今期から始まった遠隔授業の受講者の方々にも対応するため、早めにファイルを置いていますが、授業直前に内容を修正する可能性があります。2010年4月19日

#7

Sequence Motifs

Topics:

- Sequence motif representation
  - Regular expression, Profile matrix, Hidden Markov Model (HMM)
- Extracting a fixed-length motif
  - Relative Entropy Score
  - Approximation algorithm using Gibbs sampler
- Motif databases
  - PROSITE, BLOCKS, PRODOM, PFAM
  - integrated motif search system Interpro

Sequence Motif

Motif = A “common pattern”
well-conserved in a group of homologous sequences

How to discover

1) Multiple alignment first, and then extract conserved regions.
2) Fix motif length *L*, and search the optimal one with *L.*

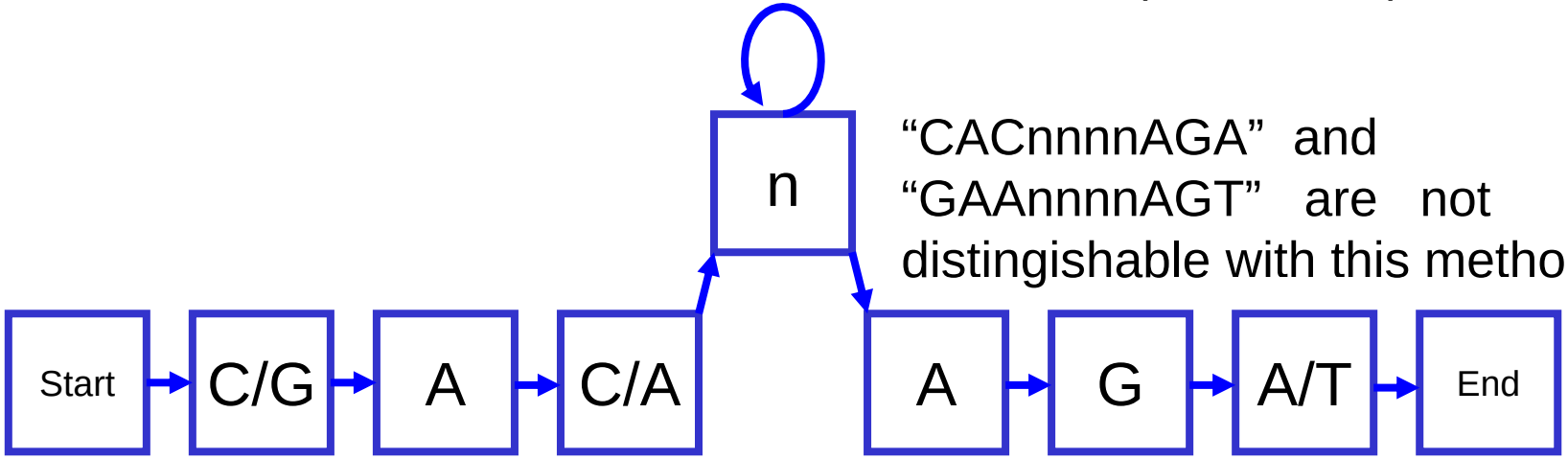
How to represent

1) Regular expression
2) Profile matrix
3) Hidden Markov Model (HMM), etc.

Motif Representation: Regular Expression

Position	1	2	3	4	5	6	7	8	9	10	11
gene A	C	A	C	a	a	a	c	g	A	G	A
gene B	C	A	C	a	t	g	g	-	A	G	T
gene C	C	A	A	t	c	t	a	-	A	G	A
gene D	G	A	C	c	g	c	t	-	A	G	A
gene E	C	A	C	a	c	t	-	-	A	G	A

example DNA sequences



"CACnnnnAGA" and "GAAnnnnAGT" are not distingishable with this method.

[C G]- A - [C A] - [ATGC]* - A - G - [A T]

Exercise

Which amino acid sequence agrees with the following motif ?
Choose one from the following four sequences.

Regular expression: C−x(2,4)−C−[LIV]−H

where

・[ ] is disjunctive OR. Any one element in [ ] can be matched.

・x(*a, b*) is a series of any spacing characters at least *a* and up to *b* characters.

・−is a connection of characters.

1 CPKRLH
2 CPKRCLVH
3 CPKRGCIH
4 CPKRGKCVH

Cited from "JSBi Bioinformatics Certificate 2007" (originally in Japanese)

Motif Representation: Profile

Position	1	2	3	4	5	6	7	8	9	10	11
gene A	C	A	C	a	a	a	c	g	A	G	A
gene B	C	A	C	a	t	g	g	-	A	G	T
gene C	C	A	A	t	c	t	a	-	A	G	A
gene D	G	A	C	c	g	c	t	-	A	G	A
gene E	C	A	C	a	c	t	-	-	A	G	A

	1	2	3	4	5	6	7	8	9	10	11
A	0	1.0	0.2						1.0	0	0.8
C	0.8	0	0.8			0.25			0	0	0
G	0.2	0	0						0	1.0	0
T	0	0	0						0	0	0.2
-	0	0	0	0	0	0.2	0.75	0	0	0	0
--	0	0	0	0	0	0	0	1.0	0	0	0

Basically, only fixed length motif can be represented.

- is gap opening probability, and -- is gap closing probability

Exercise

A motif is represented by a PSSM (Position Specific Score Matrix).
Which DNA sequence get the highest score with the PSSM?
Choose one from the following four sequences.

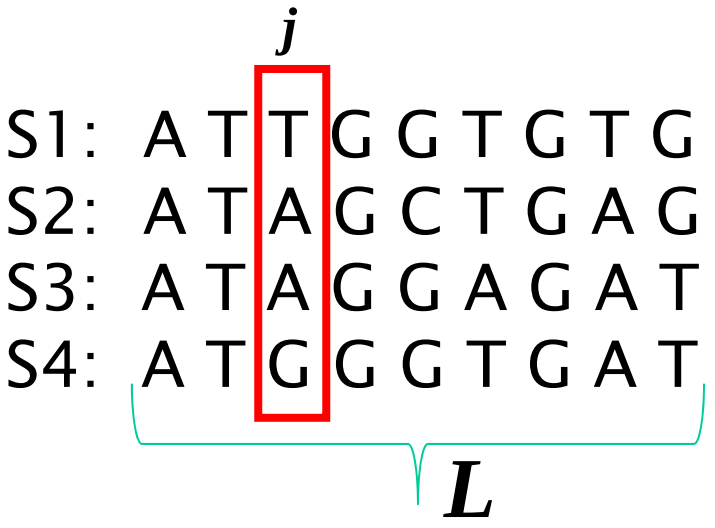
	position				
	1	2	3	4	5
A	6	-3	-3	0	-3
C	-9	0	-5	-3	6
G	-3	7	-4	-7	0
T	2	-3	0	0	-3

1 AGTAC
2 CACGA
3 TCTTG
4 TGTTC

Cited from “JSBi Bioinformatics Certificate 2007” (originally in Japanese)

Relative Entropy

$$\sum_{j=1}^{L}\sum_{a\in\Sigma} f_j(a)\cdot\log\frac{f_j(a)}{p(a)}$$



L: Length of sequence

a: character （for DNA sequence $a \in \Sigma=\{A,T,G,C\}$）

$p(a)$: background probability of a

$f_j(a)$: frequency of character a at position j in a motif

- also called as Kullback-Leibler Divergence

Exercise

Calculate relative entropy score for the following sequence motif (L=5) with four DNA sequences. The score is defined by the equation below. Here, background probabilities are given as p(A)=p(G)=p(T)=p(C) = 0.25.

$$\text{Relative Entropy Score} = \sum_{j=1}^{L} \sum_{a \in \Sigma} f_j(a) \cdot \log_2 \frac{f_j(a)}{p(a)}$$

Motif Sequence

	1	2	3	4	5
Seq. 1 :	A	T	A	T	G
Seq. 2 :	A	T	T	T	G
Seq. 3 :	A	A	G	T	C
Seq. 4 :	A	A	C	T	C

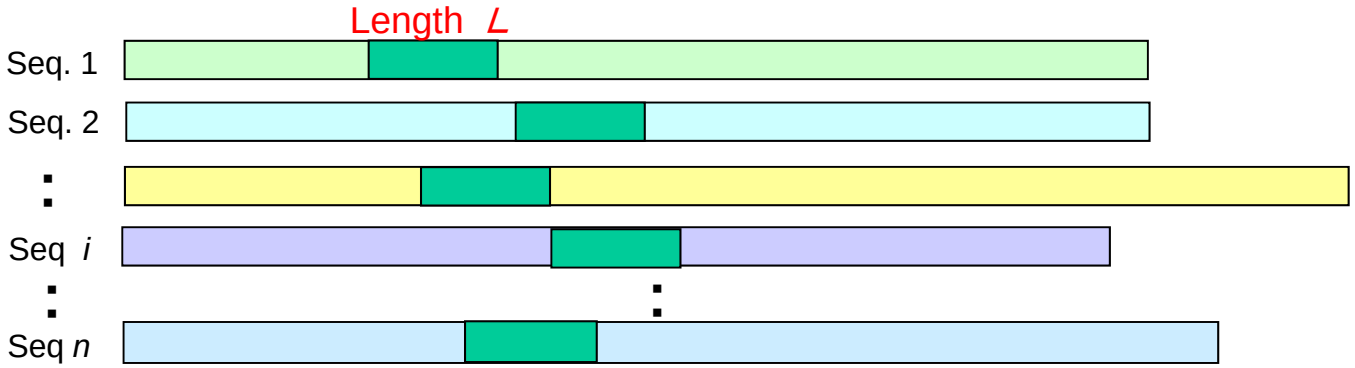
$f_j(a)$

(0.25 × log2 (0.25 / 0.25)) × 4 = 0.25 × 0 × 4 = 0.0

(0.5 × log2 (0.5 / 0.25)) × 2 = 0.5 × 1 × 2 = 1.0

(1.0 × log2 (1.0 / 0.25)) × 1 = 1.0 × 2 × 1 = 2.0

Extracting a fixed-length motif



Goal: Maximize the Relative Entropy value defined below, by sliding short windows with length *L* on each of *n* biological sequences.
OOPS: exactly One Occurrence of a motif Per Sequence
(cf. ZOOP: Zero or One Occurrence Per Sequence)

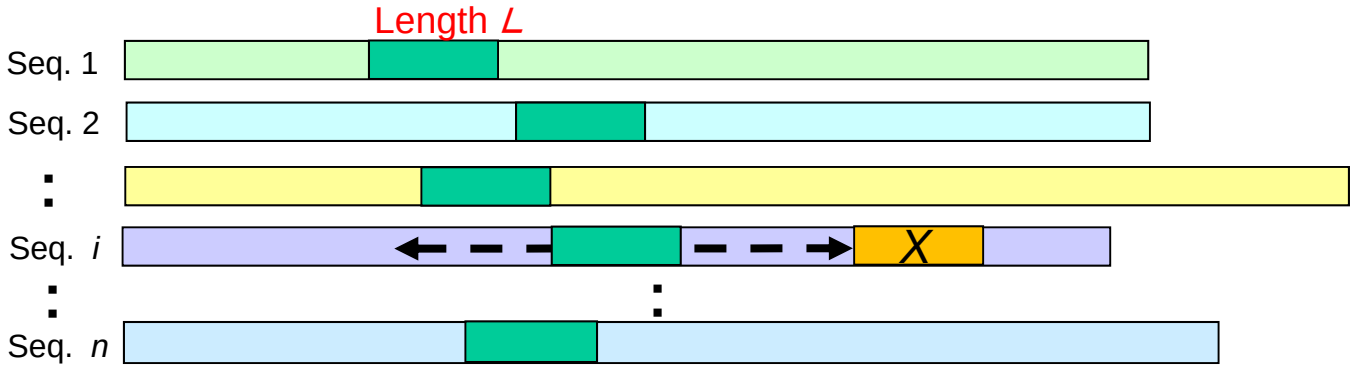
$$\textbf{Relative Entropy} = \sum_{j=1}^{L} \sum_{a \in \Sigma} f_j(a) \cdot \log \frac{f_j(a)}{p(a)}$$

a: character, *p(a)*: background probability, *fj(a)*: frequency of character a at motif position *j*

For larger number of *n*, rigorous global optimization requires exponential time.
Approximated methods are required, just like as multiple alignment.

Extracting a fixed-length motif

Approximation algorithm (using Gibbs sampling)



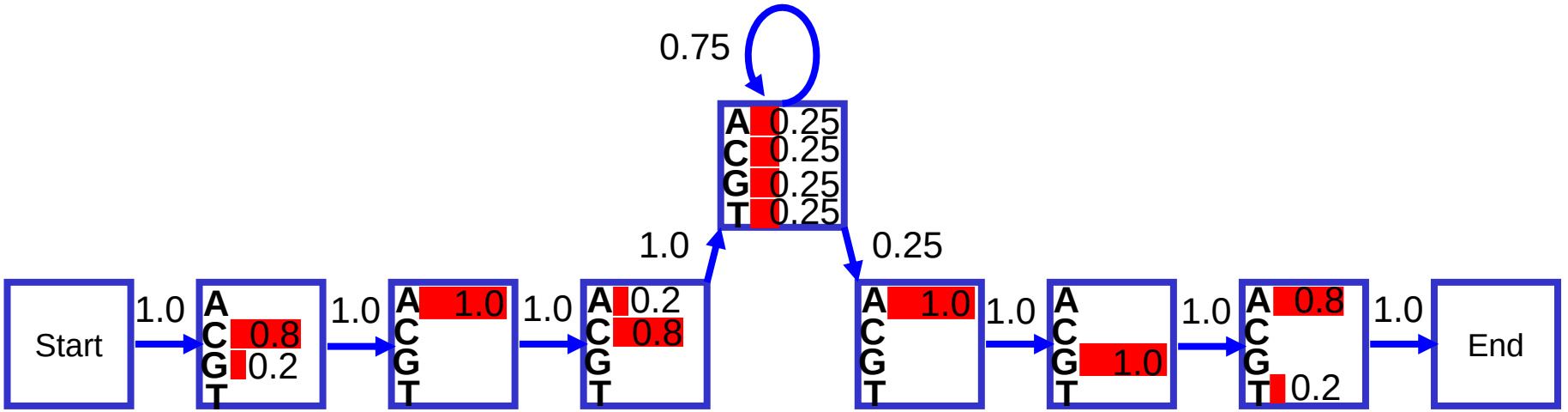
Step 1: Randomly choose an initial subsequence of length *L* on each seq (*1 to n).*
Step 2: (just like as the iterative improvement method of multiple alignment ...)
randomly choose one sequence from *n* sequences.（seq *i* hearinafter）
Step 3: On selected seq *i* , update the position of selected motif subseq. to X
so that X shows better similarity with other n-1 selected subsequences.。
Next position X is stochastically selected with a probability proportional to

$$score(x)=\prod_{j=1}^{L}\frac{f_j(x[j])}{p(x[j])}$$

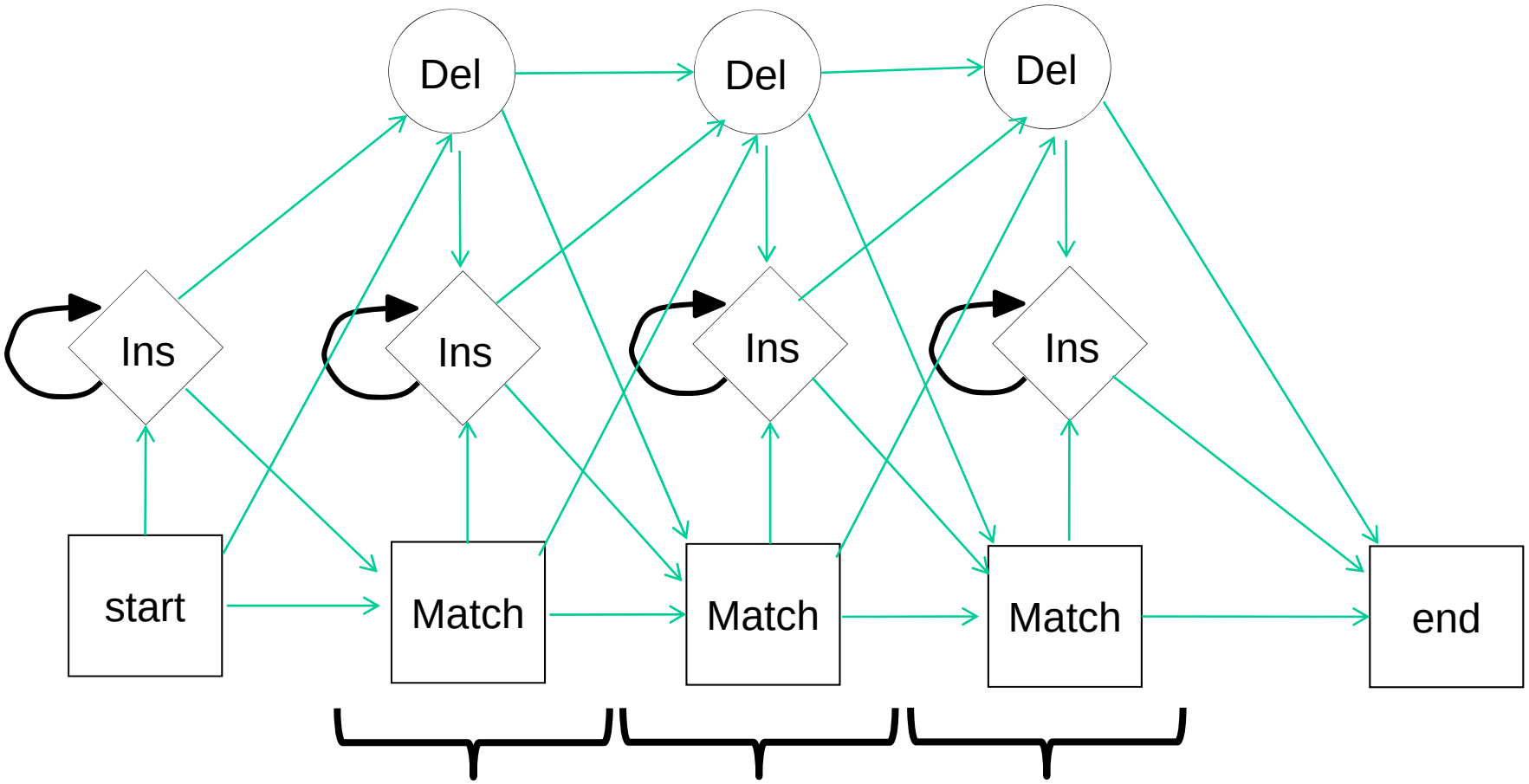
Repeat step 2 and 3 enough times, and stop when no improvement observed.
The final result depends on initial states in step1, and random numbers in step2 & 3.

Motif Representation: HMM

1	2	3	4	5	6	7	8	9	10	11
C	A	C	a	a	a	c	g	A	G	A
C	A	C	a	t	g	g	-	A	G	T
C	A	A	t	c	t	a	-	A	G	A
G	A	C	c	g	c	t	-	A	G	A
C	A	C	a	c	t	-	-	A	G	A



Profile HMM



corresponding to
each column position
in a sequence motif

Motif DB (1): PROSITE

http://expasy.org/prosite/

Swiss Institute of Bioinformatics, Geneva

Expert Protein Analyzer System = EXPASY Dr Amos Bairoch.

1)signature pattern

OR any*8

example (Zinc Finger motif):

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

PROSITE's original representation method.

(limited sensitivity, but easy to understand for human.)

2) matrix representation

Profile matix. (higher performance)

Bucher P., Bairoch A. : "A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation", ISMB-94; Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology. pp53-61, AAAIPress, Menlo Park, (1994).
MEDLINE: 7584418

Example: PDOC50020 (WW / rsp5 / WWP)

The “WW motif” is found on several proteins including “dystrophin”.
Mutations in the dystrophin lead to muscular dystrophy of Duchenne or Becker type.

ExPASy Home page | Site Map | Search ExPASy | Contact us | PROSITE | Swiss-Prot

Search Swiss-Prot/TrEMBL for Go Clear

Home | ScanProsite | ProRule | Documents | Downloads

PROSITE documentation PDOC50020

WW/rsp5/WWP domain signature and profile

Description:

The WW domain [1,2,3,4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain β-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- Dystrophn, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C- terminal globular domain. Dystrophin form

1) Pattern PS01159

2) Matrix PS50020

any but S or A

W - x(9,11) - [VFY] - [FYW] - x(6,7) - [GSTNE] - [GSTQCR] - [FYW] - {R} - {SA} - P

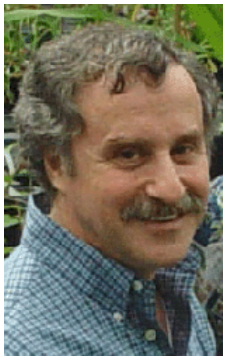
known to interact with proline (P) rich fragment of **[AP]-P-P-[AP]-Y**

Motif DB (2): BLOCKS

http://www.blocks.fhcrc.org/

Fred Hutchinson Cancer Research Center, Seattle, USA

non-gap blocks obtained from multiple sequence alignment.
BLOSUM matrix were generated from BLOCKS.
Recent versions are made from Interpro results.



Steven Henikoff

Henikoff, S. and Henikoff ,J.G.: “Automated assembly of protein blocks for database searching”, *Nucleic Acids Res.,* 19, pp.6565-6572. (1991).

J.G. Henikoff, E.A. Greene, S. Pietrokovski & S. Henikoff: "Increased coverage of protein families with the blocks database servers", *Nucl. Acids Res.* 28, pp.228-230 (2000).

Motif DB (3): PRODOM

INRA/CNRS France

http://protein.toulouse.inra.fr/prodom.html

http://www.biochem.ucl.ac.uk/bsm/dbbrowser/protocol/prodomqry.html

Automatically generated using PSI-BLAST search on SwissProt + TREMBL amino acid sequence databases.

Corpet F., Gouzy J., Kahn D.

“Recent improvements of the ProDom database of protein domain families.”

Nucleic Acids Res., 27, pp.263-267 (1999) .

MEDLINE: 99063708.