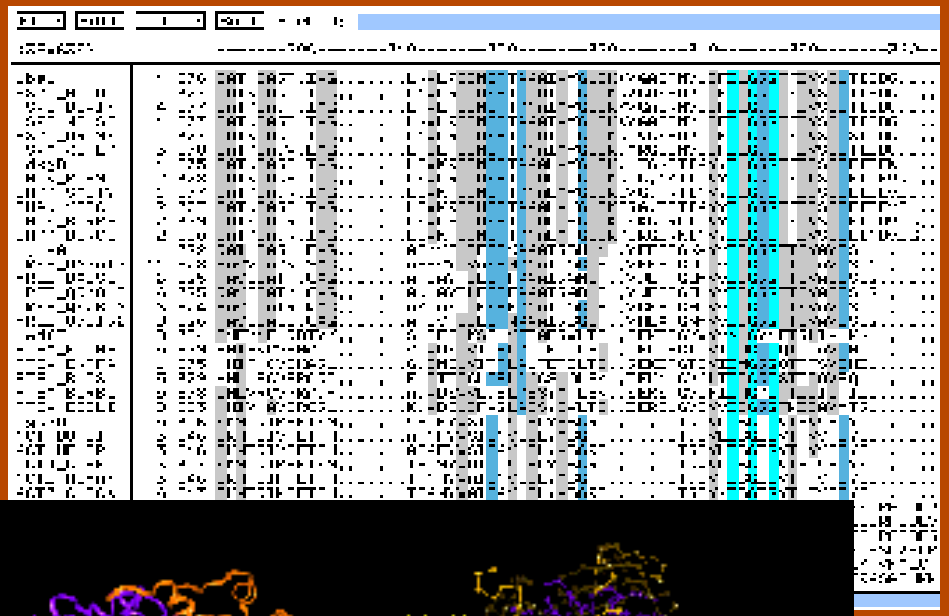
Motifs, Profiles and Domains

Michael Tress
Protein Design Group
Centro Nacional de Biotecnología, CSIC

Comparing Two Proteins

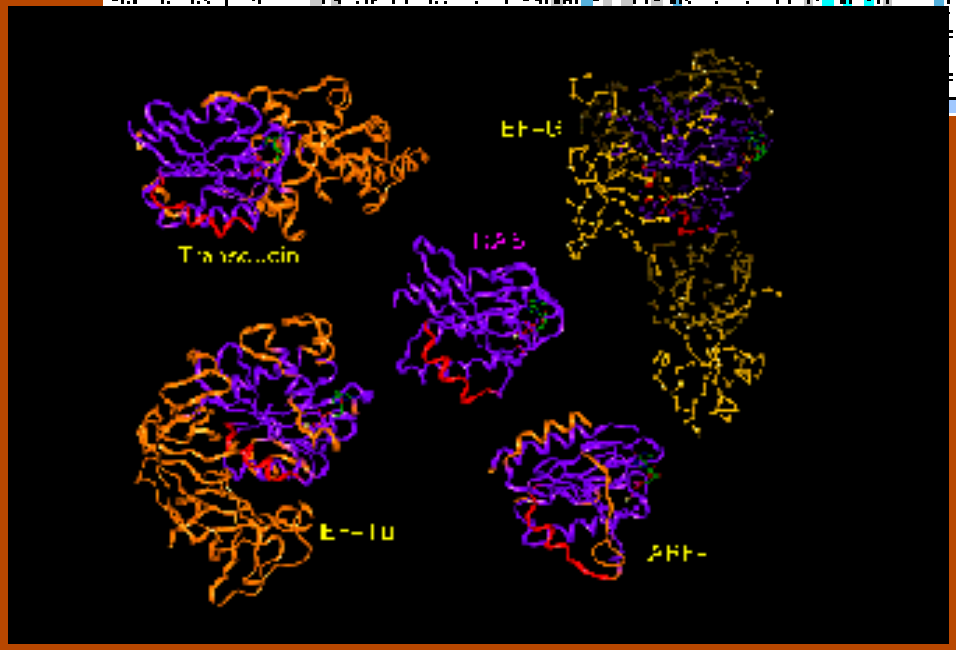
Sequence Alignment

Determining the pattern of evolution and identifying conserved regions that may be of structural or functional importance.



Structure Alignment

Structure is more conserved than sequence. If structural similarity existsiit is still possible to infer common ancestry, even in the absence of sequence similarity

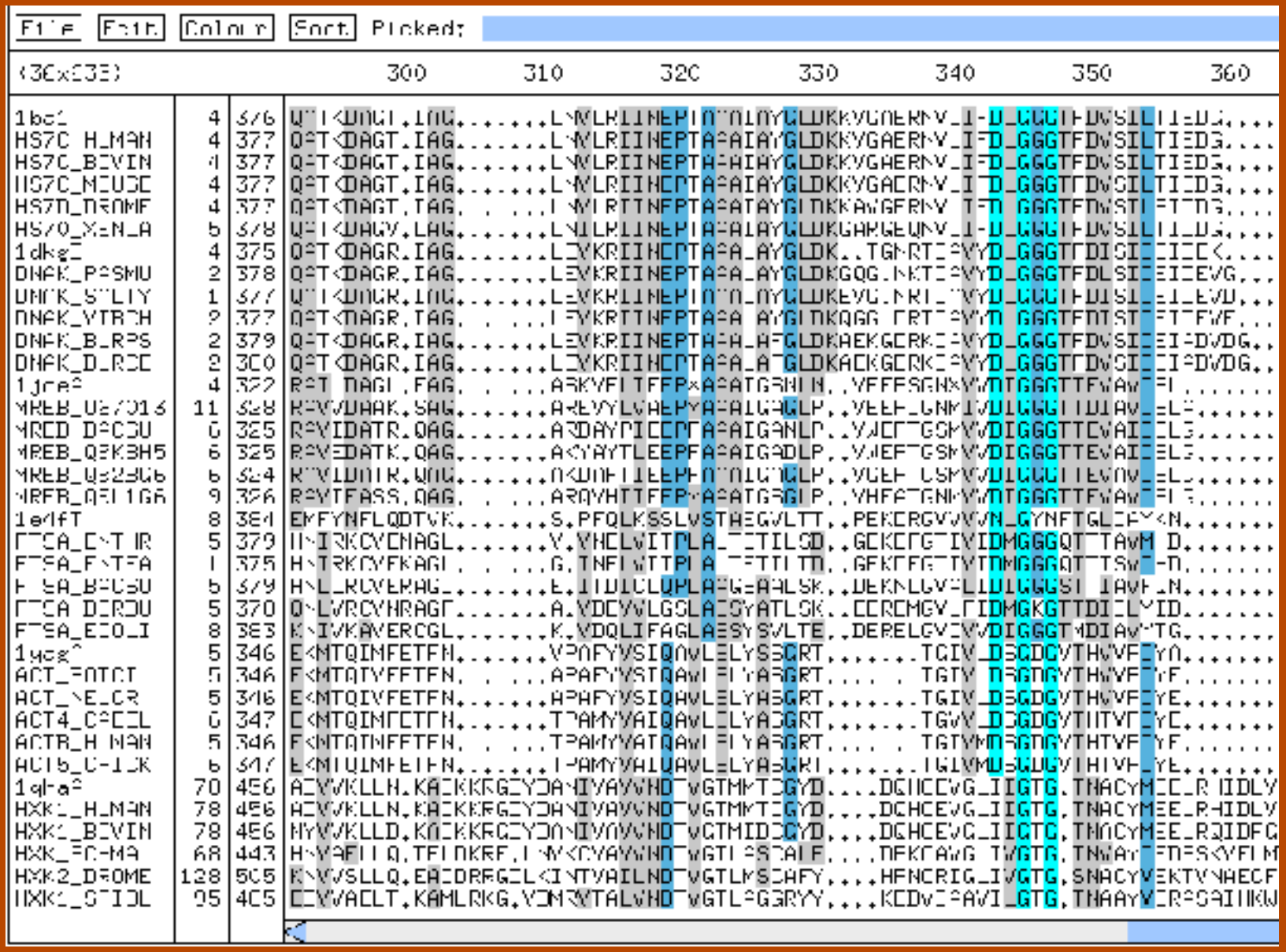


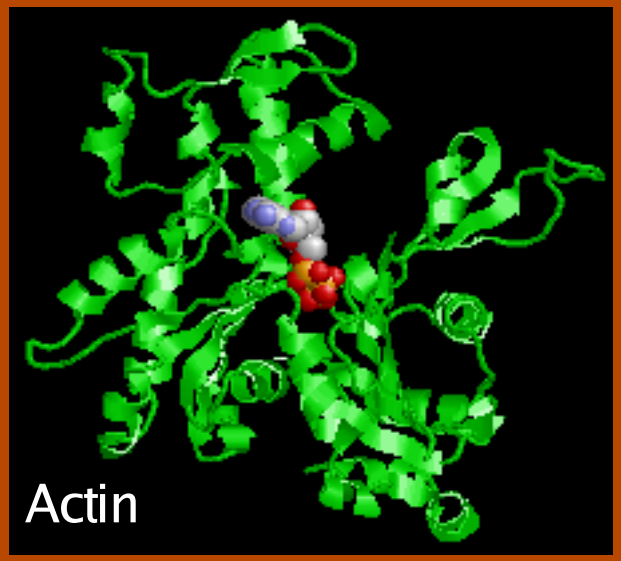
Sequence Analysis

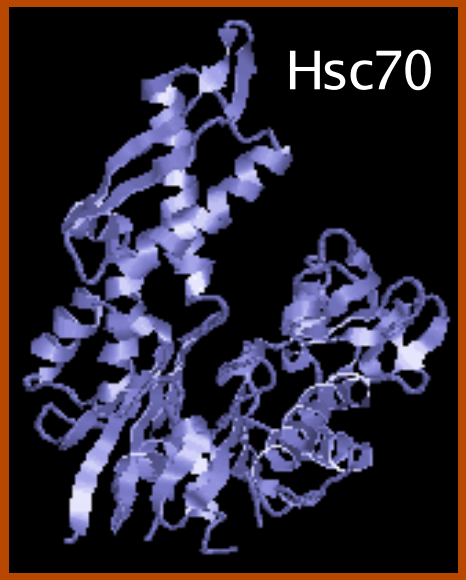
Aims

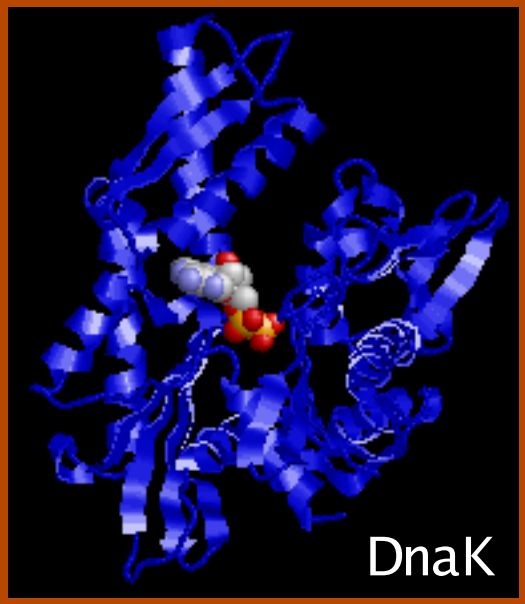
- To identify similarity between sequences.
- To deduce common ancestry from sequence similarity.
- Sequence classification
- Sequence similarity based function prediction.
- Sequence similarity based structure prediction
- The prediction of functional and physical interaction between proteins

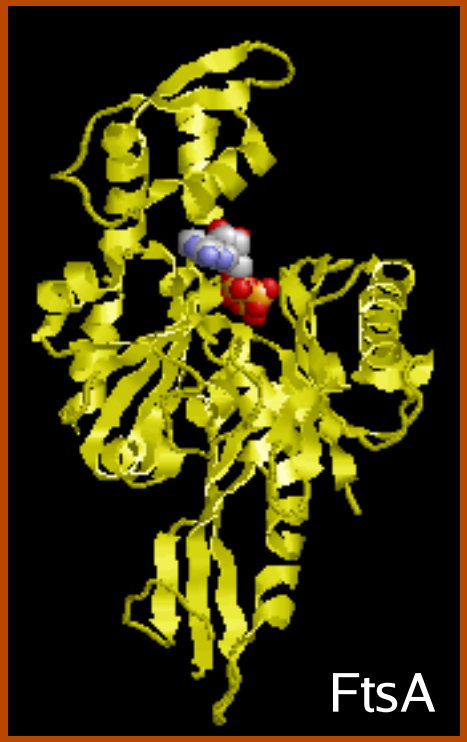
Chaperones (dnak), Proteins involved in bacterial cell wall production (ftsA, mreB), hexokinases (hxk), actin (act)

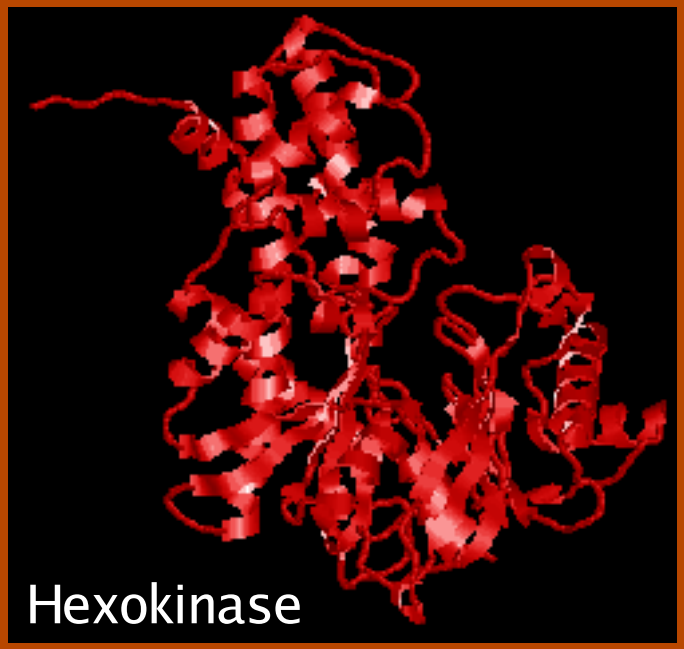



Actin


Hsc70


DnaK

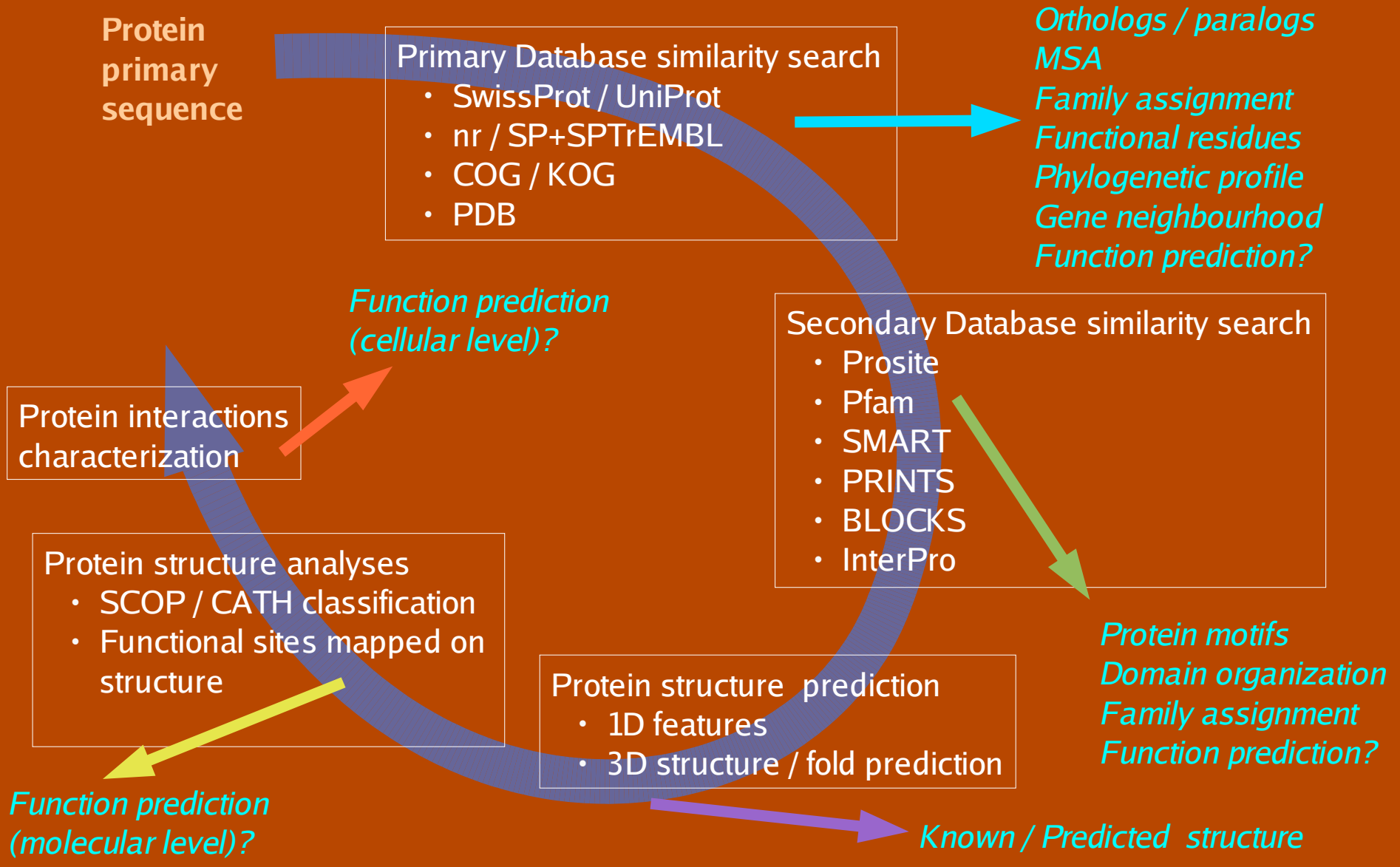

FtsA


Hexokinase

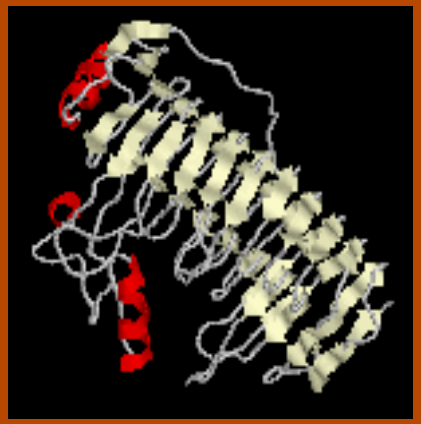

MreB

FUNCTION PREDICTION PROTOCOL

Based on sequence similarity, structural analyses and information about interacting partners.



Why Function Prediction is Important



The huge quantity of sequenced deposited in the protein databanks by genome sequencing projects will be dwarfed by the sequences from the environmental sequencing projects that are currently underway.

There is a growing imbalance between the number of sequences in databases and the information about structure and function regarding those sequences.

Functional and structural predictions could save a lot of time and effort and money, and they provide information about systems that are not experimentally treatable.

Why Pattern Recognition Methods Work. Mostly.

Sequence

The amino acid sequence of a protein determines its structure – in general similar sequences fold into similar structures.

Chaperonins

Structure

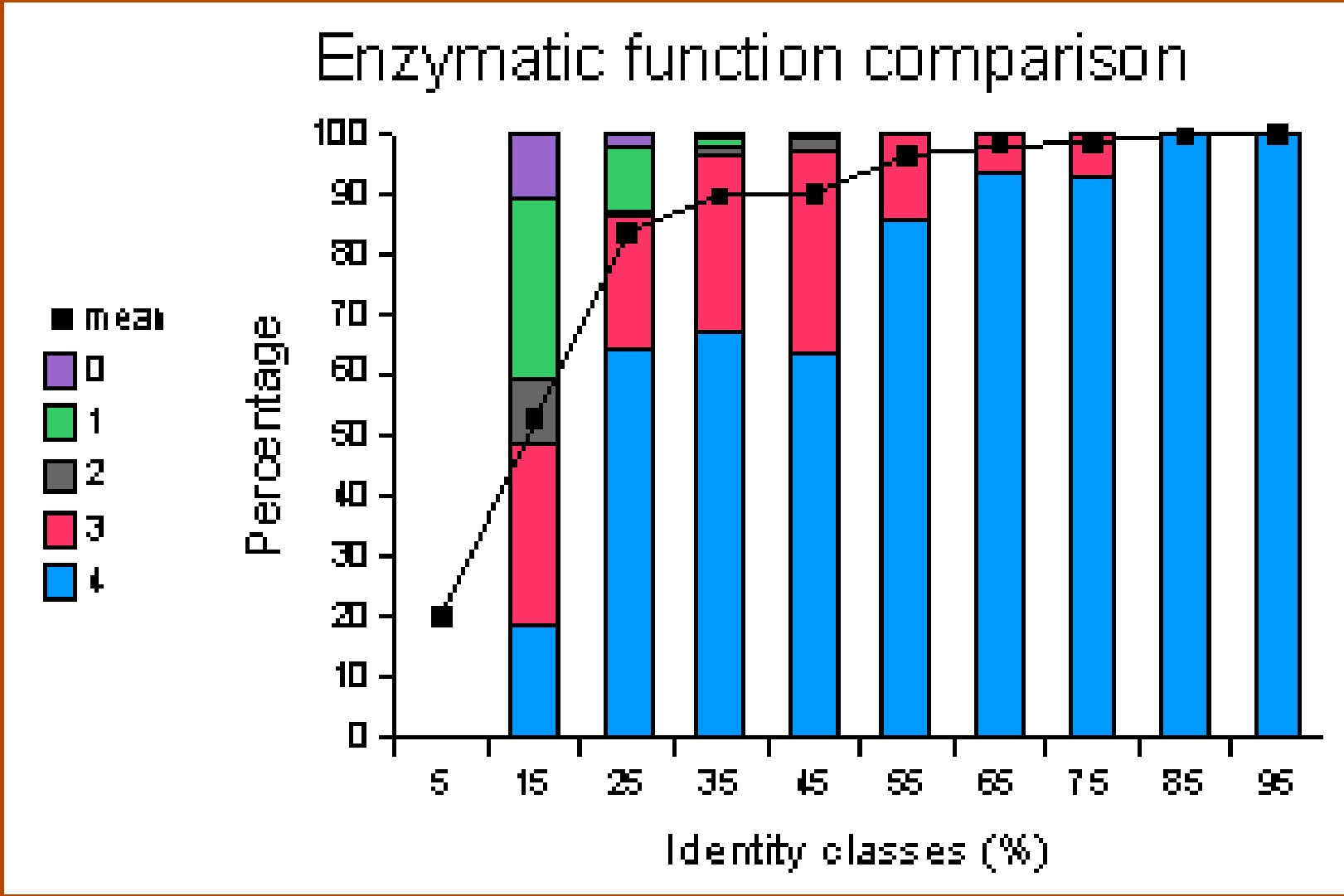
The structure of a protein determines its function. And in general similar protein structures perform similar functions.

Location

Function

Therefore, function prediction should be a simple matter of detecting similarity.

However, the relation between sequence similarity and function is far from simple.



Two Conclusions Drawn from Comparisons of Protein Sequences

First, protein sequences can be grouped in clusters or families based on sequence similarity.

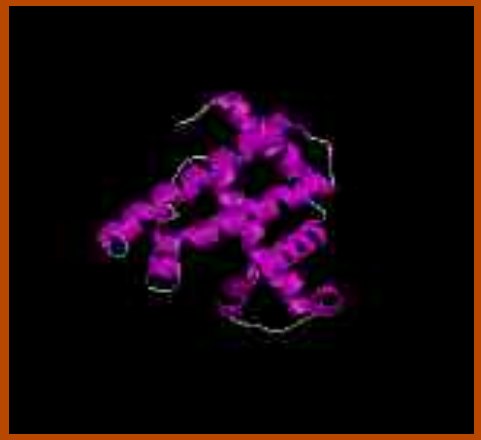
Second, this similarity may be restricted to short stretches of sequence. These conserved regions of sequence are generally regions of functional importance.

Clusters

Clustering techniques aim to identify groups of sequences that are maximally similar between them, and minimally similar to sequences in other clusters.

The terms superfamily, family and subfamily are often used. All the members of a superfamily, family or subfamily, are evolutionarily related.

The members of a superfamily generally have a similar structure, even though it may not be possible to detect sequence similarity. The members of a subfamily have clear sequence similarity and the same function.



1 Patterns and Profiles

Constructed from multiple alignments of evolutionarily related sequences. They include consensus sequences, regular expressions, profiles (PSSMs) and hidden Markov models (HMMs).

2. Conserved sub-sequences

Domains are stretches of sequence that appear as modules within proteins. They are usually more obvious structurally. Certain domains can be found in a wide range of proteins. Domain shuffling is an accepted mechanism of protein evolution.

Motifs are short, conserved sub-sequences that usually correspond to active or functional sites.

Motifs

Proteins from the same family are often characterised by short conserved regions that are usually related with function, such as binding sites or active sites.

The structural and functional restrictions placed on these motifs means that they are usually conserved even over large evolutionary distances.

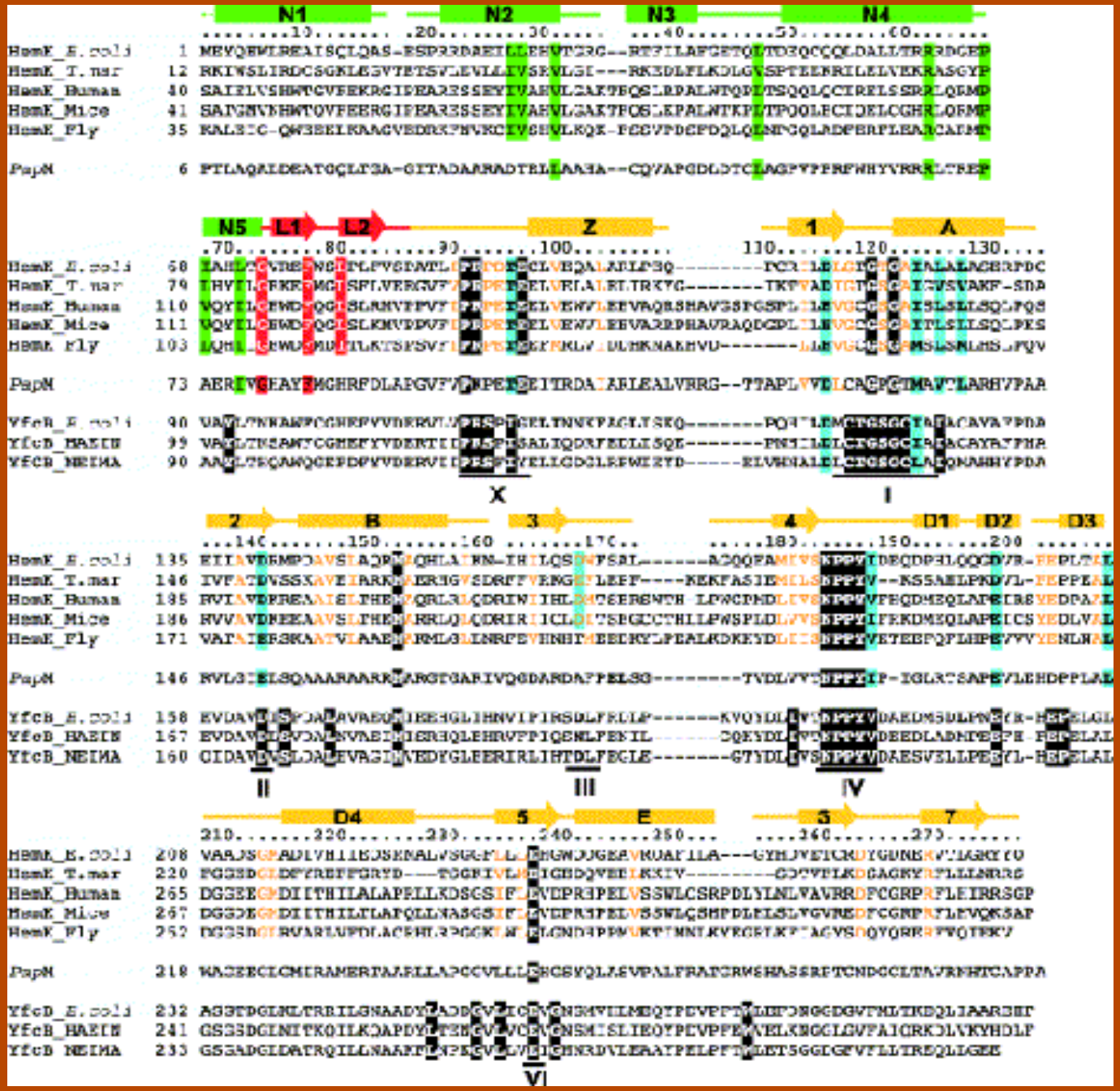
However, they are not usually detectable by sequence search techniques such as BLAST or FASTA.

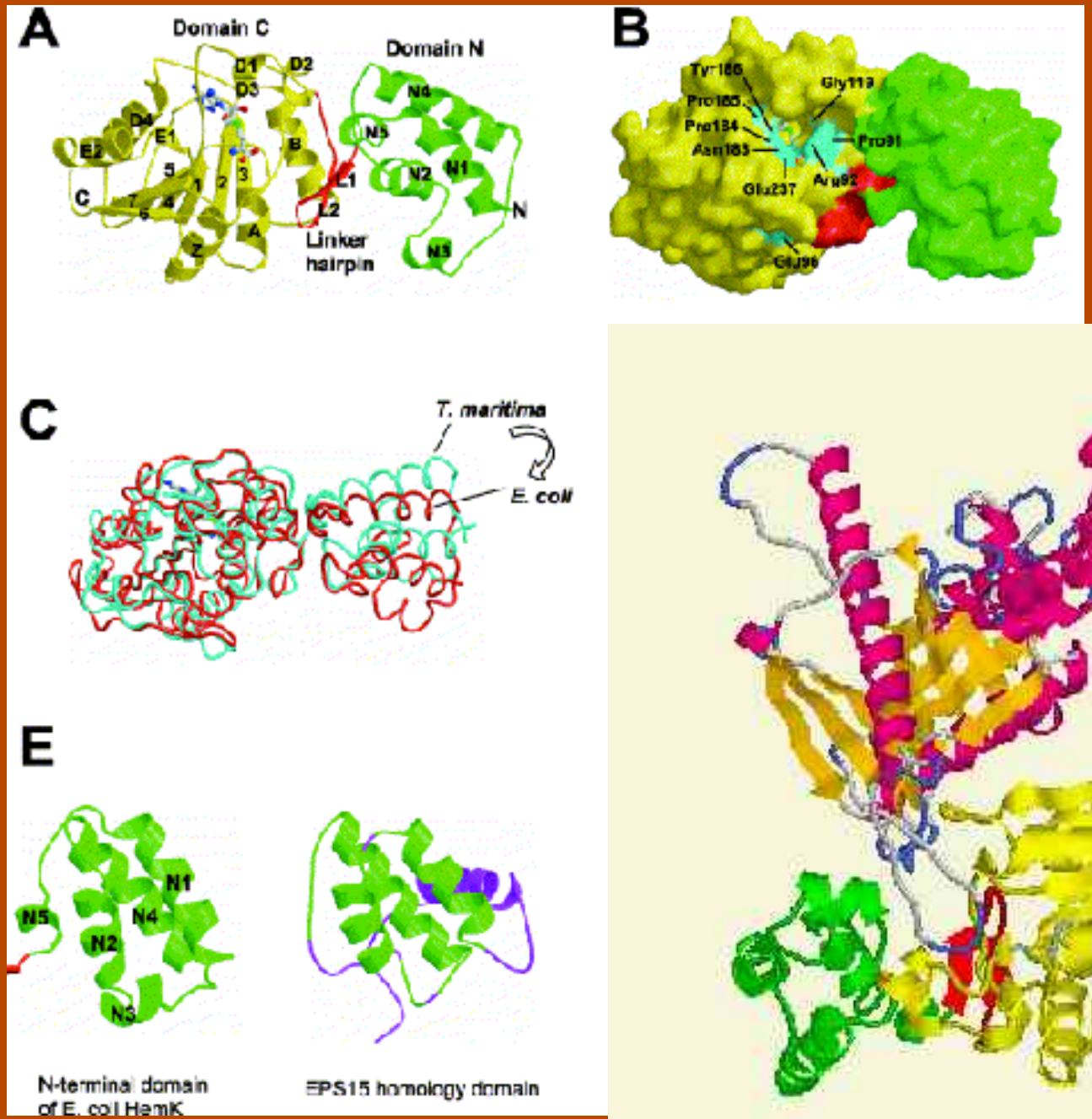
Instead there are motif databases, and tools for searching for motifs in sequences and for searching for sequences with a certain motif.

HemK

- an adenine methyltransferase?

The "NPPY" motif is characteristic of the DNA methylase family.

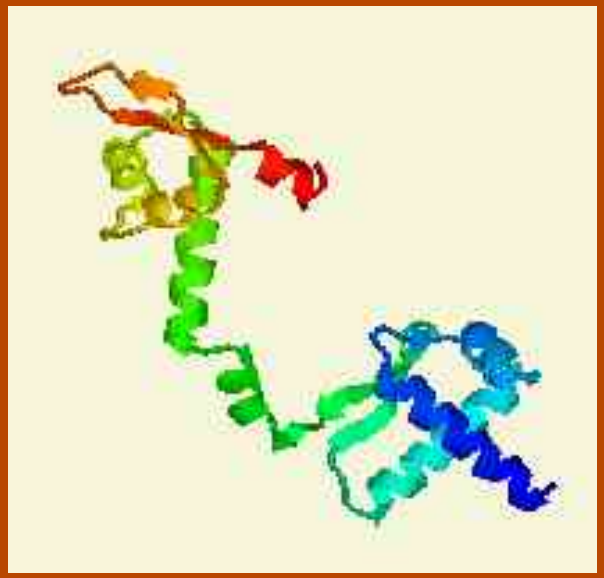


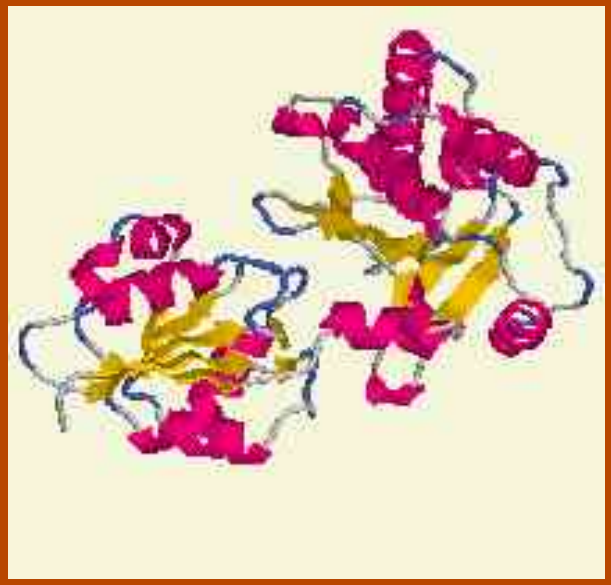


However HemK turns out to be the first known glutamine methyltransferase

Domains

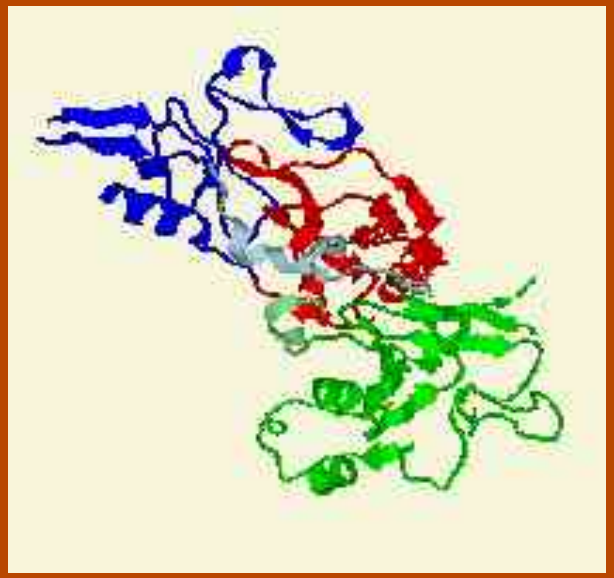
Domains are usually visible structurally, sometimes they are more obvious than others though ...





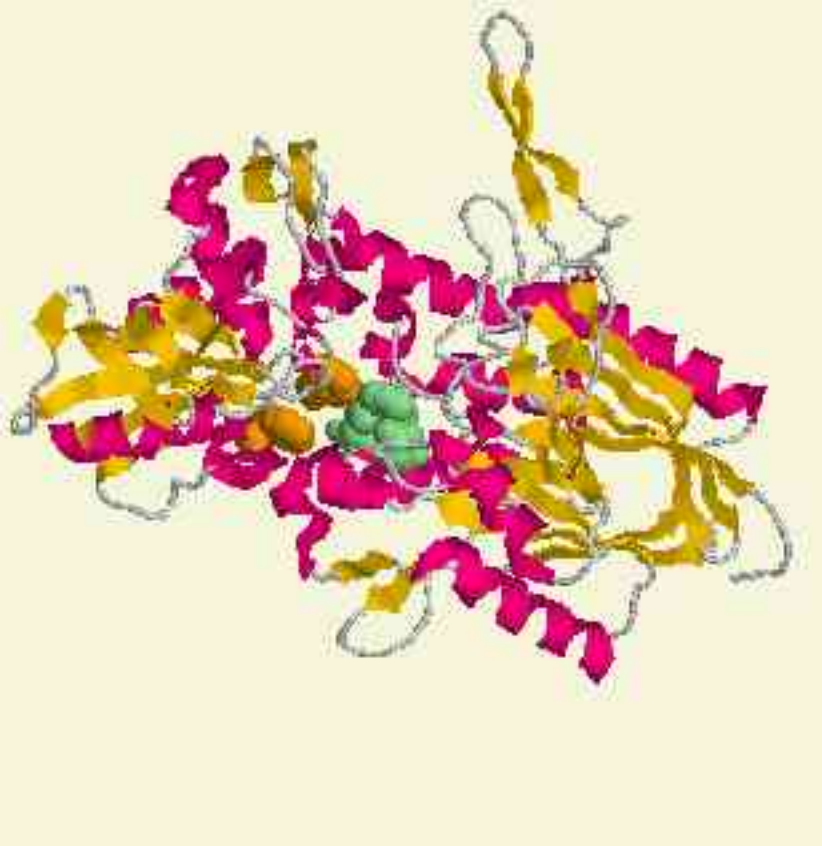
Diphtine synthase

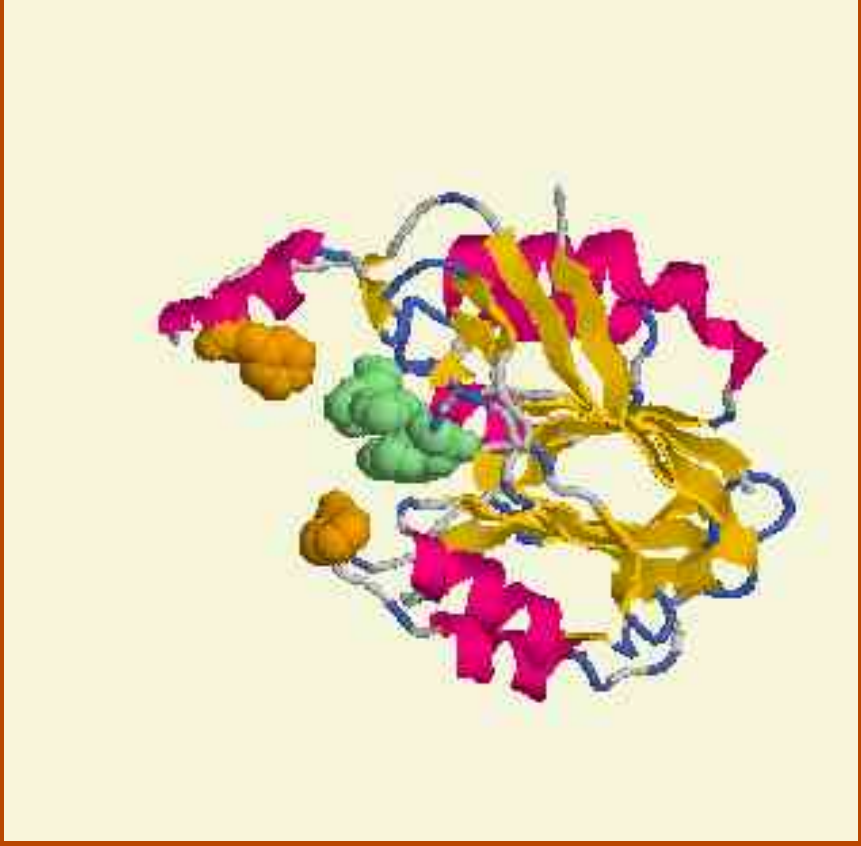
Heat shock operon repressor HrcA



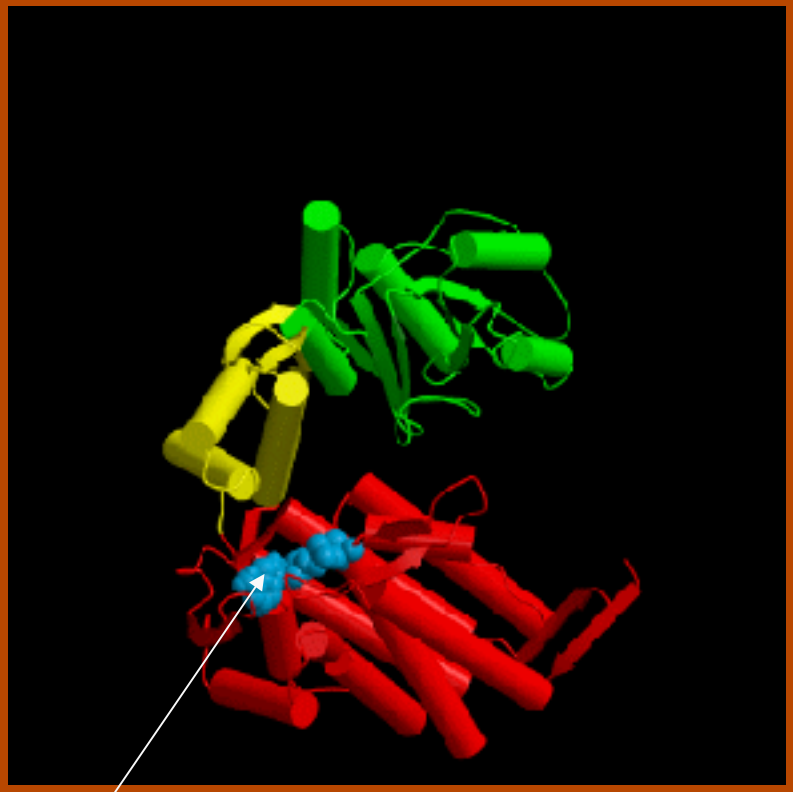
Domains

Domains often are associated with a single function, but sometimes both domains are involved – eq 1hpuD and T0200

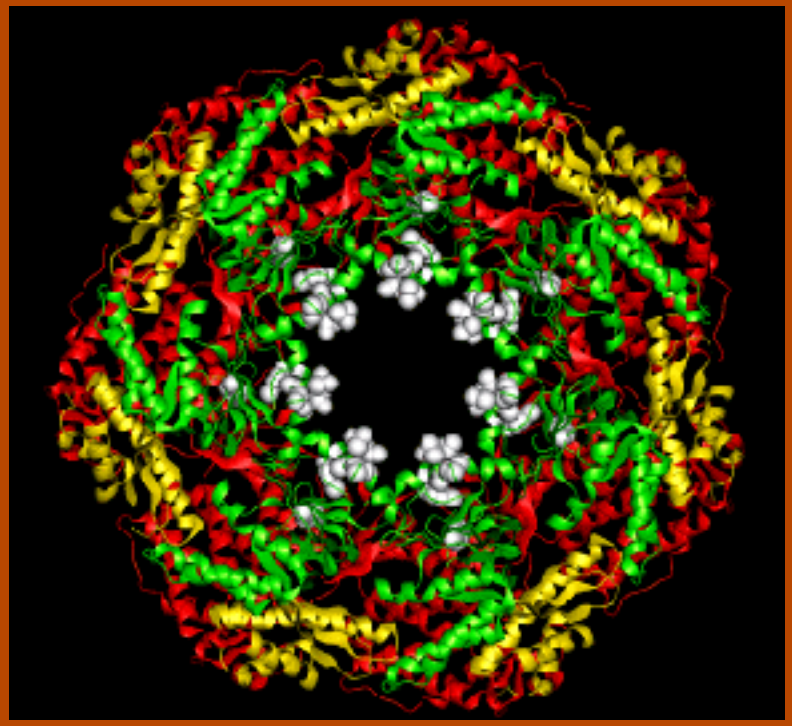




Domains and Chains



subunit



heptamer

Molecular chaperonin GroEL

(Dr Jianpeng Ma, Harvard Univ.)

Consensus Sequences

ALRDF**ATHD**DF
SMTAE**ATHD**SI
ECDQA**ATHE**AS

80% XXXXX**ATH**XXX

50% XXXXX**ATHD**XX

Regular Expressions

ALRDF**ATHD**DF
SMTAE**ATHD**SI
ECDQA**ATHE**AS
XXXXX**ATH[DE]**xx

- N terminal: <, C-terminal: >
- Any amino acid: x
- Ambiguity: [A,B…] A or B…
- or {A,B..} anything but A or B…
- Repeats: A(2,4) - A-A o A-A-A o A-A-A-A

[AC]-x-V-x(4)-{E,D}

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

PSI-BLAST

PSI-BLAST (Position Specific Iterated BLAST) is accesible at the NCBI web server. Like BLAST, PSI-BLAST performs database searches with query sequences as input.

After the first BLAST search a multiple sequence alignment is constructed from significant local alignments found in the first pass.

From that multiple alignment a profile (or PSSM) is built .

Profiles

	F	K	L	L	S	H	C	L	L	V
	F	K	A	F	G	Q	T	M	F	Q
	Y	P	I	V	G	Q	E	L	L	G
	F	P	V	V	K	E	A	I	L	K
	F	K	V	L	A	A	V	I	A	D
	L	E	F	I	S	E	C	I	I	Q
	F	K	L	L	G	N	V	L	V	C
A	-18	-10	**-1**	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-28	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Profiles are built based on the frequency of each amino acid at each position in the alignment. The relative frequency of each amino acid (or the no appearance) are weighted based on observed frequencies.

A is less probable than M because although it doesnt appear we know that M is simliar to L, I, V y F.

PSI-BLAST II

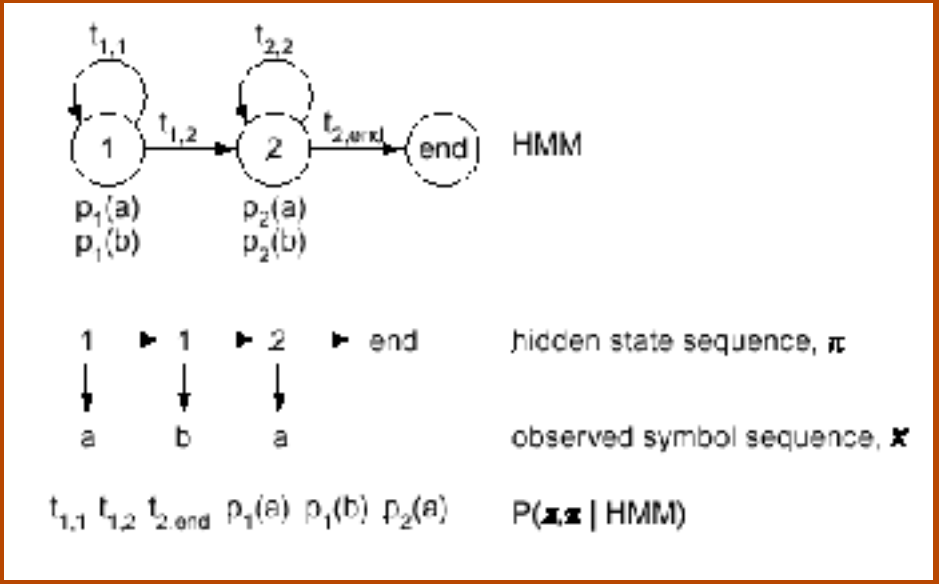
This profile is then used to search the database again, and any new significant hits are incorporated to the profile for further searches of the database.

The process iterates an arbitrary number of times or until no new sequences can be found in the database.

Hidden Markov Models I

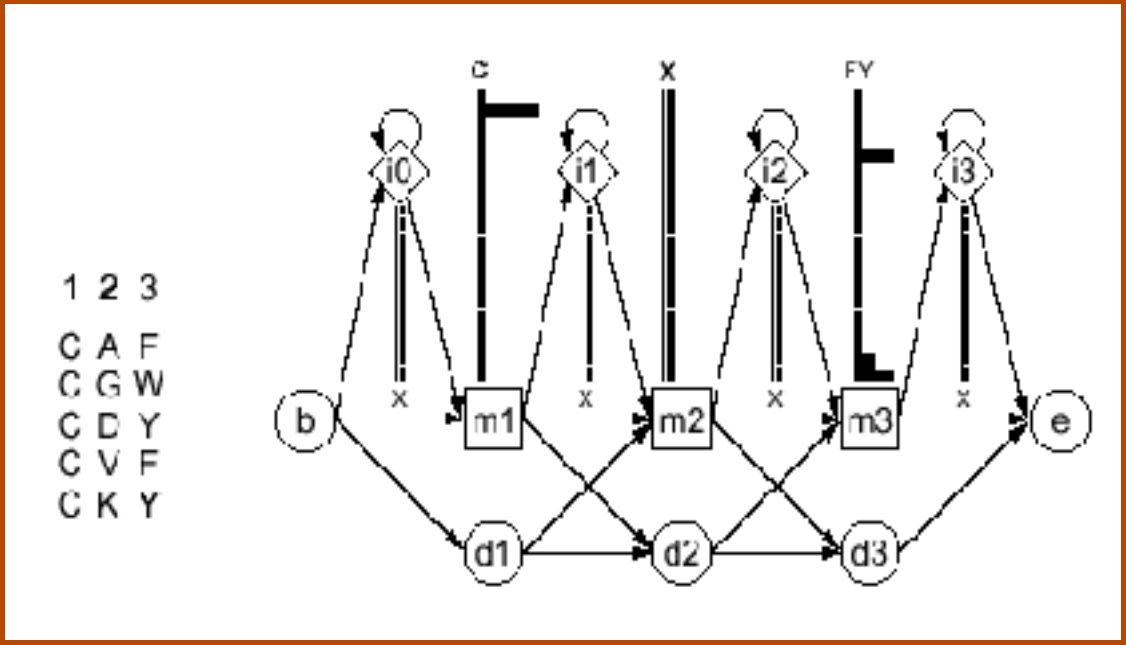
Hidden Markov models (HMMs) are statistical models of sequences.

They are similar to PSSMs, but the weights are calculated according to a probabilistic model that takes into account the previous positions.



Hidden Markov Models II

The model has three residues or states (m_1, m_2, m_3) with 20 probabilities to be a residue (bars), one state which is insertion (eg i_0) and one state which relates to deletion (eg d_1). The arrows represent the probability of the transition between each state.



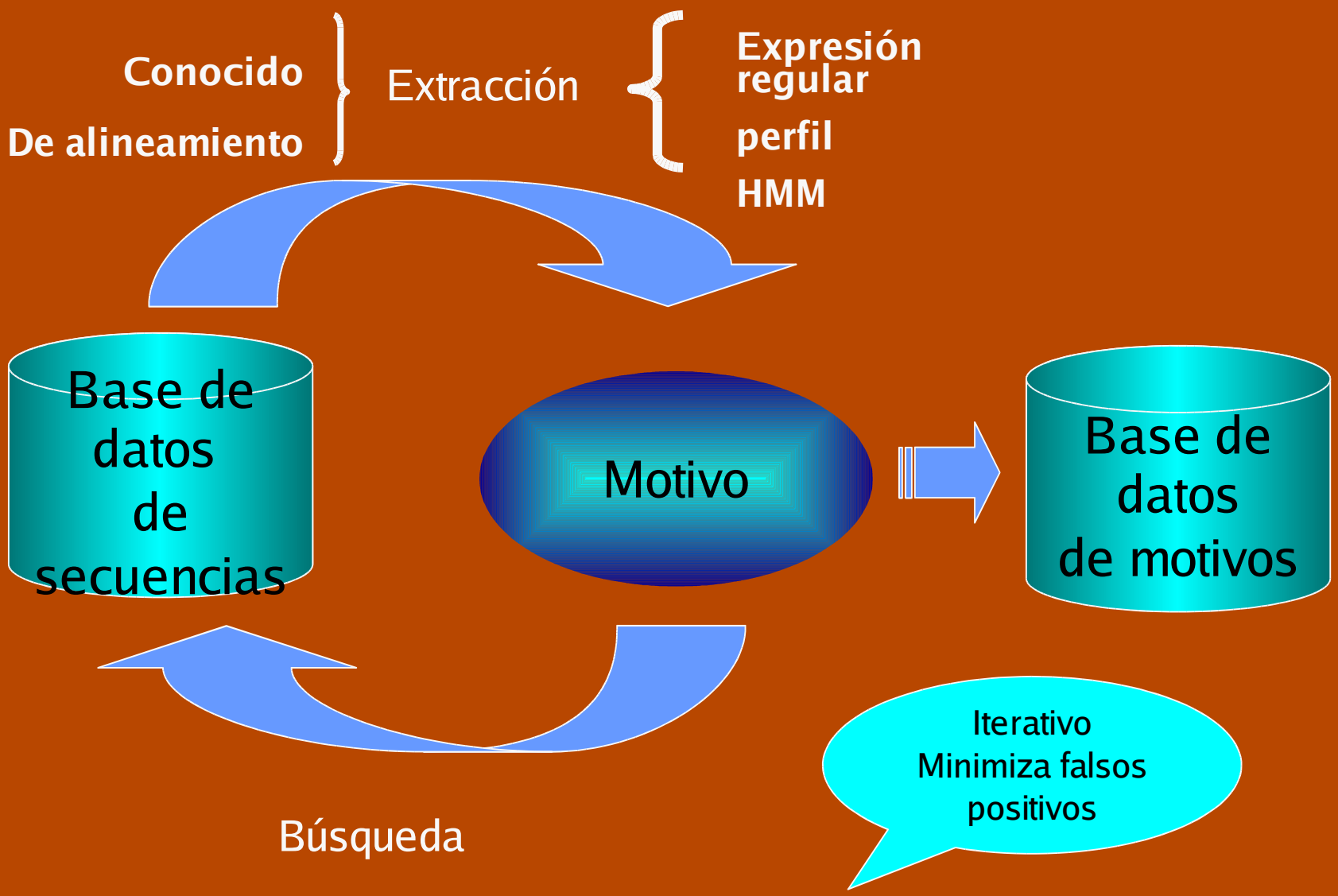
Motif Construction

Known motifs (published or found in known families)

Eg PROSITE

Empirical motifs: obtained directly from alignments. In some cases their significance is unknown.

Eg Pfam.

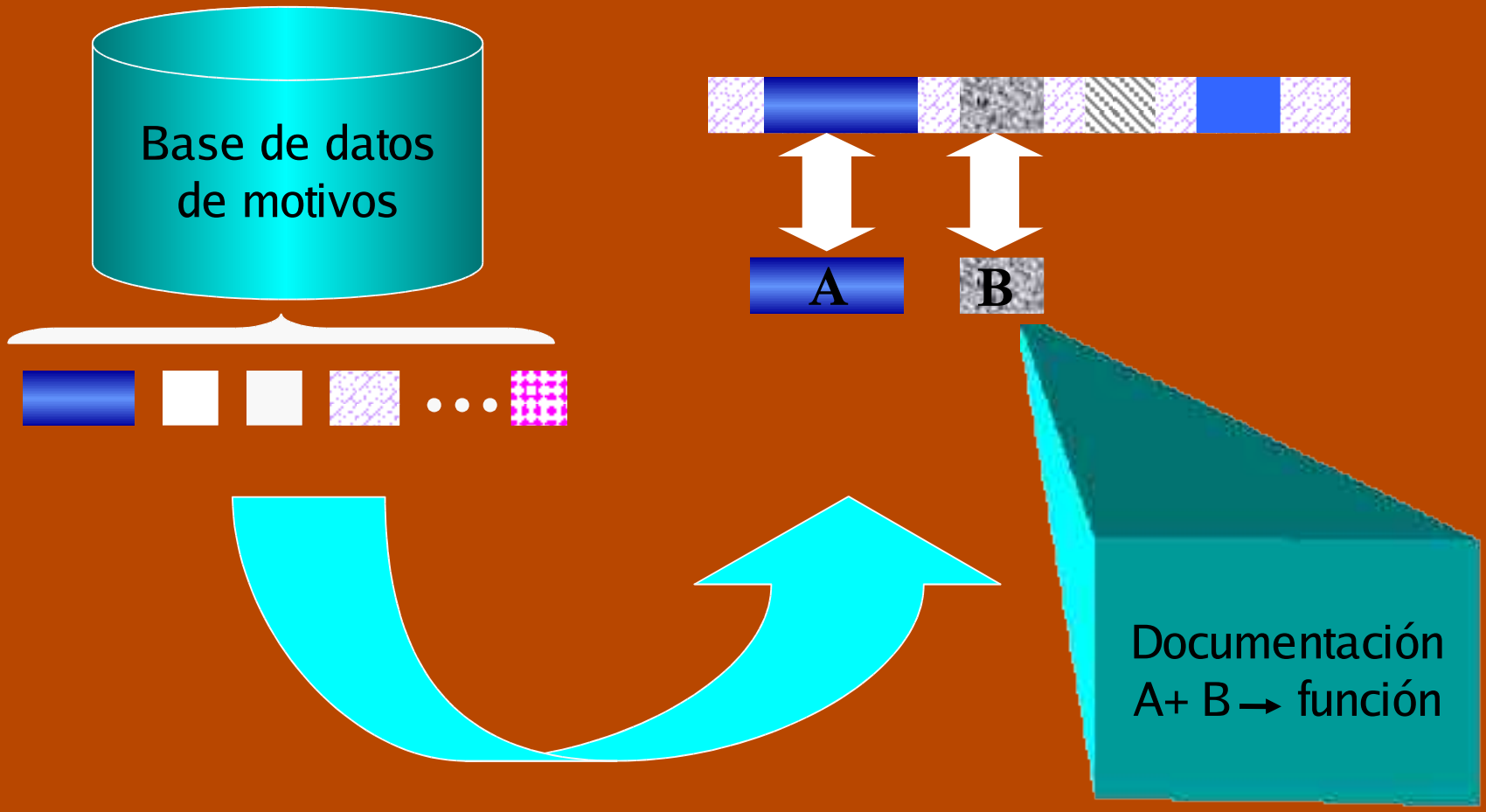

Conocido
De alineamiento
Extracción
Expresión regular
perfil
HMM
Base de datos de secuencias
Motivo
Base de datos de motivos
Búsqueda
Iterativo
Minimiza falsos positivos

Using Databases and Tools Based on Motifs and Domains

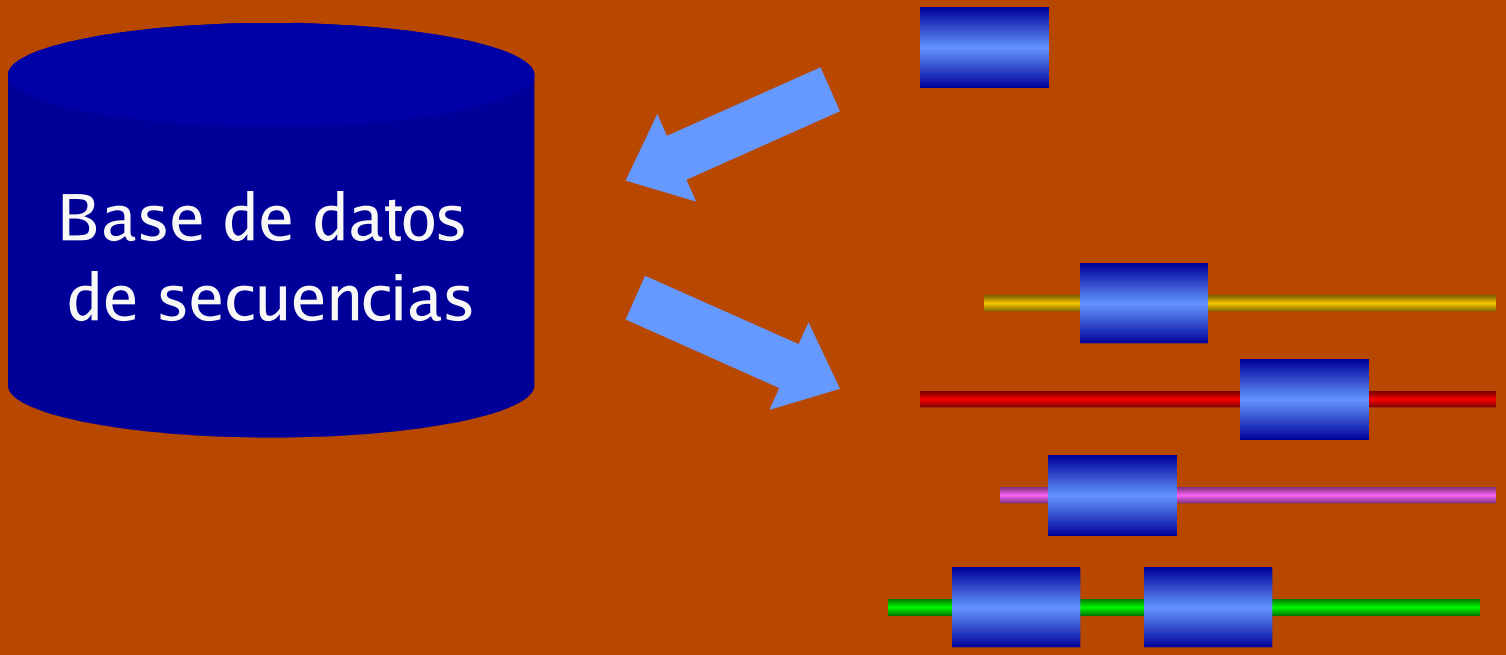
1. Function Prediction: Identifying motifs in a query sequence

2. Finding remote similarities: Searching all sequences in a database with a certain motif.

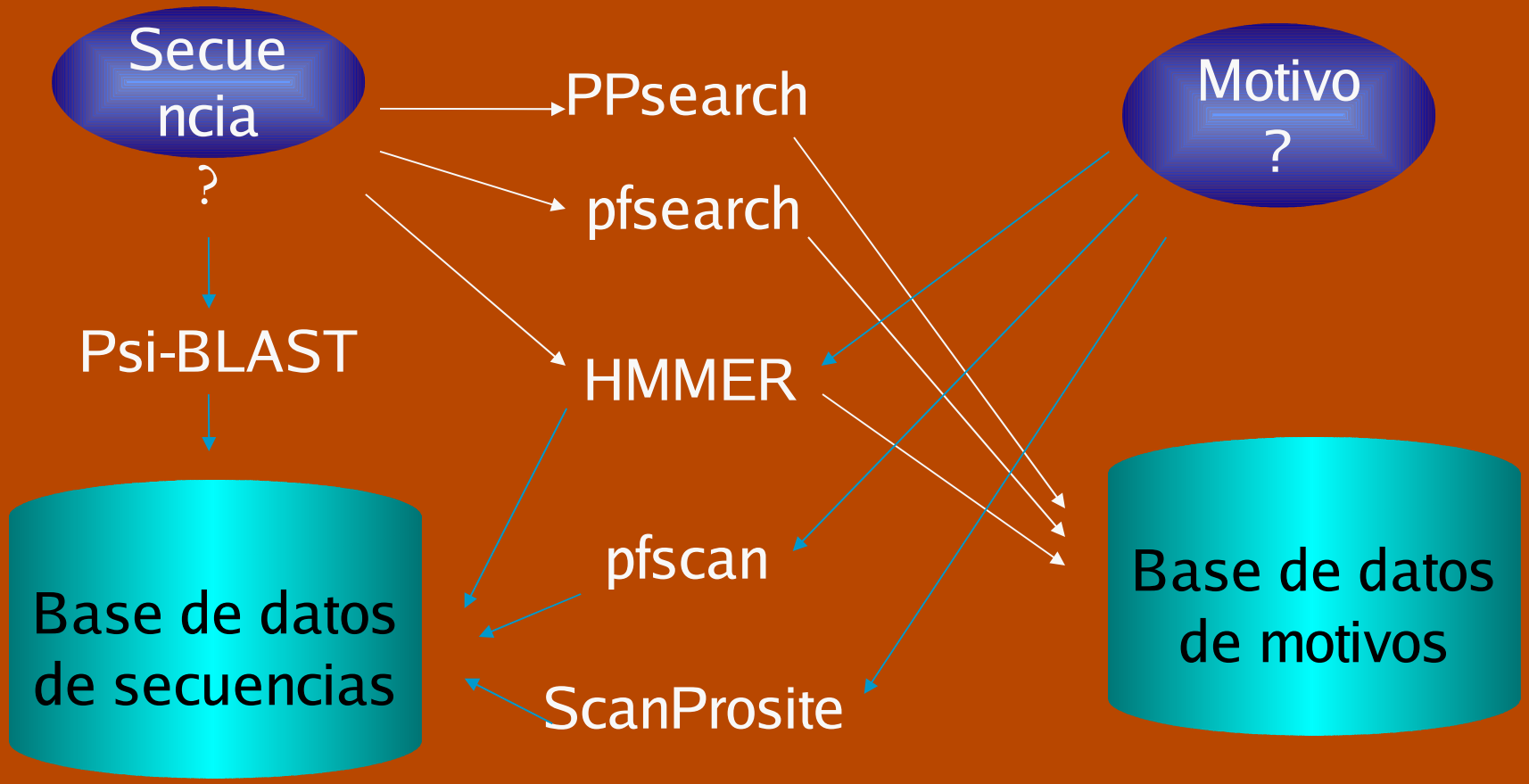
Function Prediction



Searching for Remote Similarity with Motifs



Tools


Secue
ncia
?
Psi-BLAST
Base de datos
de secuencias
PPsearch
pfsearch
HMMER
pfscan
ScanProsite
Motivo
?
Base de datos
de motivos

Databases

PROSITE Regular Expressions and profiles based on known motifs from SwissProt.

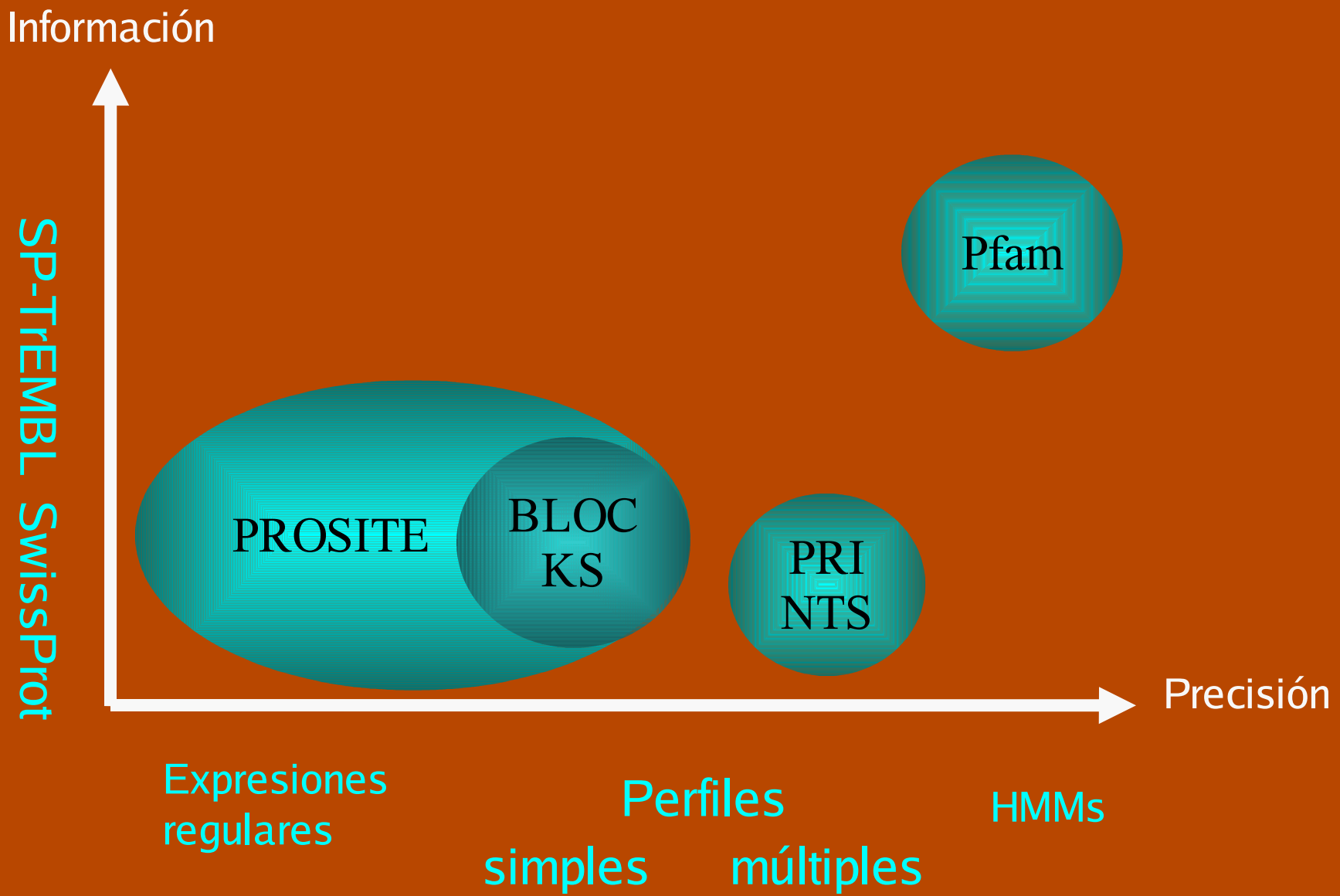
BLOCKS Profiles, based on PROSITE.

PRINTS Profiles, based on known motifs

Pfam Profiles from HMM. Motifs generated utomatically using SwissProt + SP-TrEMBL

SMART Domains identified with PSI-BLAST and HMMs.

InterPro Hierarchical database of domains and motifs that integrates information from Prosite, PRINTS, Pfam, Prodom, SMART and TIGRFAMs.


Información
SP-TrEMBL SwissProt
Pfam
PROSITE
BLOCKS
PRINTS
Precisión
Expresiones regulares
Perfiles
simples
múltiples
HMMs

This presentation contains material from:

Manuel J Gómez, CAB
Oswaldo Trelles, UMA
Joaquín Dopazo, CNIO
Paulino Gómez Puertas, CBM