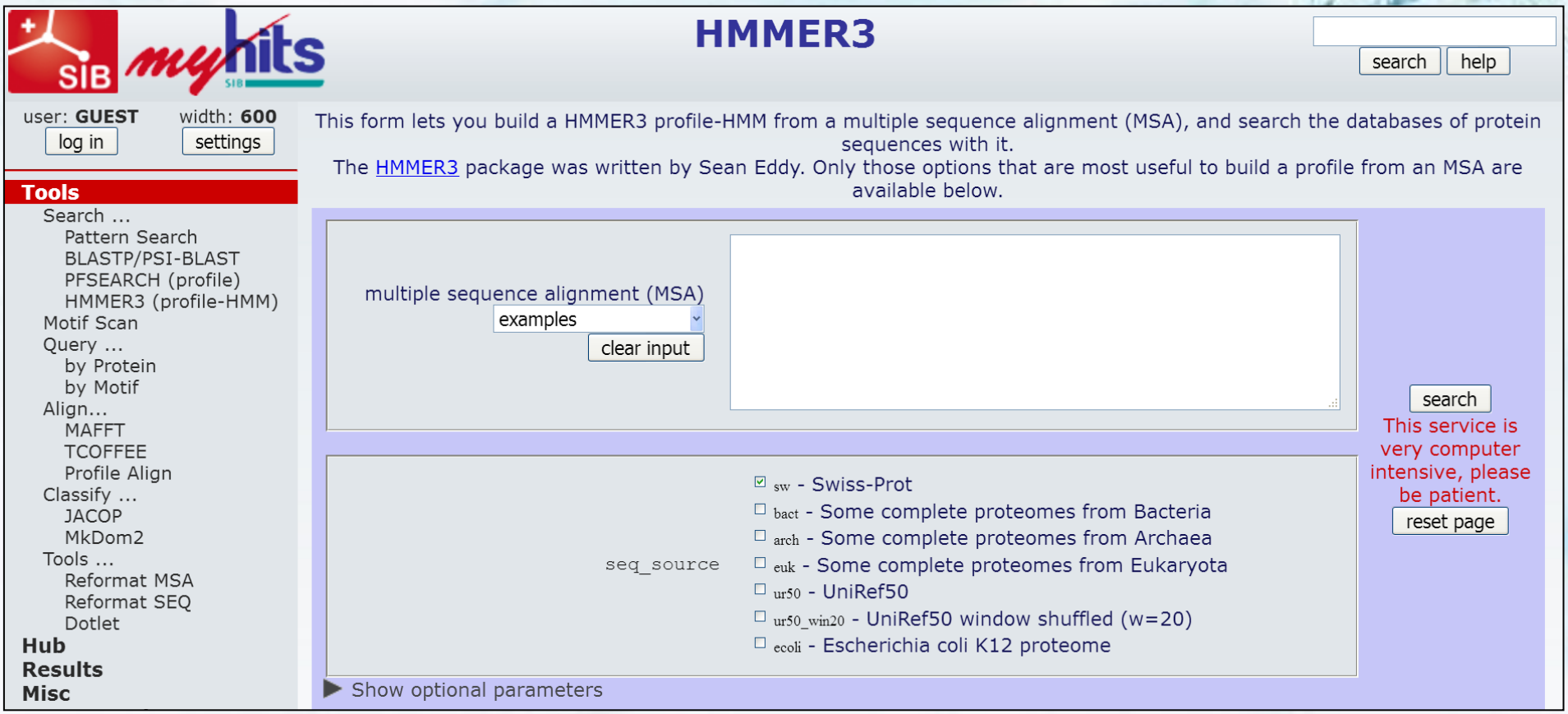
Disordered:
Regions that are conserved, yet are either shown or predicted to contain bias sequence composition and/or are intrinsically disordered (non-globular).

Related Pfam entries are grouped together into *clans*; the relationship may be defined by similarity of sequence, structure or profile-HMM.

Los registros de Pfam se clasifican en 6 grupos

5. Modelos de Markov Ocultos (HMM)

http://myhits.isb-sib.ch/cgi-bin/hmmer3_search



HMMER3 lets you build a profile-HMM from a multiple sequence alignment (MSA), and search the databases of protein sequences with it.

La herramienta HMMER3 (SIB)

5. Modelos de Markov Ocultos (HMM)

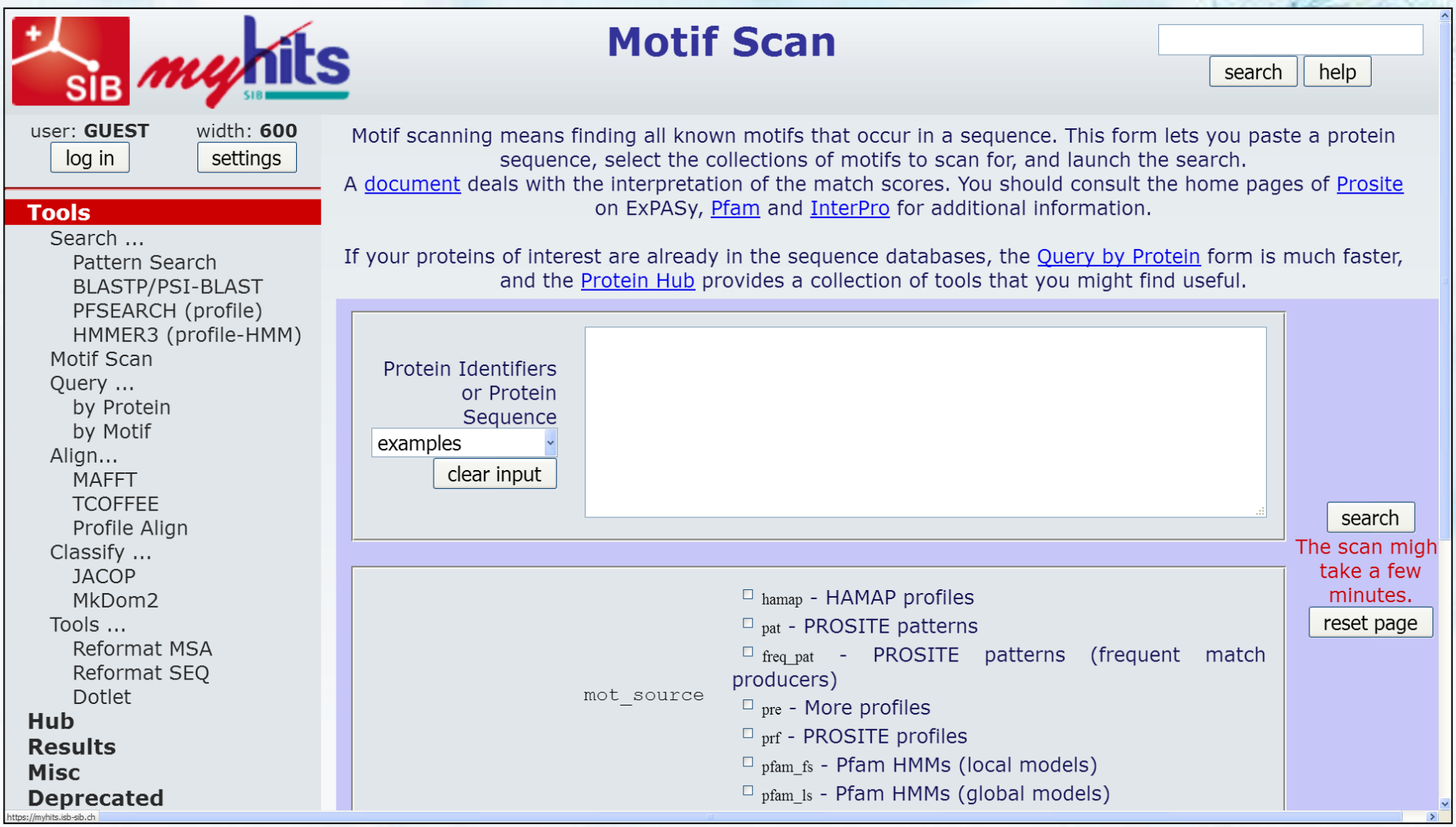
http://www.genome.jp/tools/motif/

1. Search with a protein query sequence against Motif Libraries
2. Align a protein sequnce with a profile library given by a user. (PROSITE or HMMER format)
3. Search with a profile against protein sequence databases
4. Search a protein sequence pattern (regular expression) against sequence databases
5. Generate a profile from a set of multiple aligned sequences

La herramienta MOTIF Search

5. Modelos de Markov Ocultos (HMM)

http://myhits.isb-sib.ch/cgi-bin/motif_scan



La herramienta Motif Scan (SIB)

interpro

¿Tengo que buscar en todas las BD de proteínas?

Why do we need InterPro?

Protein signature databases have become vital tools for classification of protein sequences in order to infer their function. Many protein signature-based resources are now available, each with their own strengths and weaknesses.

We need InterPro because it:

- Reduces redundancy and simplifies protein sequence analysis by integrating signatures from different member databases that represent the same protein family, domain or site.
- Unites the member databases, capitalising on their individual strengths to produce a powerful classification tool.
- Provides a single convenient searchable location, allowing simultaneous querying of all member databases.
- Adds information (including descriptive abstracts and Gene Ontology terms) to the signatures, which may be used to annotate the proteins they match.

¿Por qué resulta tan útil InterPro?

The InterPro Consortium

The following databases make up the InterPro Consortium:

CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

CDD is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domain models, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases.

MobiDB offers a centralized resource for annotations of intrinsic protein disorder. The database features three levels of annotation: manually curated, indirect and predicted. The different sources present a clear tradeoff between quality and coverage. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest.

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.

El consorcio InterPro (1)

Pfam	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at EMBL-EBI, Hinxton, UK.
PIRSF	PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.
PRINTS	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.
ProDom	ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.
prosite	PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.
SFLD	SFLD (Structure-Function Linkage Database) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities.
SMART	SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.
Superfamily	SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.
	TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

El consorcio InterPro (2)

What are entry types and why are they important?

Each InterPro entry is assigned one of a number of types which tell you what you can infer when a protein matches the entry. The entry types are:

Homologous Superfamily

A homologous superfamily is a group of proteins that share a common evolutionary origin, reflected by similarity in their structure. Since superfamily members often display very low similarity at the sequence level, this type of InterPro entry is usually based on a collection of underlying hidden Markov models, rather than a single signature.

Family

A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.

Domain

Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain.

Repeat

A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.

Site

A match to an InterPro entry of this type indicates a short sequence that contains one or more conserved residues. The type of sites covered by InterPro are active sites, binding sites, post-translational modification sites and conserved sites.

Tipos de registro en InterPro

InterPro is used to provide annotation for UniProtKB

- The sequences stored in the universal protein database UniProtKB are analysed regularly using InterPro. In this way, InterPro helps to provide annotation for uncharacterised sequences in the UniProtKB database (Figure 3).

InterPro can be used to analyse any protein sequence

- Users can also choose to analyse their own sequences for predictions about their function and/or the presence of certain domains and sequence features.

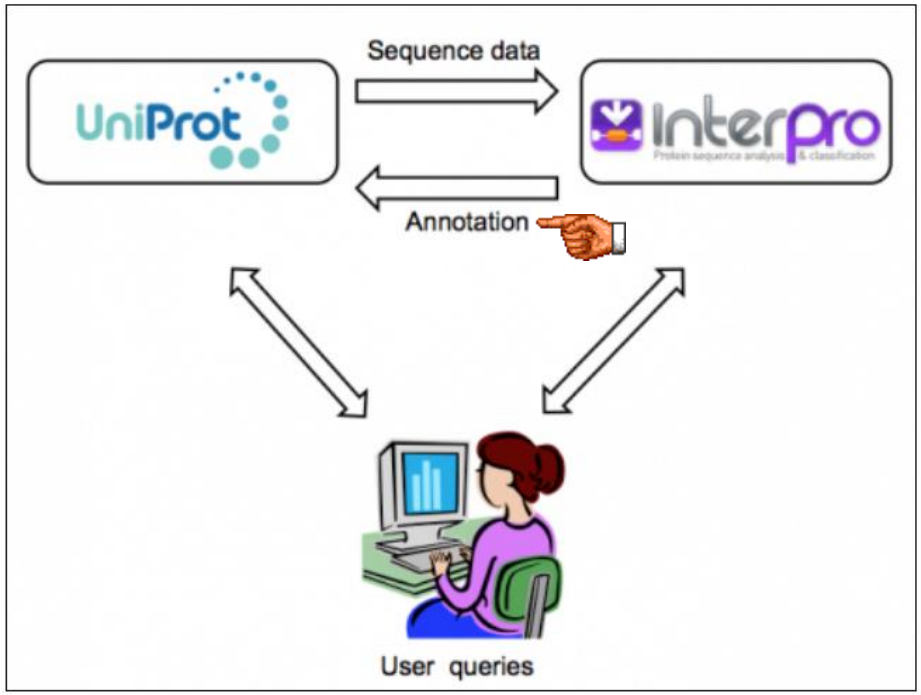
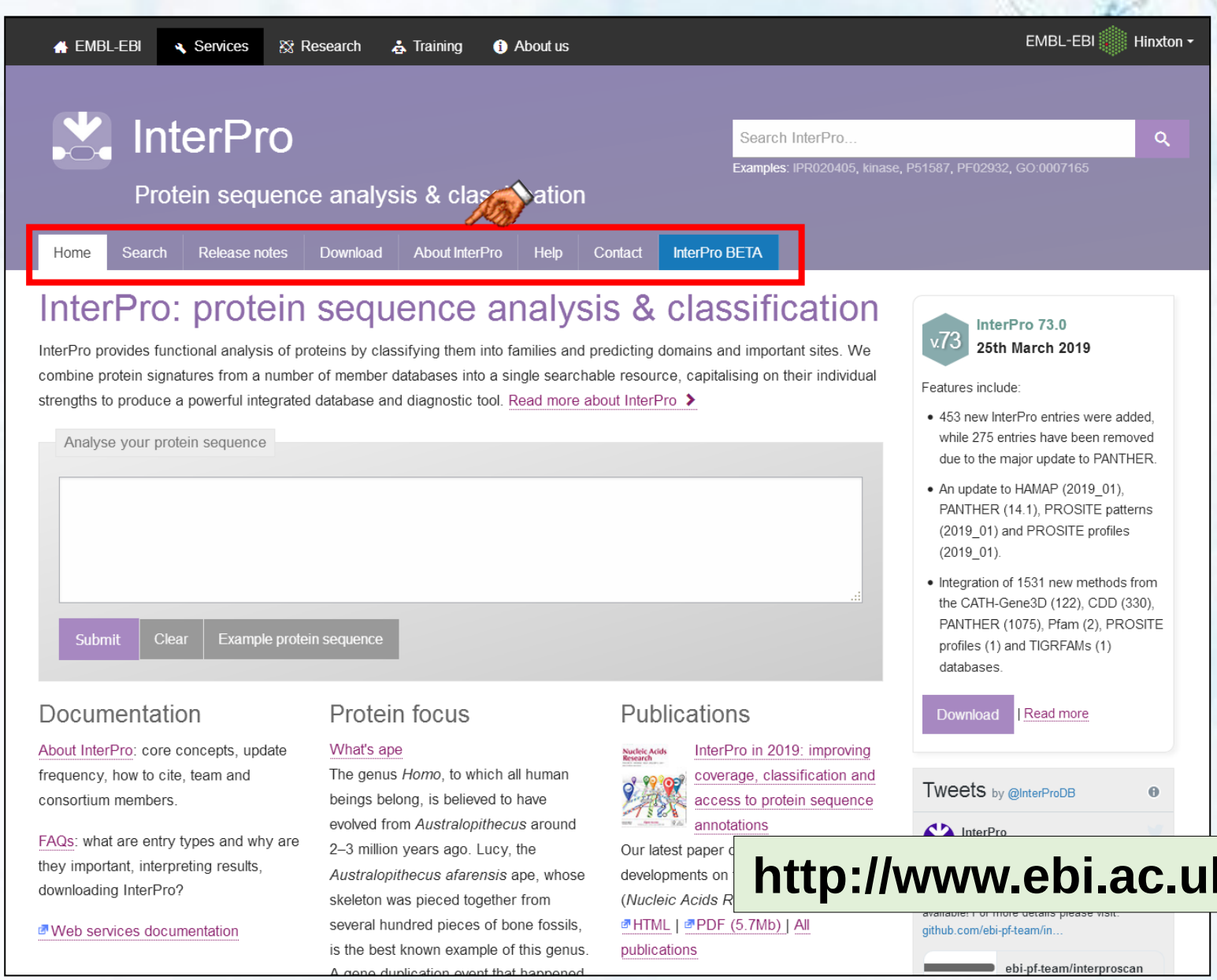


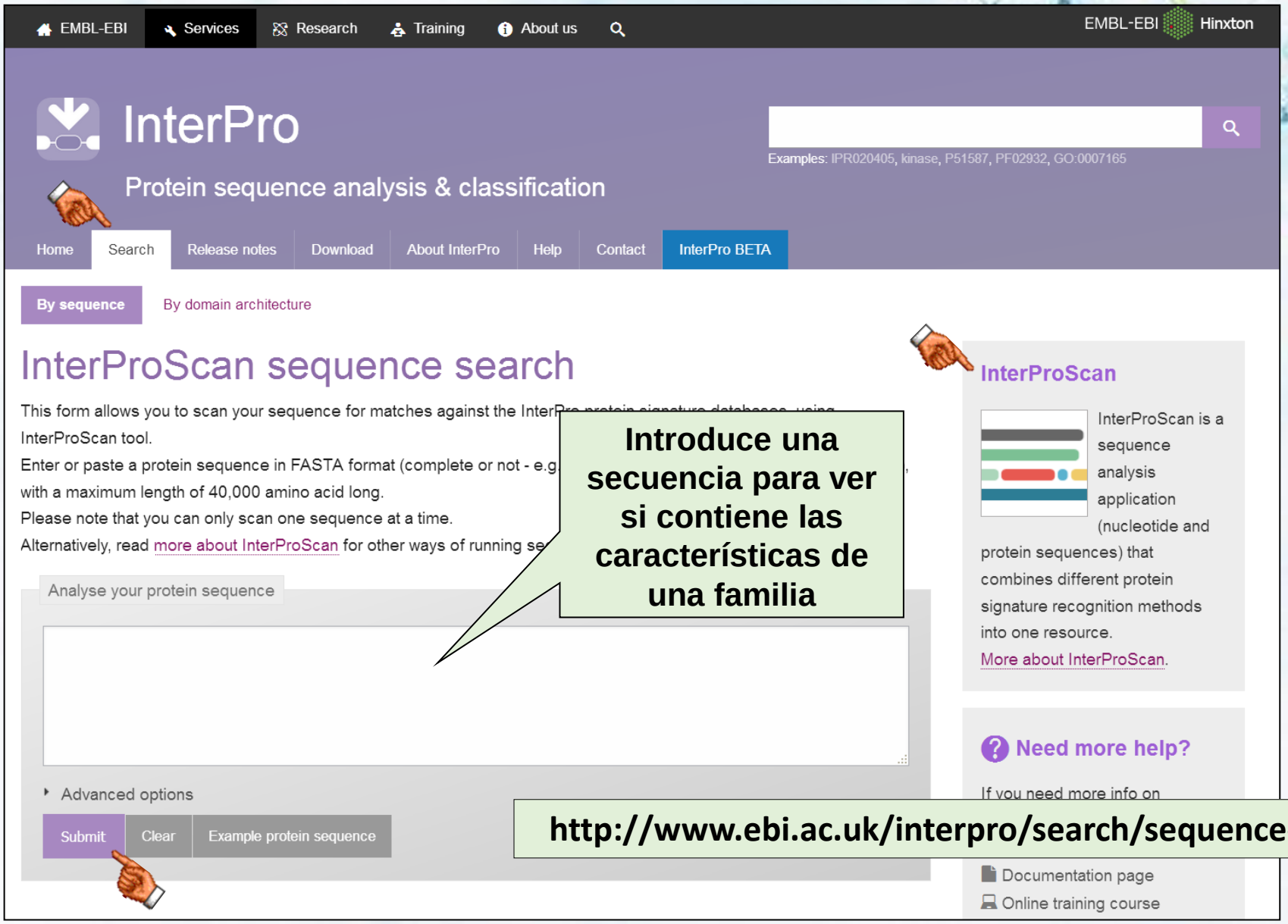
Figure 3. Flow of data between InterPro and UniProt databases.

InterPro permite introducir anotaciones en UniProtKB



http://www.ebi.ac.uk/interpro/

Página principal de InterPro



La herramienta InterProScan

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI Hinxton

InterPro

Protein sequence analysis & classification

Search InterPro...

Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home | Search | Release notes | Download | About InterPro | Help | Contact | InterPro BETA

InterProScan

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases, that make up the InterPro consortium.

Download the latest version

InterProScan 5.34-73.0

Linux (64-bit) - System requirements

There are no versions planned for Windows or Apple operating systems. This is due to constraints in the various third-party binaries that InterProScan runs.

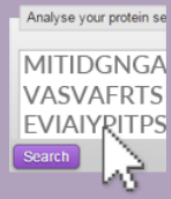
Download InterProScan

Version 5.34-73.0 - 64 bit Linux - 9.0GB

MD5:60792491c9eb5835f825c45572e530c5

Older versions of InterProScan are not supported anymore. We highly recommend you to update to the latest version.
For more information on downloading, installing and running InterProScan please see the InterProScan wiki.

Sequence search



Click here to scan your protein sequence and discover the domains it contains and the family to which it belongs.

Related Publications

InterProScan 5 user's guide

Full documentation to help on installation, configuration, migration.

HTML

InterProScan 5: genome-scale protein function classification

El programa InterProScan