

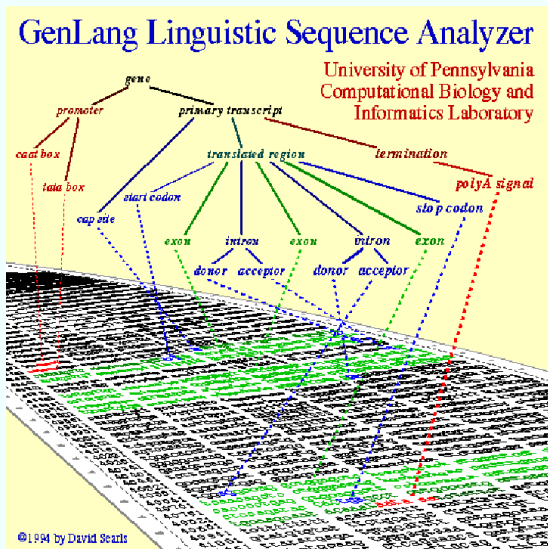
Modelling Biological Sequences by Grammatical Inference

François Coste



ICGI 2010 Tutorial Day

Motivation: learning the grammar of DNA



This talk: *The Paleolithic* of Modelling Biological Sequences by Grammatical Inference



artistic view of old stone age Glyptodon by Heinrich Harder

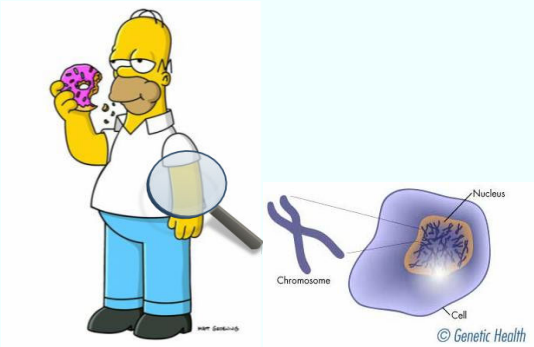
Focus on working tools and specificity of modelling biosequences

Outline

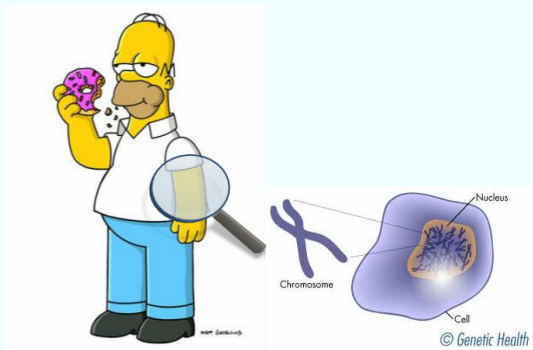
- 1 Molecular genetics
- 2 Modelling a set of conserved sequences
- 3 Inference of grammatical structure

- 1 Molecular genetics
 - Central “dogma”
 - Technology
 - Sequences annotation
- 2 Modelling a set of conserved sequences
- 3 Inference of grammatical structure

Genetic heredity



Genetic heredity



Genome

Organism's hereditary information

DNA

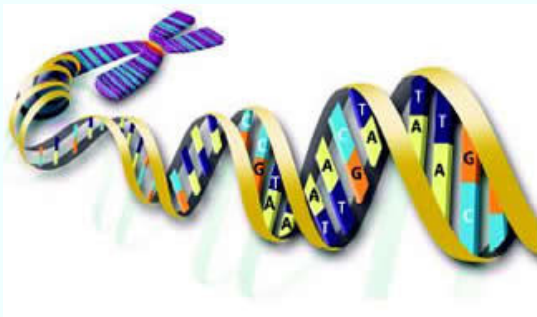
Genetic heredity



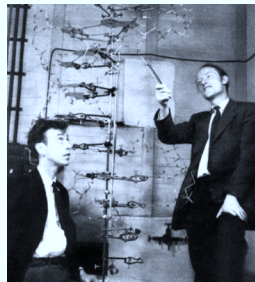
Genome

Organism's hereditary information

DNA structure



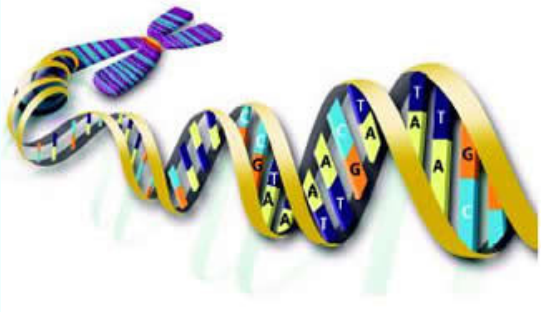
source: U.S. Department of Energy



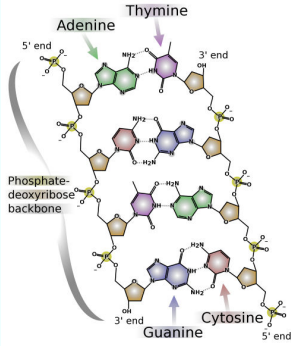
“We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.”

John Watson and Francis Crick in *Molecular Structure of Nucleic Acids. A structure for deoxyribose nucleic acid*. Nature No. 4356, April 25, 1953

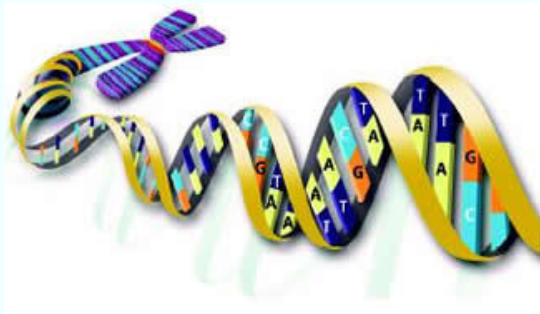
DNA sequence



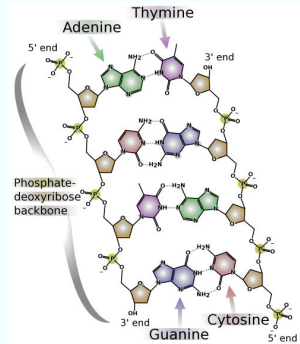
source: U.S. Department of Energy



DNA sequence



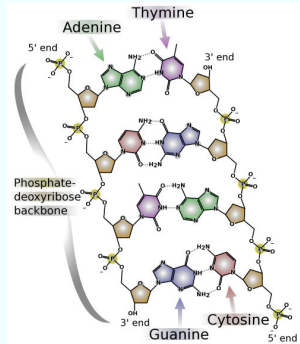
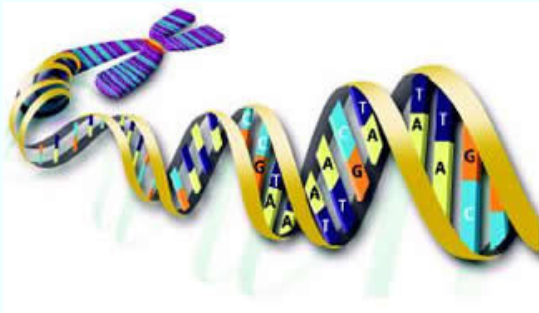
source: U.S. Department of Energy



DNA information

Sequence over a 4 letter alphabet {A,C,G,T} (nucleotide bases)

DNA sequence



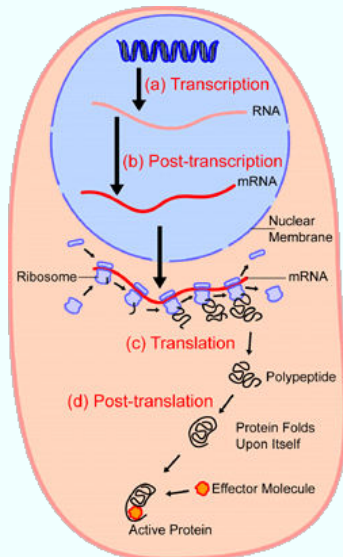
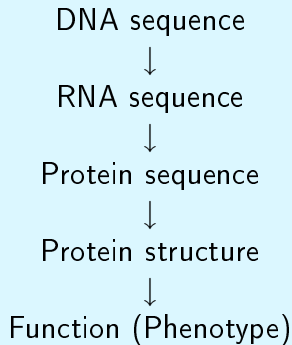
source: U.S. Department of Energy

```
ACTGGATCACAGGTCTATCACCTATTAACCACTCACGGGAGCT
CTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCACGCGAT
AGCATTGCGAGACGCTG...
```

DNA information

Sequence over a 4 letter alphabet {A,C,G,T} (nucleotide bases)

From DNA to Function



see also http://www.youtube.com/watch?v=ZjRCmUO_dhY

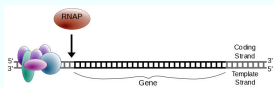
DNA to RNA

Transcription

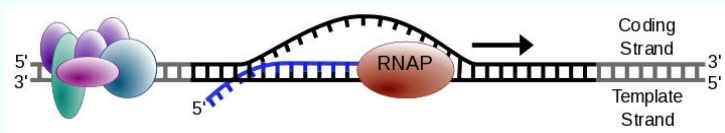
(double strand) DNA sequence {A, T, C, G}



(simple) RNA sequence {A, U, C, G}



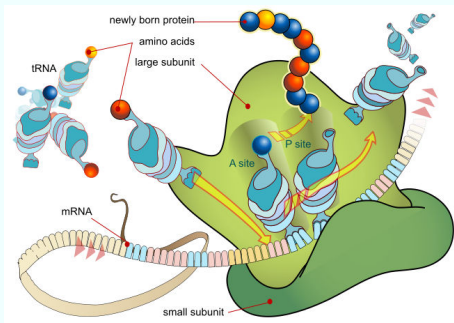
RNA-Polymerase



RNA to Protein

Translation

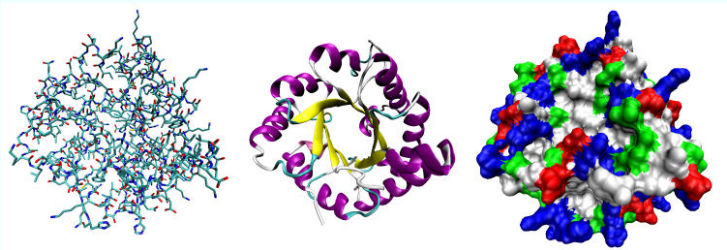
RNA sequence {A, U, C, G}
↓
Protein sequence {20 amino acids}



Ribosome

		Second letter								
		U	C	A	G					
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U	
	UUC		UCC		UAC		UGC		C	
	UUA		UCA		UAA		UGA		Stop	A
	UUG		UCG		UAG		UGG		Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U	
	CUC		CCC		CAC		CGC		C	
	CUA		CCA		CAA		CGA		A	
	CUG		CCG		CAG		CGG		G	
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U	
	AUC		ACC		AAC		AGC		C	
	AUA		ACA		AAA		AGA		A	
	AUG		ACG		AAG		AGG		G	
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U	
	GUC		GCC		GAC		GGC		C	
	GUA		GCA		GAA		GGA		A	
	GUG		GCG		GAG		GGG		G	

Proteins



Homer Simpson, 70kg

- Water: 42kg
- Protein: 12kg
- Fat: 12kg
- Bone Minerals: 3.5kg
- Carbohydrate: 0.5kg



Protein functions:

Structure, Transport, Antibodies,
Muscles, Enzymes, Hormones, Brain

...

Reading the Book of Life

Genome Sequencing

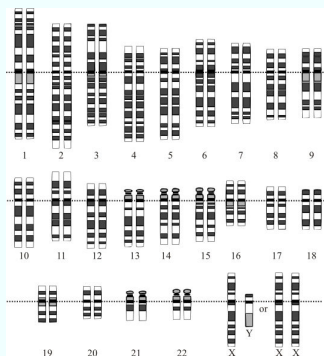


First draft of the human genome (2001)



First draft of the human genome (2001)

Consortium 20 labs, 6 countries



3.2 billion bp, ~ 25000 genes (previous estimates was ~ 100,000, function unknown for 50% of predicted genes) covering about 3% genome length, >50% repeat sequences. . .

Genome sequencing projects

Genome Project Statistic:

Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	794	580	568	1942
Archaea	75	5	33	113
Bacteria	720	575	535	1830
Eukaryotes	38	249	270	557
Animals	5	109	119	233
Mammals	2	38	41	81
Birds		2	13	15
Fishes		13	13	26
Insects	2	25	19	46
Flatworms		2	3	5
Roundworms	1	13	11	25
Amphibians		1		1
Reptiles		1		1
Other animals		16	22	38
Plants	6	23	74	103
Land plants	3	19	69	91
Green Algae	3	4	4	11
Fungi	18	82	36	136
Ascomycetes	15	63	25	103
Basidiomycetes	1	12	8	21
Other fungi	2	7	3	12
Protists	9	33	37	79
Apicomplexans	5	10	4	19
Kinetoplasts	3	2	3	8
Other protists	1	20	30	51
total:	832	829	838	2499

Revised: Sep 07, 2010

New generation sequencing



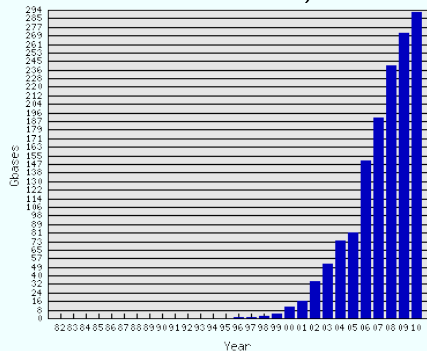
Next-generation sequencing machines at the Wellcome Trust Sanger Institute. Wellcome Library, London.

- Human Genome Project: ~ 10 years, $< \$3$ billion USD
- 2008: 2 weeks, \$60,000 USD
- Archon X Prize for Genomics:
 - < 2013 : 100 human genomes in ten days for \$10,000 USD each
- 1000\$ USD personal genome. . .

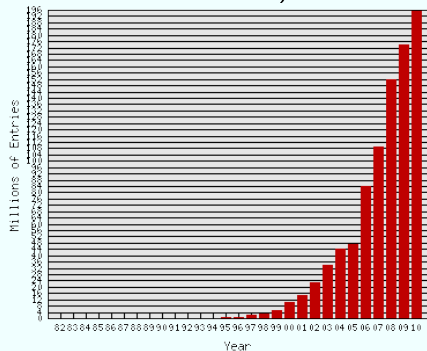
EMBL Database Growth



Total nucleotides (current
292,065,553,621)



Number of entries (current
195,230,921)



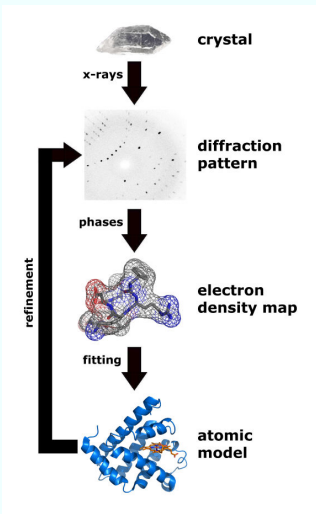
source: <http://www.ebi.ac.uk/embl/Services/DBStats/> Sep 6, 2010

Protein structure determination



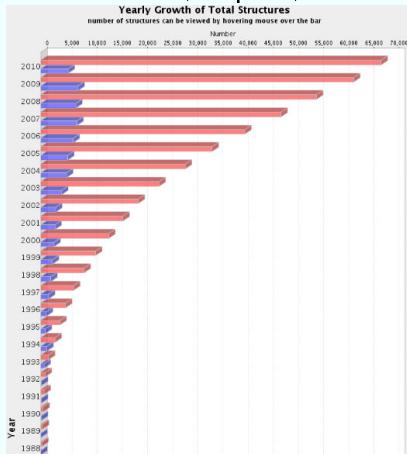
Structures in RCSP Protein Data Bank
www.pdb.org

- 90% by X-ray crystallography
9% Nuclear Magnetic Resonance
- Proteins: 62773, Nucleic Acids: 2172,
Protein/NA: 2811 Others: 38
Total: 67794
- Mainly globular proteins



PDB database Growth

67,794 structures, Sep 07, 2010 [▶ embl](#)

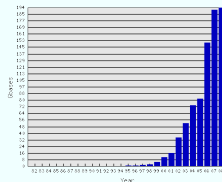


<http://www.rcsb.org/pdb/statistics/holdings.do>

<http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total>

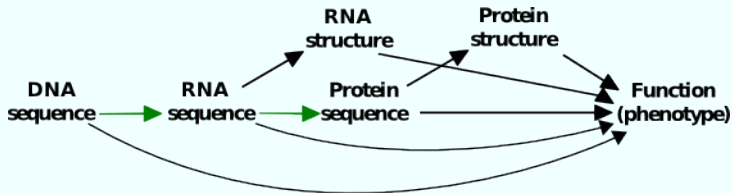
Sequences annotation

High throughput production of raw sequences



Problem

Function(s) of these sequences ?



How to predict the function of a sequence ?

2 main approaches:

1. Find a similar sequence

needs:

- a database of annotated sequences
GenBank (NCBI), EMBL Nucleotide Sequence Database, DNA Data Bank of Japan (DDBJ), ...
- a tool returning similar sequences
BLAST ...

2. Find a model accepting the sequence

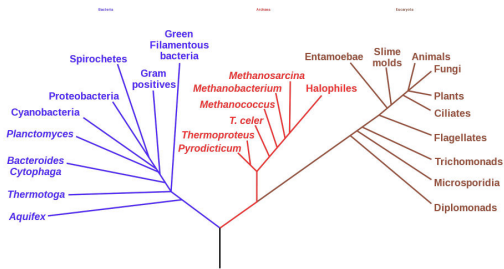
needs:

- a database of model for each “function” of interest
TRANSFAC, JASPAR, PROSITE, INTERPRO
- a tool returning models recognizing the sequence
ScanProsite, InterProScan, ...

Conservation



Phylogenetic Tree of Life

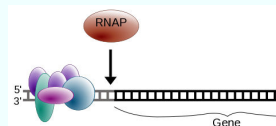
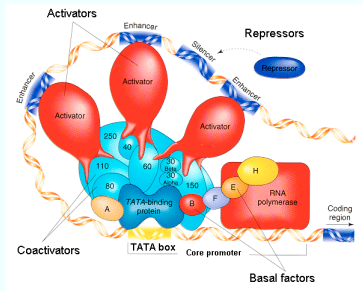


Key idea

Evolution + Natural Selection \Rightarrow Conservation is meaningful

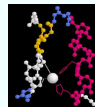
Conserved sequence

- DNA: TATA-box (5'-TATAAA-3')



- Protein: zinc finger

P41696	C	P	Y	C	H	R	L	F	S	Q	A	T	H	L	E	V	H	V	R	S	H	I	G	Y	K	P	F	V	C	D
D6W2H2	C	P	Y	C	H	R	L	F	S	Q	A	T	H	L	E	V	H	V	R	S	H	I	G	Y	K	P	F	V	C	D
A6ZNW1	C	P	Y	C	H	R	L	F	S	Q	A	T	H	L	E	V	H	V	R	S	H	I	G	Y	K	P	F	V	C	D
C7GMM3	C	P	Y	C	H	R	L	F	S	Q	A	T	H	L	E	V	H	V	R	S	H	I	G	Y	K	P	F	V	C	D
C5DS E2	C	P	Y	C	H	R	F	F	T	Q	S	T	H	L	E	V	H	V	R	S	H	I	G	Y	K	P	F	L	C	E
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*



From identity to similarity

Blocks

```
P41696 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD  
D6W2H2 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD  
A6ZMW1 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD  
C7GMM3 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD  
C5DSE2 CPYCHRFFTSATHLEVHVRSHIGYKPFVCD  
*****:*:*:*****:*****:*
```

From identity to similarity

Blocks

```
P41696 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
D6W2H2 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
A6ZNW1 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
C7GMM3 CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
C5DSE2 CPYCHRFFTSATHLEVHVRSHIGYKPFVCE
*****:*:*:*****:*****:*
```

BLOSUM62 substitution matrix Frequently observed substitutions receive positive scores and seldom observed substitutions are given negative scores (log odds ratio)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4							I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

From identity to similarity

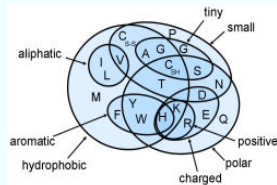
Blocks

```

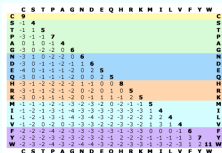
P41696  CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
D6W2H2  CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
A6ZNW1  CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
C7GMM3  CPYCHRLFSQATHLEVHVRSHIGYKPFVCD
C5DSE2  CPYCHRFFTSQATHLEVHVRSHIGYKPFVCE
*****:*:*:*****:*****:
    
```

BLOSUM62 substitution matrix Frequently observed substitutions receive positive scores and seldom observed substitutions are given negative scores (log odds ratio)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4							I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-2	-2	-3	-3	-2	1	3	1	4					V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	-1	6				F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-1	-1	-1	3	7				Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W



Aligning two proteins



Smith&Waterman, Blast, ...

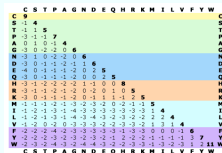
```
AAB24882      TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881      -----YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                ****: .***: * *:*** * :****.:* *****.

AAB24882      PSHLQYHERITHIGEKPYECHQCGQAFKCSLLQRHKRIHTGEKPYE-CNQCGKAFAQ- 116
AAB24881      HSHLQCHKRITHIGEKPYECNQCGKAFSQHGLLQRHKRIHTGEKPYMNVINMVKPLHNS 98
                **** * :*****:***:**: .*****: * : * : :
```

Alignment score:

$$\sum_p (\text{Blosum}(a_{1,p}, a_{2,p})) + f(\text{gap})$$

Aligning two proteins



Smith&Waterman, Blast, ...

```
AAB24882      TYHMCQFHCRVYNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCCKAFPT 60
AAB24881      -----YECNQCGKAFQAHSCLKCHYRTHIGKPYECNQCGKAFSK 40
                ****: .***: * *:*** * :****.:* *****.

AAB24882      PSHLQYHERITHIGKPYECHQCGQAFKKCSLLQRHKRTHIGKPYE-CNQCGKAFQA- 116
AAB24881      HSHLQCHKRTHIGKPYECNQCGKAFSQHGLLQRHKRTHIGKPYMNVINMVKPLHNS 98
                **** *:*****:***:**.: .*****: *.::
```

Alignment score:

$$\sum_p (\text{Blosum}(a_{1,p}, a_{2,p})) + f(\text{gap})$$

+ significativity (E-value) → tool for approach 1.

Multiple sequence alignment

ClustalW2, ...

```
ABC3G_LAGLA/285-305    Cfs..CaekVaeflqenpHvnL..H
ABRU_DROME/546-567    Cpk..CgkiYrsahtlrthledk.H
ACE1_TRIRE/402-424    CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHPO/445-467    ClyngCnkrIarkynvesHigt..H
ACE2_SCHPO/475-495    Cdl..CkagFvrhhdLkrHlri..H
ACE2_YEAST/605-627    ClypnCnkVfkrrynirshigt..H
ACE2_YEAST/635-657    CdfpgCtkaFvrnhdLirHkis..H
ADNP2_HUMAN/772-793   Clf..CpctFhdikglseHsrnr.H
ADNP2_HUMAN/877-899   Cpf..CfpgFvtteayelHlkerhH
ADNP2_MOUSE/802-823   Clf..CpctFhdvrglveHsrtk.H
```

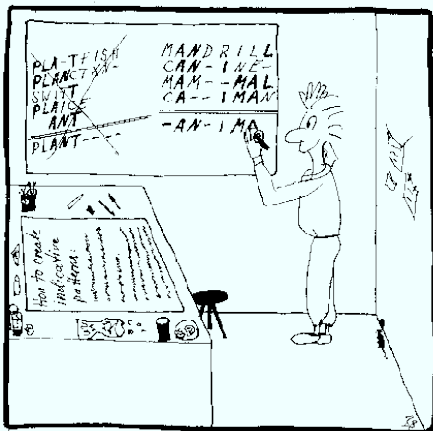
Multiple sequence alignment

ClustalW2, ...

```
ABC3G_LAGLA/285-305    Cfs..CaekVaeflqenpHvnL..H
ABRU_DROME/546-567    Cpk..CgkiYrsahtlrthledk.H
ACE1_TRIRE/402-424    CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHPO/445-467    ClyngCnkrIarkynvesHiqt..H
ACE2_SCHPO/475-495    Cdl..CkagFvrhhdLkrHlri..H
ACE2_YEAST/605-627    ClypnCnkvFkrrynirsHiqt..H
ACE2_YEAST/635-657    CdfpgCtkaFvrnhdlirHkis..H
ADNP2_HUMAN/772-793   Clf..CpctFhdikglseHsrnr.H
ADNP2_HUMAN/877-899   Cpf..CfpgFvtteayelHlkerhH
ADNP2_MOUSE/802-823   Clf..CpctFhdvrglveHsrtk.H
```

Can be used to build models of the sequence family for approach 2.
(exact models and their probabilistic mate)

How we develop Prosite patterns!



Brigitte Boeckmann / 1995

1 Molecular genetics

2 Modelling a set of conserved sequences

- Patterns
- Allowing insertion
- Allowing insertions and deletions
- With covariations

3 Inference of grammatical structure

Modelling a set of aligned sequences

Task: Model a family of (bio)sequences

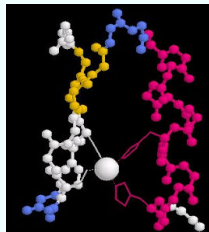
(same function, structure, localization, expression, ...)

Goals: Understand, Predict, Generate

Assumption: Characterization of family from sequence only

Available: A **representative sample** of the family and a **multiple alignment** of its conserved region (from bio experiments or a computational alignment tool)

```
ABC3G_LAGLA/285-305    Cfs..CaekVaeflqenpHvnL..H
ABRU_DROME/546-567    Cpk..CgkiYrsahtLrtHledk.H
ACE1_TRIRE/402-424    CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHP0/445-467    ClyngCnkrIarkynvesHigt..H
ACE2_SCHP0/475-495    CdL..CkagFvrhhdLkrHlri..H
ACE2_YEAST/605-627    ClypnCnkvFkrrynirsHigt..H
ACE2_YEAST/635-657    CdfpgCtkaFvrnhdlirHkis..H
ADNP2_HUMAN/772-793  Clf..CpctFhdikglseHsrnr.H
ADNP2_HUMAN/877-899  Cpf..CfpgFvtteayelHlkerhH
ADNP2_MOUSE/802-823  Clf..CpctFhdvrglveHsrtk.H
```



Modelling a set of aligned sequences

Task: Model a family of (bio)sequences

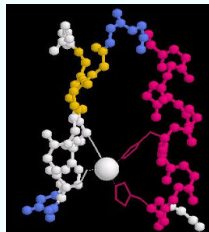
(same function, structure, localization, expression, ...)

Goals: Understand, Predict, Generate

Assumption: Characterization of family from sequence only

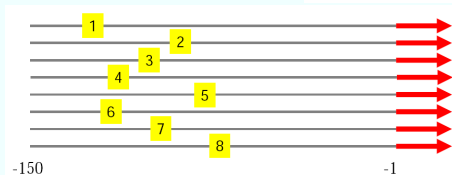
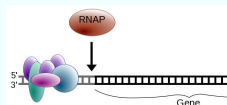
Available: A **representative sample** of the family and a **multiple alignment** of its conserved region (from bio experiments or a computational alignment tool)

```
ABC3G_LAGLA/285-305    Cfs..CaekVaeflqenpHvnL..H
ABRU_DROME/546-567    Cpk..CgkiYrsahtLrtHledk.H
ACE1_TRIRE/402-424    CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHP0/445-467    ClyngCnkrIarkynvesHigt..H
ACE2_SCHP0/475-495    Cdl..CkagFvrhhdLkrHlri..H
ACE2_YEAST/605-627    ClypnCnkvFkrrynirsHigt..H
ACE2_YEAST/635-657    CdfpgCtkaFvrnhdlirHkis..H
ADNP2_HUMAN/772-793  Clf..CpctFhdikglseHsrnr.H
ADNP2_HUMAN/877-899  Cpf..CfpgFvtteayelHlkerhH
ADNP2_MOUSE/802-823  Clf..CpctFhdvrglveHsrtk.H
```



How to model this conservation ?

Example: Modelling binding sites



Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

example from Maximilian Haussler, INRIA tech. report RR-5714



Consensus motif

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

B R M C W W W H R W G G B M

Consensus motif (sequence), IUPAC code

[TCG] [ATG] [AC] C [AT] [AT] [AT] [ATC] [ATG] [AT] G G [TCG] [AC]

Consensus motif

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

B R M C W W W H R W G G B M

Consensus motif (sequence), IUPAC code

[TCG] [ATG] [AC] C [AT] [AT] [AT] [ATC] [ATG] [AT] G G [TCG] [AC]



GTACATTTGAAGTA

vs

TAACATAAATGGGA

?

Consensus motif

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

B R M C W A W H R W G G B M

Consensus motif (sequence), IUPAC code

[TCG] [ATG] [AC] C [AT] A [AT] [ATC] [ATG] [AT] G G [TCG] [AC]



GTACATTTGAAGTA

vs

TAAC TATAATGGGA

?

Consensus motif

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

B R M C W A W H R W G G B M

Consensus motif (sequence), IUPAC code

[TCG] [ATG] [AC] C [AT] A [AT] [ATC] [ATG] [AT] G G [TCG] [AC]

+ allow a limited number of errors



GTACATTTGAAGTA vs TAACTATAATGGGA ?

Consensus sequence

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

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

Consensus sequence
+ allow a limited number of errors

Consensus sequence

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

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

Consensus sequence
+ allow a limited number of errors



Some positions mutate more easily...
Some substitutions occur more likely...

Position frequency matrix

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Position frequency matrix

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

→

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position frequency matrix (PFM)

Position frequency matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position frequency matrix (PFM)

Position frequency matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position frequency matrix (PFM)

Exercises

- How to use the PFM to assign probabilities to sequences ?
 - $P(\text{GTACATTTGAAGTA}) = ?$
 - $P(\text{TAACTATAATGGGA}) = ?$
 - $P(\text{AAACTATAATGGGA}) = ?$
- Significativity of these probabilities ?
- What if the nucleotides composition is biased ?

Position probabilities and weights at each position/column

Notation: $O(x)$ frequency/count of nucleotide x in the column \vec{O} of the PFM
Maximum likelihood estimation of probability of x :

$$\hat{P}(x|\vec{O}) = \frac{O(x)}{\sum_{x'} O(x')}$$

or rather use pseudo-counts:

$$\hat{P}(x|\vec{O}) = \frac{O(x) + 1}{\sum_{x'} (O(x') + 1)}$$

Weight of x (*Log odds*)

$$W(x|\vec{O}) = \log \frac{\hat{P}(x|\vec{O})}{P_0(x)}$$

where $P_0(x)$: background probability of x

Position weight matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position frequency matrix

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	-0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	-0.66	-1.93	-1.93	1.07	-0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

Position weight matrix (PWM)

(with $p(A) = p(T) = p(G) = p(C) = 1/4$)

Scoring a site

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	-0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	-0.66	-1.93	-1.93	1.07	-0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

Position weight matrix

0.45	-0.66	0.79	1.68	0.45	-0.66	0.79	0.45	-0.66	0.79	0.00	1.68	-0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

TRANSFAC

AC M00231

XX VSMEP2_02

DE 26.01.1996 (created); ewi.

DE 26.01.1996 (updated); ewi.

DE Copyright (C), Biobase GmbH.

CO

XX MEF-2

NA

XX

DE myogenic MADS factor MEF-2

XX

DE T09505 MEF-2; Species: mouse, Mus musculus.

DE T01005 MEF-2; Species: human, Homo sapiens.

XX

PO	A	C	G	T	N
01	5	28	25	42	N
02	16	32	31	21	N
03	18	36	27	19	N
04	19	25	33	23	N
05	22	12	43	23	N
06	33	9	21	37	N
07	20	4	43	33	N
08	3	85	3	9	N
09	3	8	0	89	N
10	85	0	0	15	N
11	57	0	0	41	N
12	91	0	1	8	N
13	96	0	1	4	N
14	93	0	1	6	N
15	0	0	0	100	N
16	100	0	0	0	N
17	9	0	90	1	N
18	34	46	11	9	N
19	36	28	8	28	N
20	20	37	15	28	N
21	30	34	13	23	N
22	23	23	22	32	N

DE 194 selected sequences binding MEF-2 activity of skeletal muscle

XX

DE sequences binding to protein that reacts with anti-MEF-2A antibodies were isolated by 9 cycles of selection and amplification from a pool of random 48-mers; nucleotide distribution given in percentages

XX

RN [1]

RX MEDLINE; 26007455.

RA Andres V., Cervera M., Mahdavi V.

RE Determination of the consensus binding site for MEF2 expressed in muscle and brain reveals tissue-specific sequence constraints

RT J. Biol. Chem. 270:23246-23249 (1995).

RL

XX

//

Information content

Relative entropy (Kullback-Leibler distance) of matrix wrt background distribution

- Information content of one column/position:

$$|C_{pos} = \sum_{\alpha \in \Sigma} f_{pos,\alpha} \log_2 \frac{f_{pos,\alpha}}{f_{\alpha}}$$

(ADN, 1/4: between 0 and 2 bits)

- Information content of a matrix:

$$|C_{matrix} = \sum_{pos=1}^{len} |C_{pos}$$

(ADN, 1/4: max = len × 2)

Most surprising set of conserved words

Information content

Examples

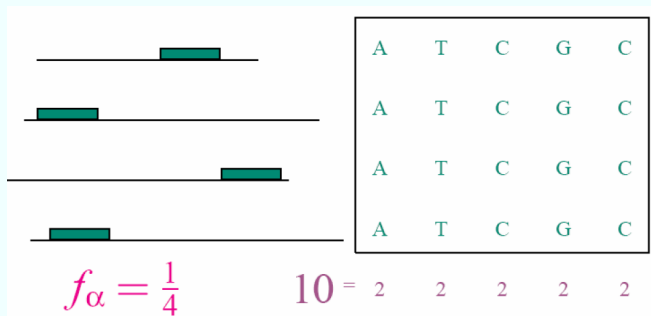
$$\sum_{pos=1}^{len} \sum_{\alpha \in \Sigma} f_{pos,\alpha} \log_2 \frac{f_{pos,\alpha}}{f_{\alpha}}$$



Information content

Examples

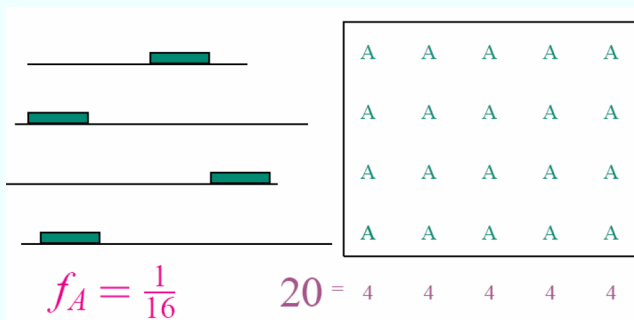
$$\sum_{pos=1}^{len} \sum_{\alpha \in \Sigma} f_{pos,\alpha} \log_2 \frac{f_{pos,\alpha}}{f_{\alpha}}$$



Information content

Examples

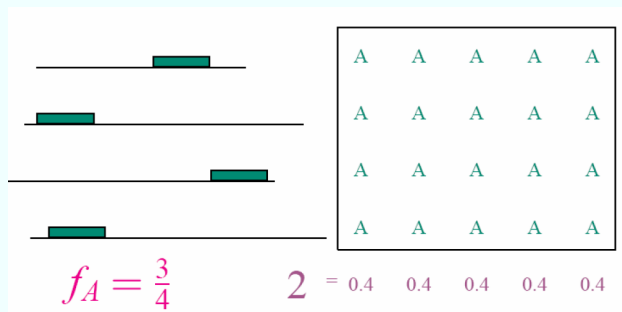
$$\sum_{pos=1}^{len} \sum_{\alpha \in \Sigma} f_{pos,\alpha} \log_2 \frac{f_{pos,\alpha}}{f_{\alpha}}$$



Information content

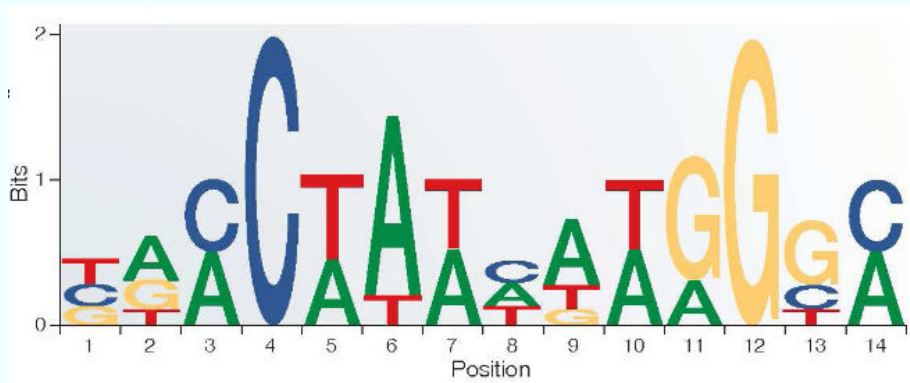
Examples

$$\sum_{pos=1}^{len} \sum_{\alpha \in \Sigma} f_{pos,\alpha} \log_2 \frac{f_{pos,\alpha}}{f_{\alpha}}$$



A graphical view

Sequence Logos



Motif discovery as local alignment

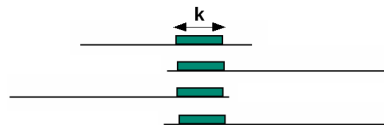
Input:

Data: A set of sequences

Parameter: k length of alignment

Output:

A set of word of length k , one in each sequence, s.t. the information content of the matrix is maximal

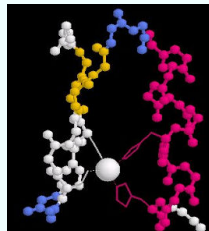


Heuristics:

- Expectation-Maximization: MEME, Bailey, 1995
- Gibbs Sampling: Lawrence et al, 1993, Thijs et al, 2001
- Greedy: (w)consensus, Hertz et al, 1999
- Projection: Buhler et al, 2000

Modelling insertions

```
ABC3G_LAGLA/285-305      Cfs..CaekVaeflqenpHvnl..H
ABRU_DROME/546-567      Cpk..CgkiYrsahtlrthledk.H
ACE1_TRIRE/402-424      CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHPO/445-467      ClyngCnkrIarkynvesHiqt..H
ACE2_SCHPO/475-495      Cdl..CkagFvrhhdldrHlri..H
ACE2_YEAST/605-627      ClypnCnkvFkrrynirsHiqt..H
ACE2_YEAST/635-657      CdfpgCtkaFvrnhdlirHkis..H
ADNP2_HUMAN/772-793     Clf..CpctFhdikglseHsrnr.H
ADNP2_HUMAN/877-899     Cpf..CfpgFvtteayelHlkerhH
ADNP2_MOUSE/802-823     Clf..CpctFhdvrglveHsrthk.H
```



- Prosite patterns
- Prosite profiles

Prosite's pattern syntax

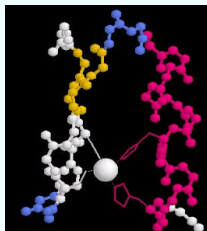
F	single-letter amino acid symbol: F
x	wildcard character: ny amino acid
[GST]	list: G, S, or T
{F}	exclusion list: not F
[ST] (2)	multiplier: two S or T in a row
x (0, 1)	range multiplier: between zero and times any amino acid
<G	N-terminal anchor: G at the beginning of the sequence
G>	C-terminal anchor: G at the end of the sequence
-	pattern element separator

Examples: [AC] -x-V-x (4) - {ED}
 <A-x- [ST] (2) -x (0, 1) -V

Exact models: Prosite patterns

ABC3G_LAGLA/285-305
ABRU_DROME/546-567
ACE1_TRIRE/402-424
ACE2_SCHPO/445-467
ACE2_SCHPO/475-495
ACE2_YEAST/605-627
ACE2_YEAST/635-657
ADNP2_HUMAN/772-793
ADNP2_HUMAN/877-899
ADNP2_MOUSE/802-823

Cfs..CaekVaeflqenpHvnL..H
Cpk..CgkiYrsahtlrthledk.H
CrepGctkeFkrpcdltkHekt..H
ClyngCnkrIarkynvesHiqt..H
CdL..CkagFvrhhdLkrHlri..H
ClypnCnkvFkrrynirsHiqt..H
CdfpgCtkaFvrnhdlirHkis..H
Clf..CpctFhdikglseHsrnr.H
Cpf..CfpgFvtteayelHlkerhH
Clf..CpctFhdvrglveHsrtk.H



Prosite's C2H2 Pattern:

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

 [Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#)

Entry: **PS00028**

General information about the entry

Entry name	ZINC_FINGER_C2H2_1
Accession number	PS00028
Entry type	PATTERN
Date	APR-1990 (CREATED); JUN-1994 (DATA UPDATE); DEC-2007 (INFO UPDATE).
PROSITE Documentation	PDOC00028

Name and characterization of the entry

Description	Zinc finger C2H2 type domain signature.
Pattern	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.

Numerical results

- UniProtKB/Swiss-Prot release number: **54.7**, total number of sequence entries in that release: **333445**.
- Total number of hits in UniProtKB/Swiss-Prot: **11060 hits in 1794 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **10810 hits in 1609 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **26 hits in 12 different sequences**
- Number of false hits (on unrelated proteins): **224 hits in 173 different sequences**
- Number of known missed hits: **60**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **97.97 %**
- Recall (true hits / (true hits + false negatives)): **99.45 %**

Probabilistic version: Prosite's Profile



Entry: PS50157

[Home](#) [ScanProsite](#) [ProRules](#) [Documents](#) [Downloads](#) [Links](#)

General information about the entry

Entry name	ZINC_FINGER_C2H2_2
Accession number	PS50157
Entry type	MATRIX
Date	DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); DEC-2007 (INFO UPDATE).
PROSITE	
Documentation	PF0000028
Associated ProRule	PRU00042

Name and characterization of the entry

Description Zinc finger C2H2 type domain profile.

```

/GENERAL_NAME=ALPHABETICALLY_SORTED_ZINC_FINGER_C2H2_2; LENGTH=48;
/REGIONS=REGIONS=PROSITE; M1=1; M2=4;
/REGULATION_FACTOR=FUNCTIONAL_LENGTH; R1=0.6489; R2=0.6287318; WEIGHT=LogR;
/COMP_SIZE=LENGTH; SCORE=1; IL2=SCORE*5; MODEL=PS50157;
/COMP_SIZE=LENGTH; IL2=SCORE*5; MODEL=PS50157;
/DEFAULT: D=-28; I=-28; M1=-58; M2=-50; M3=-105; M4=-105; M5=-105; M6=-105;

```

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	T	V	Y	X	
/I:	M1=0; M2=-185; M3=-105;																				
/M:	PS50157; M1=-23; -21; -24; -21; -21; 39; -28; 10; -2; -17; 2; 0; -17; -28; -19; -11; -18; -10; -7; 12; -48; -23;																				
/M:	PS50157; M2=-4; -5; -23; -5; 6; -28; -18; -9; -14; 7; -15; -7; -5; -9; 3; 2; -2; -2; -9; -25; -12; 4;																				
/M:	PS50157; M3=-18; -20; 118; -36; -30; -28; -30; -30; -38; -30; -28; -26; -26; -40; -30; -38; -18; -10; -18; -56; -30; -38;																				
/M:	PS50157; M4=-5; 3; -24; 3; 6; -22; -11; -9; -28; 1; -21; -16; 4; -3; 1; -3; 5; 2; -18; -29; 15; 3;																				
/I:	I=12; M5=0; M6=-30; M7=0; M8=-30;																				
/M:	PS50157; M9=-2; -24; 1; 14; -18; -17; -4; -15; -1; -11; -8; -5; -12; 4; -5; -5; 8; -12; -24; -9; 8;																				
/M:	PS50157; M10=-18; -20; 119; -18; -30; -28; -30; -38; -30; -28; -28; -28; -40; -30; -38; -18; -10; -18; -56; -29; -18;																				
/M:	PS50157; M11=-1; -1; -24; -1; -7; -24; 16; -11; -33; -11; -27; -18; 4; -15; -10; -12; 1; -13; -27; -24; -3; -9;																				
/M:	PS50157; M12=-18; -2; -28; -3; 8; -23; -15; -7; -24; 16; -24; -8; -1; -12; 10; 27; -9; -5; -11; -19; -8; 8;																				
/M:	PS50157; M13=1; -7; -9; -11; -7; -17; -7; -14; -18; -8; -18; -11; -4; -15; -6; -5; 8; 4; -7; -15; -15; -7;																				
/M:	PS50157; M14=-19; -29; -19; -37; -28; 7; -29; -17; 8; -28; 9; 6; -28; -18; -36; -19; -19; -9; -1; 9; 31; -24;																				
/M:	PS50157; M15=1; -5; -17; -9; -6; -18; -11; -10; -14; -5; -14; -10; 9; -12; -4; 9; 8; 7; -4; -27; -32; -4;																				
/M:	PS50157; M16=-18; -3; -28; -4; 0; -18; -17; 2; -19; 3; -18; -8; 0; -17; 8; 9; -1; -3; -17; -19; -5; 3;																				
/M:	PS50157; M17=-4; -6; -22; -4; 0; -19; -13; -5; -18; 7; -18; -9; 1; -18; 2; 8; 2; -2; -34; -25; -11; 8;																				
/M:	PS50157; M18=2; -1; -18; -1; -1; -18; -4; -9; -10; -7; -22; -14; 4; -12; -2; 7; 18; 7; -13; -29; -12; 0;																				
/M:	PS50157; M19=-5; -5; -28; 3; 0; -18; -10; 8; -18; -5; -18; -11; 9; -18; 1; -4; 4; -1; -17; -27; -7; -2;																				
/M:	PS50157; M20=-11; -29; -29; -36; -20; 12; -29; -19; 17; -27; 45; 18; -28; -20; -20; -18; -28; -10; 9; -18; 2; -24;																				
/M:	PS50157; M21=-6; -6; -22; -16; -8; -18; -20; -8; -7; 2; -9; -3; -2; -18; 6; 3; -1; 0; -8; -24; -24; -5;																				
/M:	PS50157; M22=-6; -6; -23; -7; 0; -19; -17; -7; -14; 7; -13; -6; -2; -18; 4; 12; -3; -3; -18; -24; -10; 8;																				
/M:	PS50157; M23=-2; 0; -9; 8; 0; -28; -20; 99; -38; -15; -28; 6; 18; -28; 18; 8; -18; -20; -33; -16; 20; 8;																				
/M:	PS50157; M24=-18; -10; -25; -12; 1; -18; -22; -2; -4; 1; -5; 6; -9; -17; 13; 5; -9; -8; -9; -19; -4; 8;																				
/M:	PS50157; M25=-11; -8; -28; -9; 0; -19; -19; -4; -21; 20; -14; -4; -2; -17; 6; 35; -6; -7; -14; -21; -9; 8;																				
/I:	I=12; M26=0; M27=-29; M28=-29;																				
/M:	PS50157; M29=-3; -16; -17; -21; -17; -4; -25; -29; 13; -15; 2; 3; -12; -18; -14; -14; -2; 9; 13; -25; -7; -17;																				
/M:	PS50157; M30=-24; 0; -19; 8; 0; -24; -26; 97; -38; -10; -28; 6; 18; -24; 16; 8; -18; -20; -34; -30; 10; 8;																				
/M:	PS50157; M31=2; -2; -13; -18; 6; -14; -15; -15; -12; -7; -13; 9; 0; -12; -6; 8; 14; 25; -3; -29; -12; -8;																				
/M:	PS50157; M32=-3; -4; -27; -4; -8; -27; 34; -34; -32; -9; -25; -14; 2; -15; -10; -9; 8; -13; -24; -23; -23; -18;																				
/M:	PS50157; M33=-9; 8; -27; 12; -15; -24; -17; -3; -24; 7; -18; -15; -1; -4; 12; 8; -2; -4; -22; -28; -18; -22;																				
/M:	PS50157; M34=-11; -1; -28; -2; 6; -25; -17; -8; -27; 32; -25; -10; 1; -11; 8; 28; -7; -8; -19; -22; -11; 1;																				
/M:	PS50157; M35=-7; -14; -32; -9; -1; -24; -17; -15; -18; -7; -23; -15; -13; 51; -7; -13; -6; -6; -23; -28; -22; -7;																				
/I:	I=148;																				

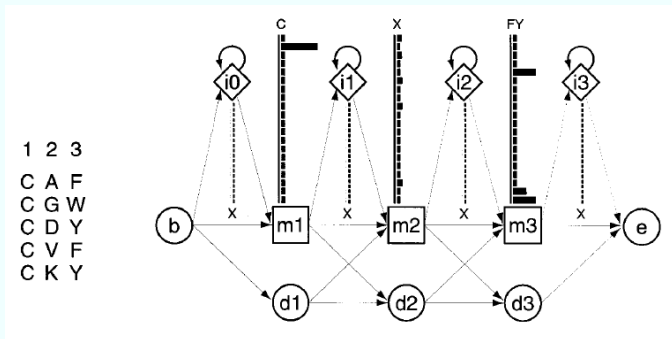
Numerical results

- UniProtKB/Swiss-Prot release number: **54.7**, total number of sequence entries in that release: **333445**
- Total number of hits in UniProtKB/Swiss-Prot: **11272 hits in 1555 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **1257 hits in 1540 different sequences**



Profile HMM

Simple left-to-right topology: PWM + insertions and deletion state



m: Match state, i: Insertion state, d: Deletion (silent) state

picture from Sean Eddy

- Each match state has its own emission probability distribution
- Handles deletions and insertions, enabling different costs to enter, leave or continue in a mode

Parameter estimation

- From aligned sequences
- (Simplified) Baum-Welsh, ...

Simple topology but still an important number of free parameters.
One needs big training sample. Prone to overspecialization.

Example: SH3 domain

source: An Introduction to Hidden Markov Models for Biological Sequences, Anders Krogh

```
GGWWRGGdy.ggtkkqLWFFPSSNTYYV
IGWLNNGyn.efttgerLDGDFPSTYV
PNWWEQGqql..nrrrIGIFPSSNYV
DEWWEQAqrrr..ddeqqrigIIVPSSK--
GEWWEKAqqs...tggqqGFIIPFNFYV
GDWWEAELArsl..skgqrGKYIPSNYV
GDWWEAELArsl..skgqrGKYIPSNYV
-DWWEAELArsl..skgqrGKYIPSNYV
GDWWEAELArsl..skgqrGKYIPSNYV
GEWWEKArsllatrnseGKYIPSNYV
GDWWEKArsllatrnseGKYIPSNYV
GEWWEKArsllatrnseGKYIPSNYV
GEWWEKArsllatrnseGKYIPSNYV
SDWWEAELArsl..skgqrGKYIPSNYV
LPWWEAELArsl..skgqrGKYIPSNYV
RDWWEAELArsl..skgqrGKYIPSNYV
EHWWEKVKkd.alngnveGKYIPSNYV
IHWWEKVKkd.alngnveGKYIPSNYV
KDWWEKVKkd.alngnveGKYIPSNYV
VGEWMPGln.e.rtrqrGDFPSTYV
PDWWEGGei..ngnrrkGVFFPASYYV
ENWWEGGei..ngnrrkGVFFPASYYV
EEWLEGec..kkgkvGIFPKVFFV
GGWWEKGDy.g.triqQYFFPSNYV
DGWWEKGDy.g.triqQYFFPSNYV
OGWWEKGDy.g.triqQYFFPSNYV
GRWWEKArrr..anggetGIIIPSNYV
GGWWEKGDy.g.triqQYFFPSNYV
GDWWEAELArsl..skgqrGKYIPSNYV
NDWWEAELArsl..skgqrGKYIPSNYV
```

Figure 4.4: An alignment of 30 short amino acid sequences chopped out of a alignment of the SH3 domain. The shaded areas are the most conserved and were chosen to be represented by the main states in the HMM. The unshaded area with

Example: SH3 domain

source: An Introduction to Hidden Markov Models for Biological Sequences, Anders Krogh

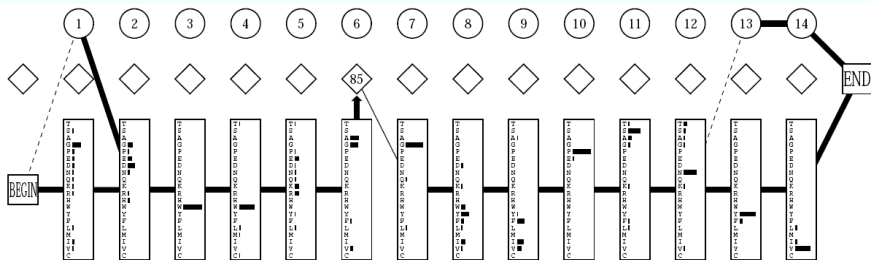


Figure 4.5: A profile HMM made from the alignment shown in Fig. 4.4. Transition lines with no arrow head are transitions from left to right. Transitions with probability zero are not shown, and those with very small probability are shown as dashed lines. Transitions from an insert state to itself is not shown; instead the probability times 100 is shown in the diamond. The numbers in the circular delete states are just position numbers. (This figure and Fig. 4.6 were generated by a program in the SAM package of programs.)

Example: SH3 domain

source: An Introduction to Hidden Markov Models for Biological Sequences,
Anders Krogh

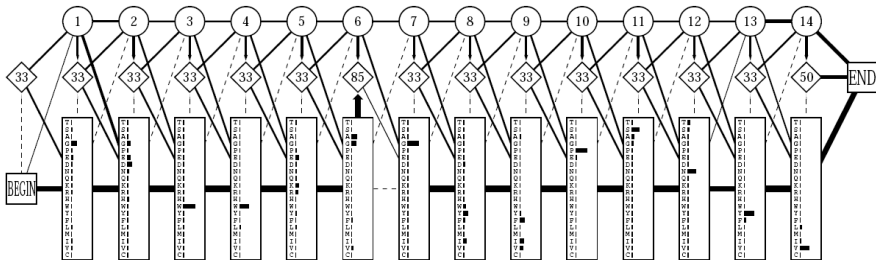


Figure 4.6: Model obtained in the same way as Fig. 4.5, but using a pseudocount of one.

Example: 2crd Charybdotoxin

source: tutorial ISMB, SAM-T98

With maximum likelihood weighting scheme:

```
2crd  XFTNVSCTTSKECW SVCQRLHNTSRGKCMNKKCRCYS  
1cmr  -----CTTSKECW SVCQRLHNTSKGWCDHRGCICES  
2bmt  XFTNVSCSASSQCWPVCKK LFGTYRGKCMNSKRCRCYS
```



Example: 2crd Charybdotoxin

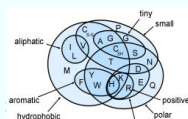
source: tutorial ISMB, SAM-T98

With maximum likelihood weighting scheme:

```
2crd  XFTNVSCTTSKECWVCQRLHNTSRGKCMNKKCRCYS
1cmr  -----CTTSKECWVCQRLHNTSKGWCDHRGCICES
2bmt  XFTNVSCSASSQCWPVCKKLFQTYRGKCMNSKRCRCYS
```



Can't we replace a L by a I ?



Adding pseudocounts

source: tutorial ISMB, SAM-T98

Add-one pseudocount estimation (Laplace rule):

$$\hat{P}(a|\vec{O}) = \frac{O(a) + 1}{\sum_{\text{amino acids } j} O(j) + 1}$$

Adding pseudocounts

source: tutorial ISMB, SAM-T98

Add-one pseudocount estimation (Laplace rule):

$$\hat{P}(a|\vec{O}) = \frac{O(a) + 1}{\sum_{\text{amino acids } j} O(j) + 1}$$

```
2crd  XFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
1cmr  -----CTTSKECWSVCQRLHNTSKGWCDHRGCICES
2bmt  XFTNVSCSASSQCWPVCKKLFGTYRGKCMNSKRCRCYS
```



Adding pseudocounts

source: tutorial ISMB, SAM-T98

Add-one pseudocount estimation (Laplace rule):

$$\hat{P}(a|\vec{O}) = \frac{O(a) + 1}{\sum_{\text{amino acids } j} O(j) + 1}$$

More sophisticated pseudocounts:

$$\hat{P}(a|\vec{O}) = \frac{O(a) + A P_0(a)}{\sum_{a'} O(a') + A}$$

where A = weight on pseudocounts
and $P_0(a)$ = background probability of a

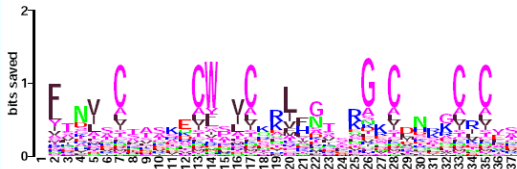
Using substitution matrix

source: tutorial ISMB, SAM-T98

Gribskov Average Score:

$$\hat{P}(a|\vec{O}) \leftarrow P_0(a) e^{\frac{\sum_b M_{a,b} O(b)}{|\vec{O}|}}$$

```
2crd  XFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
1cmr  -----CTTSKECWSVCQRLHNTSKGWCDHRGICICES
2bmt  XFTNVSCSASSQCWPVCKKLFGTYRGKCMNSKCRCSYS
```



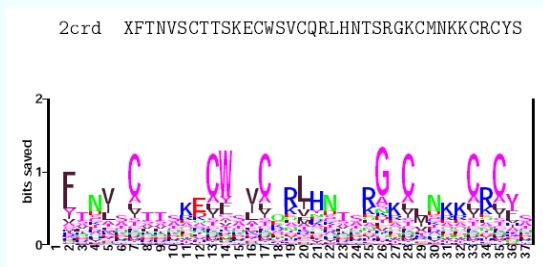
Using substitution matrix

source: tutorial ISMB, SAM-T98

Gribskov Average Score:

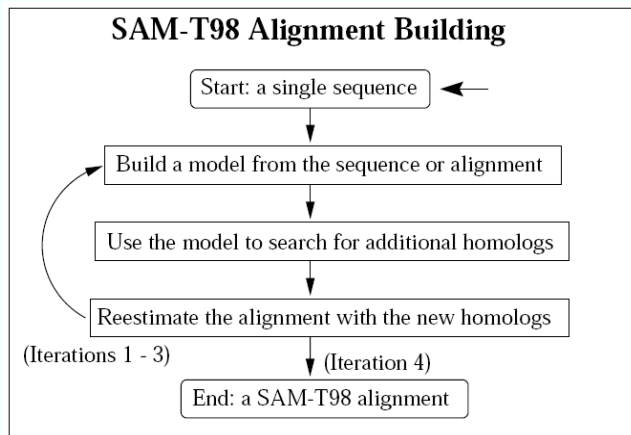
$$\hat{P}(a|\vec{O}) \leftarrow P_0(a) e^{\frac{\sum_b M_{a,b} O(b)}{|\vec{O}|}}$$

Only one sequence: \sim signal searched by Blast



SAM-T98 workflow

Starts from a single sequence !

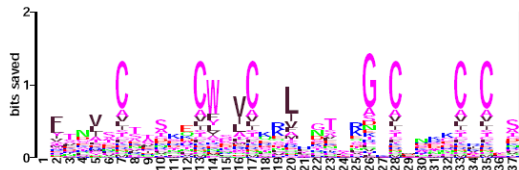


Dirichlet mixtures

source: tutorial ISMB, SAM-T98

$$\hat{P}(a|\vec{O}) = \sum_c P(c|\vec{O}) \frac{O(a) + \alpha_{c,a}}{\sum_{\text{amino acids } b} O(b) + \alpha_{c,b}}$$

```
2crd  XFTNVSC TT SKECWSVCQRLHNTSRGKCMNKKRCRCYS
1cmr  -----CTTSKECWSVCQRLHNTSKGWDHRGCICES
2bmt  XFTNVSCSASSQCWPVCKKLFGT YRGKCMNSKRCRCYS
```



Dirichlet mixtures

	uprior.9comp								
	uprior9.0	uprior9.1	uprior9.2	uprior9.3	uprior9.4	uprior9.5	uprior9.6	uprior9.7	uprior9.8
c	0.182962	0.057607	0.089823	0.079297	0.083183	0.091122	0.115962	0.06604	0.234006
$ \theta $	1.18065	1.35583	6.66436	2.08141	2.08101	2.56819	1.76606	4.98468	0.0995
A	0.270671	0.021465	0.561459	0.070143	0.041103	0.115607	0.093461	0.452171	0.005193
C	0.039848	0.0103	0.045448	0.01114	0.014794	0.037381	0.004737	0.114613	0.004039
D	0.017576	0.011741	0.438366	0.019479	0.00561	0.012414	0.387252	0.06246	0.006722
E	0.016415	0.010883	0.764167	0.094657	0.010216	0.018179	0.347841	0.115702	0.006121
F	0.014268	0.385651	0.087364	0.013162	0.153602	0.051778	0.010822	0.284246	0.003468
G	0.131916	0.016416	0.259114	0.048038	0.007797	0.017255	0.105877	0.140204	0.016931
H	0.012391	0.076196	0.21494	0.077	0.007175	0.004911	0.049776	0.100358	0.003647
I	0.022599	0.035329	0.145928	0.032939	0.299635	0.796882	0.014963	0.55023	0.002184
K	0.020358	0.013921	0.762204	0.576639	0.010849	0.017074	0.094276	0.143995	0.005019
L	0.030727	0.093517	0.24732	0.072293	0.999446	0.285858	0.027761	0.700649	0.00599
M	0.015315	0.022034	0.118662	0.02824	0.210189	0.075811	0.01004	0.27658	0.001473
N	0.048298	0.028593	0.441564	0.080372	0.006127	0.014548	0.187869	0.118569	0.004158
P	0.053803	0.013086	0.174822	0.037661	0.013021	0.015092	0.050018	0.09747	0.009055
Q	0.020662	0.023011	0.53084	0.185037	0.019798	0.011382	0.110039	0.126673	0.00363
R	0.023612	0.018866	0.465529	0.506783	0.014509	0.012696	0.038668	0.143634	0.006583
S	0.216147	0.029156	0.583402	0.073732	0.012049	0.027535	0.119471	0.278983	0.003712
T	0.147226	0.018153	0.445586	0.071587	0.035799	0.088333	0.065802	0.358482	0.00369
V	0.65438	0.0361	0.22705	0.042532	0.180085	0.94434	0.02543	0.66175	0.002967
W	0.003758	0.07177	0.02951	0.011254	0.012744	0.004373	0.003215	0.061533	0.002772
Y	0.009621	0.419641	0.12109	0.028723	0.026466	0.016741	0.018742	0.199373	0.002686

small aromatic polar + charge large I/V hydrophilic hydro- highly

NP

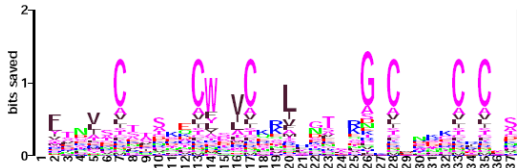
phobic conserved

Dirichlet mixtures

source: tutorial ISMB, SAM-T98

$$\hat{P}(a|\vec{O}) = \sum_c P(c|\vec{O}) \frac{O(a) + \alpha_{c,a}}{\sum_{\text{amino acids } b} O(b) + \alpha_{c,b}}$$

```
2crd  XFTNVSC TT SKECWSVCQRLHNTSRGKCMNKKRCRCYS
1cmr  -----CTTSKECWSVCQRLHNTSKGWDHRGCICES
2bmt  XFTNVSCSASSQCWPVCKKLFGT YRGKCMNSKRCRCYS
```

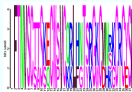


Dirichlet mixtures

source: tutorial ISMB, SAM-T98

$$\hat{P}(a|\vec{O}) = \sum_c P(c|\vec{O}) \frac{O(a) + \alpha_{c,a}}{\sum_{\text{amino acids } b} O(b) + \alpha_{c,b}}$$

```
2crd XFTNVSC TT SKECWSVCQRLHNTSRGKCMNKKRCRCYS
1cmr -----CTTSKECWSVCQRLHNTSKGWCDHRGCICES
2bmt XFTNVSCSASSQCWPVCKKLFGT YRGKCMNSKRCRCYS
```



The state-of-the-art!

PHMM: a success story on proteins

- Tools
 - HMMR
 - SAM
- PHMM Databases
 - PFAM
 - TIGRFAM
 - SUPERFAMILY
 - CATH
 - ...

And we could add the closely related tool PSI-Blast, based on profiles, and the databases relying on it ...

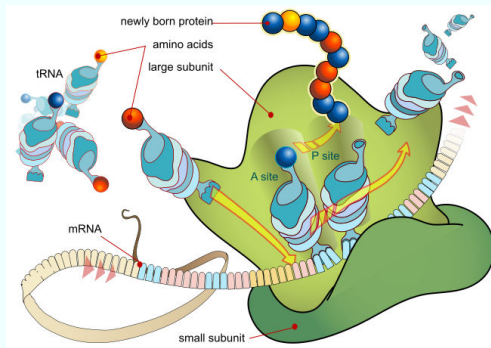
What about RNA ?

What about RNA ?

Functional (*i.e.* non-coding) RNA

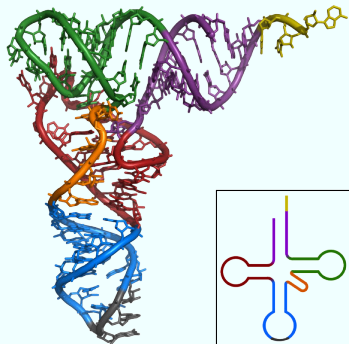
What about RNA ?

For instance, transfer RNA (many other ncRNA):



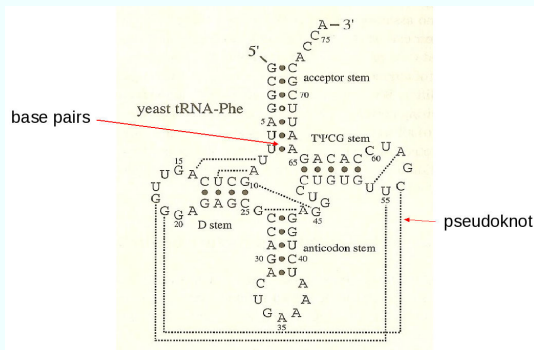
What about RNA ?

For instance, transfer RNA (many other ncRNA):



What about RNA ?

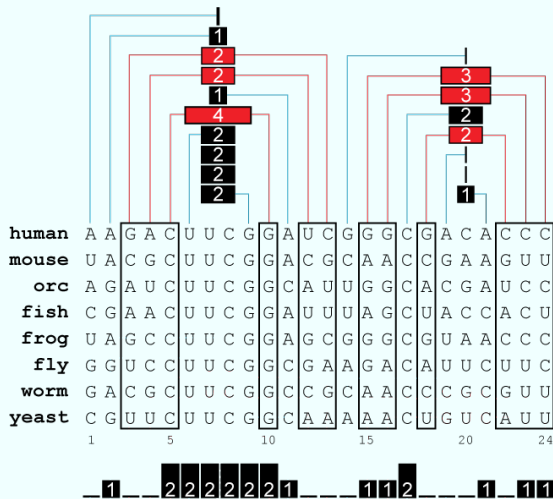
For instance, transfer RNA (many other ncRNA):



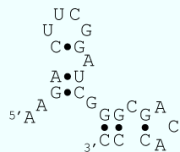
source Y. Sakakibara

RNA: conservation of secondary structure and sequence

Pairwise correlation



sequence/structure profile:
29 bits



sequence profile:
21 bits

source S. Eddy

1. Terminal alphabet is $\{A, C, G, U\}$.

2. Production rules are of the following forms :

(where, X, Y, Z are non-terminals, and $a, b \in \{A, C, G, U\}$).

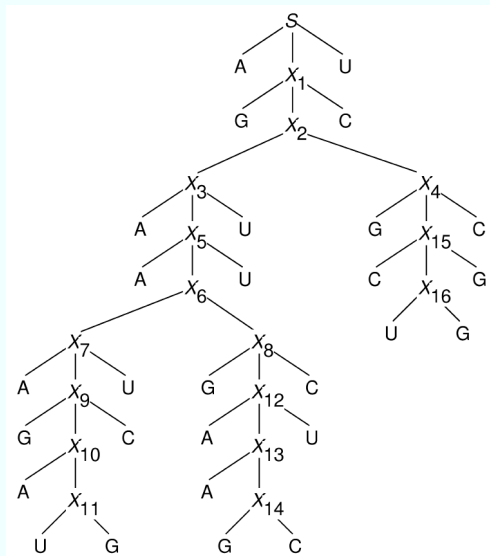
- To represent Watson-Crick complementary pair : $X \rightarrow aYb$
For examples, $X \rightarrow AYU$, $X \rightarrow UYA$, $X \rightarrow CYG$, $X \rightarrow GYC$.
- To represent single base (loops) : $X \rightarrow aY$ and $X \rightarrow a$.
- To insert gaps in alignments (deletion states) : $X \rightarrow Y$.
- To represent branching structures : $X \rightarrow YZ$.

Example

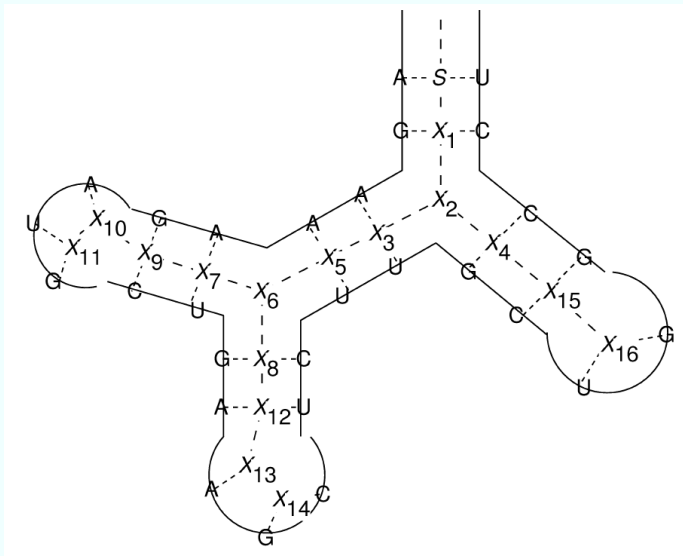
$$P = \left\{ \begin{array}{l} S \rightarrow \mathbf{A}X_1\mathbf{U}, \\ X_1 \rightarrow \mathbf{G}X_2\mathbf{C}, \quad X_2 \rightarrow X_3X_4, \quad X_3 \rightarrow \mathbf{A}X_5\mathbf{U}, \\ X_5 \rightarrow \mathbf{A}X_6\mathbf{U}, \quad X_6 \rightarrow X_7X_8, \quad X_7 \rightarrow \mathbf{A}X_9\mathbf{U}, \\ X_9 \rightarrow \mathbf{G}X_{10}\mathbf{C}, \quad X_{10} \rightarrow \mathbf{A}X_{11}, \quad X_{11} \rightarrow \mathbf{U}\mathbf{G}, \\ X_8 \rightarrow \mathbf{G}X_{12}\mathbf{C}, \quad X_{12} \rightarrow \mathbf{A}X_{13}\mathbf{U}, \quad X_{13} \rightarrow \mathbf{A}X_{14}, \\ X_{14} \rightarrow \mathbf{G}\mathbf{C}, \quad X_4 \rightarrow \mathbf{G}X_{15}\mathbf{C}, \quad X_{15} \rightarrow \mathbf{C}X_{16}\mathbf{G}, \\ X_{16} \rightarrow \mathbf{U}\mathbf{G} \end{array} \right\}$$

source Y. Sakakibara

A derivation tree



The corresponding secondary structure



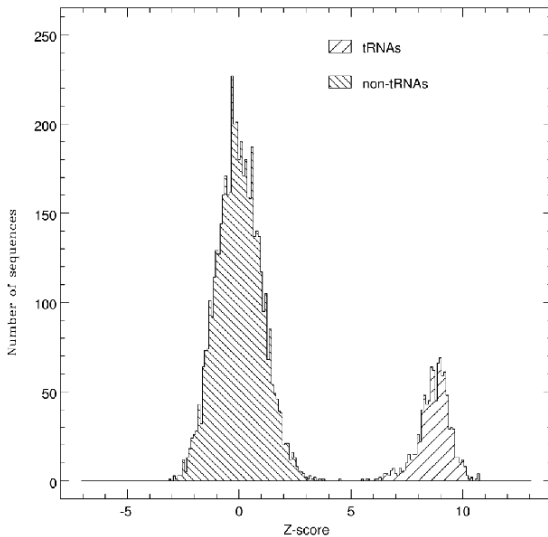
Multiple alignment

```
[      ] <   D-domain   > <   Anticodon   >< Extra ><   T-domain   >[      ]  
((((((( ((((          ))) ((((( == )))          ((((((          ))))))))))))
```

```
1 -GCCAAGGUGGCAG.AGUUcGGccUAACGCGGGGCCUGCAGAGCCGCUC---AUCGCCGGUUCAAAUCCGGCCCUUGGCU---  
2 -GGGCGUGUGGCGU.AGUC.GG.UAGCGCGCUCUUAGCAUGGGAGAGG---UCUCCGGUUCGAUUCGGACUCGUCCA---  
3 -GCCCC-AUCGUCU.AGAG.GCc.UAGGACACCUCUUUCACGGAGGCG----ACGGGGAUUCGAAUCCCCU-GGGGGU--A  
4 -GGCGGCAUAGCCA.AGC-.GG.UAAGGCCGUGGAUUGCAAUCCUCUA---UCCCCAGUUCAAAUCUGGGUGCCGCCU---  
5 -GUCUGAUUAGCGC.AACU.GG.UCAGAGCAAUGACUCUUAUUCAGUGGG---UUGUGGGUUCGAUUCACCAUCAGGCACCA  
6 -GGGCGAAUAGUGUcAGCG.GG.-AGCACACCAGACUUGCAAUUCUGGUA----GGGAGGGUUCGAGUCCUCUUUGUCCACCA  
7 -GGGGCUAUAGUUU.AACU.GG.UAAAACGCGGAUUUUGCAAUUCGUUA---UUUCAGGAUCGAGUCCUGAUAAUCUCA---  
8 -AGCUUUGUAGUUU.A--U.GU.UAAAAUGCUUGUUUGUGAUUAGAGUGA--AAU-----UGGAGCUU---
```

source Y. Sakakibara

Discrimination ability :



source Y. Sakakibara

Rfam 9.1 : Home (Janelia) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://rfam.janelia.org/ icgi 2010

Rfam 9.1 :: Home

The Rfam database of RNA alignments and CMs

HHMI
janelia farm
research campus

[rfam consortium \(cambridge\)](#) | [rfam mirror \(janelia\)](#) | [infernal](#) | [eddy lab](#) | [janelia farm home](#) | [browse rfam](#) | [sequence search](#) | [help](#)

Rfam 9.1 (December 2008, 1371 families)

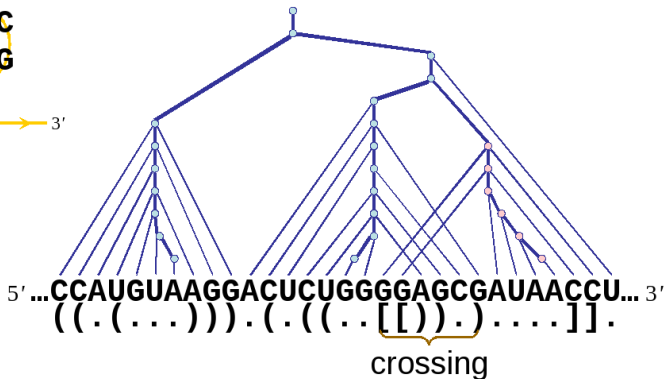
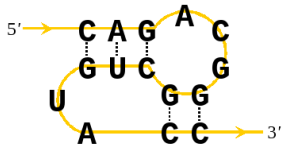
Rfam is a collection of multiple sequence alignments and covariance models covering many common non-coding RNA families. The main use of Rfam is as a source of RNA multiple alignments with consensus secondary structure annotation in a consistent format. In conjunction with the [Infernal](#) software package, Rfam covariance models (CMs) can be used to search genomes or other DNA sequence databases for homologs to known structural RNA families.

Rfam makes use of a large amount of available RNA alignment data, especially published multiple sequence alignments, and repackages these data in a single searchable and sustainable resource. We have made every effort to credit individual sources on each family page (see the [help page](#) for a list). If you find any of the data presented here useful, please also be sure to credit the primary source.

Rfam is produced by the Rfam Consortium, a collaboration between researchers at the [Wellcome Trust Sanger Institute](#) near Cambridge, UK, the [University of Manchester](#) in Manchester, UK, and [HHMI Janelia Farm](#) near Washington, DC.

[BROWSE RFAM](#) View Rfam annotation and alignments.

- ▶ Branches (parentheses) cross in pseudoknots



Conclusion of this part

So far:

- Ad-hoc simple topologies designed wrt domain knowledge
- Modelling conserved regions
- Efficient parameter weighting schemes
- It works well!

We would like also:

- More expressive topologies, learnt automatically
- Better understand chaining of the conserved regions (grammar of the DNA ?)
- Efficient parameter weighting schemes (why not, but maybe less important if the topology is good)
- and it should still work well!

- 1 Molecular genetics
- 2 Modelling a set of conserved sequences
- 3 Inference of grammatical structure
 - Pattern (motif) Discovery
 - Learning Automata

Pattern (motif) Discovery

- DNA: A lot of work on the search of over/under-represented words (topology :- ()
- Proteins: not so much tools (20 letters, substitutions)
Teiresias, Splash, Emotif, Pratt...
No need anymore of aligned sequences. Topology ? Maximum level of expressiveness: Prosite motifs (Pratt), still unable to

accept:

... R ... E ...

... E ... R ...

and reject:

... E ... E ...

... R ... R ...

(position specific pattern: $[ER]_{x(l,k)}[ER]$)

Learning local languages and application to DNA

Yokomori et al, 1994, 1998

- Splicing systems: model of recombination of DNA [Head 1987] (crossover operation)
- *Persistent* splicing languages are equivalent to *strictly locally testable* (regular) languages

$$\{L \mid \exists k, \exists A, B, C \subseteq \Sigma^k, \forall w, |w| \geq k :$$

$$w \in L \text{ iff } L_k(w) \in A, I(w) \in B, R(w) \in C\}$$

- Class of k -testable language is learnable in the limit using DFA

Sequence Encoding

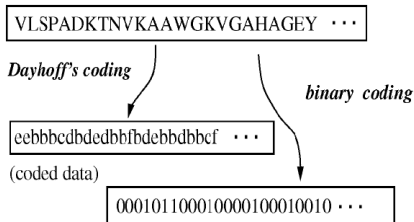


Fig. 5. Raw data of amino acid sequence from human hemoglobin is coded into new data over D_6 and B.

TRANSLATION TABLES
DAYHOFF'S CODING [9]

Amino Acids	Properties	New Symbols
C	sulfur polymerization	a
S, T, P, A, G	small	b
N, D, E, Q	acid & amide	c
H, R, K	basic	d
M, I, L, V	hydrophobic	e
F, Y, W	aromaticity	f

(a)

BINARY CODING [25]

Amino Acids	Hydropathy Index	New Symbols
A, C, F, G, I, L, M, N, S, T, V, W, Y	High	0
D, E, H, K, P, Q, R	Low	1

(b)

Experiments



Protein α -chain identification [Yokomori et al 1994, 1998]
 $\sim 95\%$ classification success (hemoglobin ?)

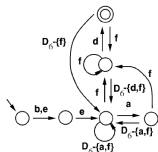


Fig. 8. A DFA $M_{2,30}$ most frequently obtained in experiments Exp(2, 30) over D_0 . This DFA was obtained more than 70 among 100 times of random experiments.

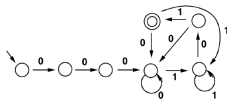
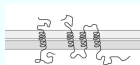


Fig. 9. A DFA $M_{3,6}$ most frequently obtained in experiments Exp(3, 6) over B. This DFA was obtained more than 10 among 100 times of random experiments.



Coiled-coil proteins [Peris et al ICGI 2006]



Transmembrane domain [Peris et al BMC Bioinf. 2008]

Learning automata on proteins

It is easy to represent a set of protein sequences by an automaton!

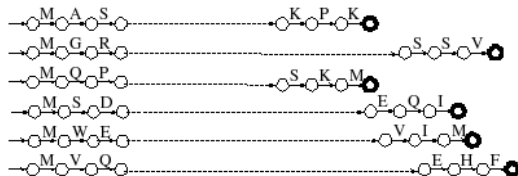
```
>AQP1_BOVIN  
MASEFKKKLFWRAVVAEFL...KPK  
>AQP3_MOUSE  
MGRQKELMNRCGE...SSV  
>AQP9_HUMAN  
MQPEGAEKGKSFKQRLVLKSSLA...SKM  
>AQP4_BOVIN  
MSDRPAATRWGKCGPLCTRES...EQI  
>AQP2_RAT  
MWELRSIAFSRAVLAEFLAT...VIM  
>AQP7_HUMAN  
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Learning automata on proteins

It is easy to represent a set of protein sequences by an automaton!

```
>AQP1_BOVIN
MASEFKKKLFWRAVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGKSFQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Maximal Canonical Automaton

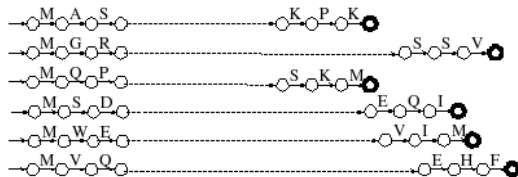


Learning automata on proteins

It is easy to represent a set of protein sequences by an automaton!

```
>AQP1_BOVIN
MASEFKKKLFWRAVVAEFL...KPK
>AQP3_MOUSE
MGRQKELMNRCGE...SSV
>AQP9_HUMAN
MQPEGAEKGKSFQRLVLKSSLA...SKM
>AQP4_BOVIN
MSDRPAATRWGKCGPLCTRES...EQI
>AQP2_RAT
MWELRSIAFSRAVLAEFLAT...VIM
>AQP7_HUMAN
MVQASGHRRSTRGSKMVSWSVP...EHF
```

Maximal Canonical Automaton



Rote learning, an inductive leap is needed...

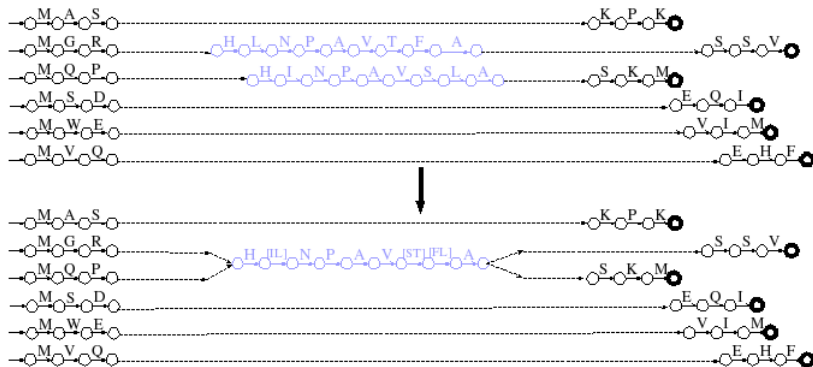
Protomata Learner

[Coste et al 2005,2006]

State (fragment) merging approach, EDSM inspiration

- Merging similar fragment pairs

Maximal Canonical Automaton

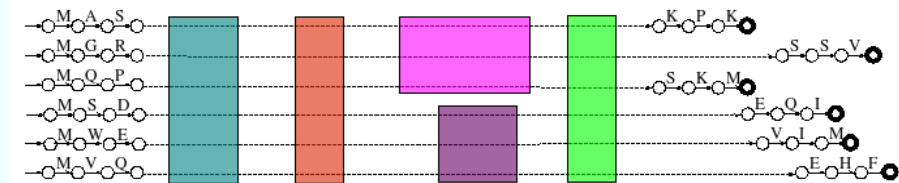


Protomata Learner

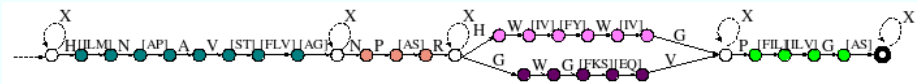
[Coste et al 2005,2006]

A new kind of alignment: Partial Local Multiple Alignment

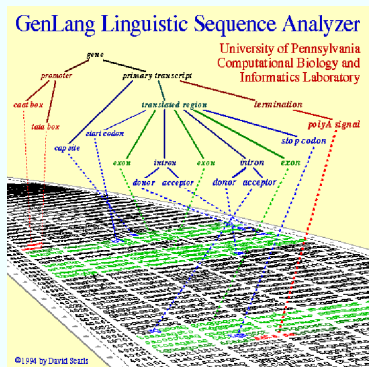
Maximal Canonical Automaton



Merge blocks, detect exceptions, merge gaps:



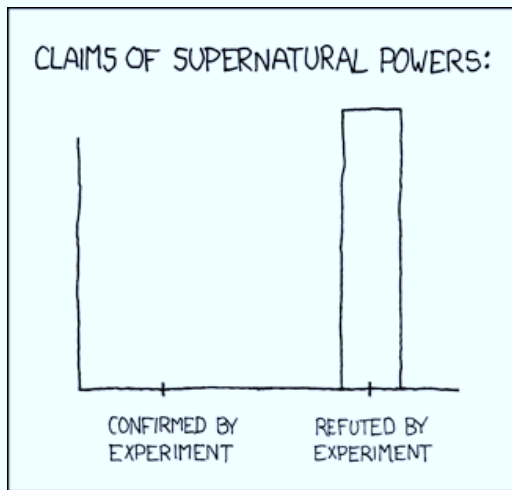
Learning the grammar of DNA sequences ?



More expressive topologies ?

- Adios, ABL, Distributional learning ?
- Hierarchical (context-free) structure inference (Sequitur ...)
- Invited talk of D. Searls on Wednesday ...

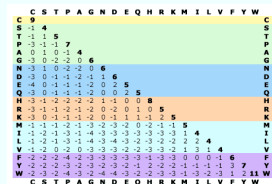
In-silico predictions. . .



“Tutorial on Modelling Biological Sequences
by Grammatical Inference: Bibliography”,
F. Coste, ICGI'2010 Tutorial Day

Appendix

Dayhoff encoding



a	C	sulfhydryl
b	G, S, T, A, P	small hydrophilic
c	D, E, N, Q	acid, acid-amide and hydrophilic
d	R, H, K	basic
e	L, V, M, I	small hydrophobic
f	Y, F, W	aromatic

Prosite's Pattern

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

Numerical results

- UniProtKB/Swiss-Prot release number: **54.7**, total number of sequence entries in that release: **333445**.
- Total number of hits in UniProtKB/Swiss-Prot: **11060 hits in 1794 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **10810 hits in 1609 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **26 hits in 12 different sequences**
- Number of false hits (on unrelated proteins): **224 hits in 173 different sequences**
- Number of known missed hits: **60**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **97.97 %**
- Recall (true hits / (true hits + false negatives)): **99.45 %**

Probabilistic version: Prosite's Profile

Matrix / Profile

```

      A   B   C   D   E   F   G   H   I   E   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:   HI=0; HI=-105; HD=-105;
/M:  SV=V*:  H=-19,-21,-24,-25,-21, 39,-28, 10,-2,-17, 2, 0,-17,-28,-19,-13,-18,-10,-7, 12, 48,-21;
/M:  SV=V*:  H=-4,-5,-23,-5, 6,-20,-18,-9,-14, 7,-15,-7,-5,-9, 3, 2,-2,-2,-9,-25,-12, 4;
/M:  SV=C*:  H=-10,-20,118,-30,-30,-20,-30,-30,-30,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30;
/M:  SV=W*:  H=-5, 3,-24, 3, 6,-22,-11,-6,-20, 1,-21,-14, 4,-1, 1,-3, 5, 2,-18,-29,-15, 3;
/I:   I=-12; HI=0; HD=-30; DI=0; DI=-30;
/M:  SV=V*:  H=-9,-2,-26, 1, 14,-18,-17,-4,-13,-1,-11,-8,-5,-12, 4,-5,-5,-8,-12,-24,-9, 8;
/M:  SV=C*:  H=-10,-20,119,-30,-30,-20,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-29,-30;
/M:  SV=G*:  H=-3,-1,-28,-1,-7,-28, 36,-11,-33,-11,-27,-18, 4,-15,-10,-12, 1,-13,-27,-24,-23,-9;
/M:  SV=V*:  H=-10,-2,-28,-3, 8,-25,-19,-7,-26, 36,-24,-8,-1,-12,10,27,-9,-9,-18,-19,-8, 8;
/M:  SV=A*:  H= 8,-7,-9,-11,-7,-17,-7,-14,-16,-6,-16,-11,-4,-15,-6, 5, 8, 4,-7,-27,-15, 7;
/M:  SV=V*:  H=-19,-29,-19,-37,-28, 71,-29,-17, 0,-28, 9, 0,-20,-30,-36,-19,-19,-9,-1, 9, 31,-28;
/M:  SV=V*:  H= 0,-5,-17,-9,-6,-16,-11,-10,-14,-3,-16,-10, 0,-12,-4, 0, 8, 7,-8,-27,-12,-6;
/M:  SV=C*:  H=-10,-3,-20,-4, 0,-18,-17, 2,-19, 3,-16,-8, 0,-17, 8, 9,-1,-3,-17,-19,-5, 3;
/M:  SV=W*:  H=-4,-4,-22,-6, 0,-19,-13,-5,-18, 7,-18,-9, 1,-10, 2, 8, 2,-2,-14,-25,-11, 0;
/M:  SV=M*:  H= 2,-1,-18,-1,-1,-18,-4,-6,-19,-7,-22,-14, 4,-12,-2,-7, 16, 7,-13,-29,-12,-2;
/M:  SV=V*:  H=-5, 5,-20, 1, 0,-18,-10, 8,-18,-5,-18,-11, 9,-16, 1,-4, 4,-1,-17,-27,-7,-1;
/M:  SV=L*:  H=-11,-29,-20,-30,-20, 12,-29,-19, 17,-27, 43, 18,-28,-29,-20,-18,-28,-10, 9,-18, 2,-20;
/M:  SV=V*:  H=-6,-6,-22,-10,-4,-15,-20,-8,-7, 2,-9,-3,-2,-16, 0, 3,-3, 0,-6,-24,-8,-3;
/M:  SV=V*:  H=-6,-6,-23,-7, 0,-19,-17,-7,-14, 7,-13,-6,-2,-16, 4, 12,-3,-3,-10,-24,-10, 0;
/M:  SV=W*:  H=-20, 0,-30, 0, 0,-20,-20,99,-30,-10,-20, 0, 10,-20, 10, 0,-10,-20,-30,-30, 20, 0;
/M:  SV=Q*:  H=-10,-10,-25,-12, 1,-16,-22,-2,-6, 1,-3, 6,-9,-17, 13, 3,-9,-8,-9,-19,-4,-6;
/M:  SV=V*:  H=-13,-8,-26,-9, 0,-19,-19,-4,-21, 20,-16,-6,-2,-17, 6, 35,-8,-7,-14,-21,-9, 0;
/I:   I=-12; HI=0; HD=-29; DI=0; DI=-29;
/M:  SV=V*:  H=-3,-16,-17,-21,-17,-6,-25,-20, 11,-15, 2, 3,-12,-18,-14,-14,-2, 9, 13,-25,-7,-17;
/M:  SV=M*:  H=-20, 0,-30, 0, 0,-20,-20,97,-30,-10,-20, 0, 10,-20, 10, 0,-10,-20,-30,-30,-19, 0;
/M:  SV=V*:  H= 1,-2,-13,-8,-6,-14,-15,-15,-12,-7,-13,-9, 0,-12,-6,-8,-14, 25,-5,-29,-12,-6;
/M:  SV=G*:  H=-3,-4,-27,-4,-8,-27, 34,-14,-32,-9,-25,-16, 2,-15,-10,-9, 0,-13,-24,-23,-23,-10;
/M:  SV=V*:  H=-9, 6,-27, 12, 33,-26,-17,-3,-24, 7,-18,-15,-1,-6, 12, 0,-2,-8,-22,-28,-16, 22;
/M:  SV=V*:  H=-11,-1,-28,-2, 6,-25,-17,-6,-27, 32,-25,-10, 1,-11, 8, 28,-7,-8,-19,-22,-11, 6;
/M:  SV=W*:  H=-7,-14,-32,-9,-1,-24,-17,-15,-18,-7,-23,-15,-13, 51,-7,-13,-6,-6,-23,-28,-22,-7;
/I:   KI=0;

```

Numerical results

- UniProtKB/Swiss-Prot release number: **54.7**, total number of sequence entries in that release: **333445**.
- Total number of hits in UniProtKB/Swiss-Prot: **11272 hits in 1555 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **11257 hits in 1540 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **6 hits in 6 different sequences**
- Number of false hits (on unrelated proteins): **9 hits in 9 different sequences**
- Number of known missed hits: **126**
- Number of protein sequences which belong to the set under consideration but which are not in the UniProtKB/Swiss-Prot database: **126**