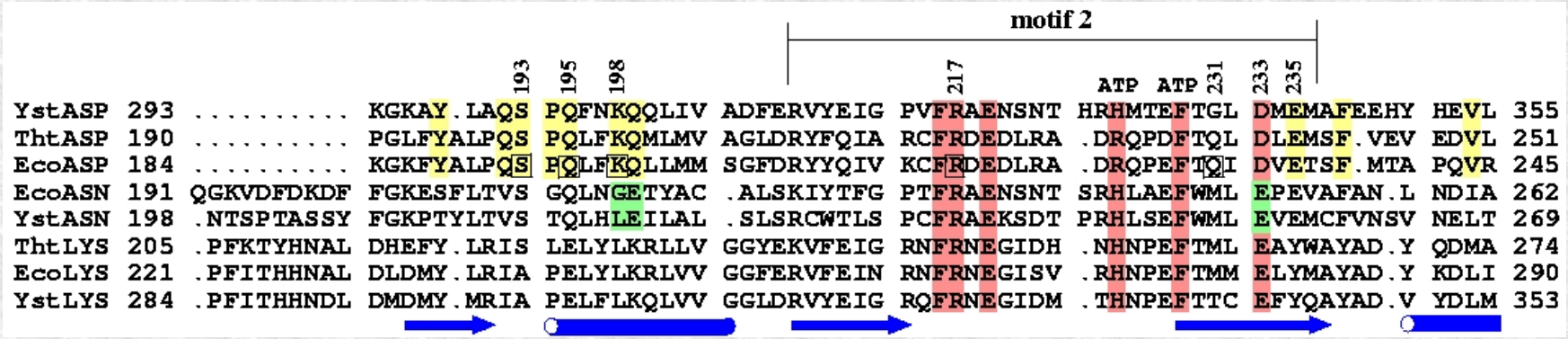
Sequence comparaison. II.



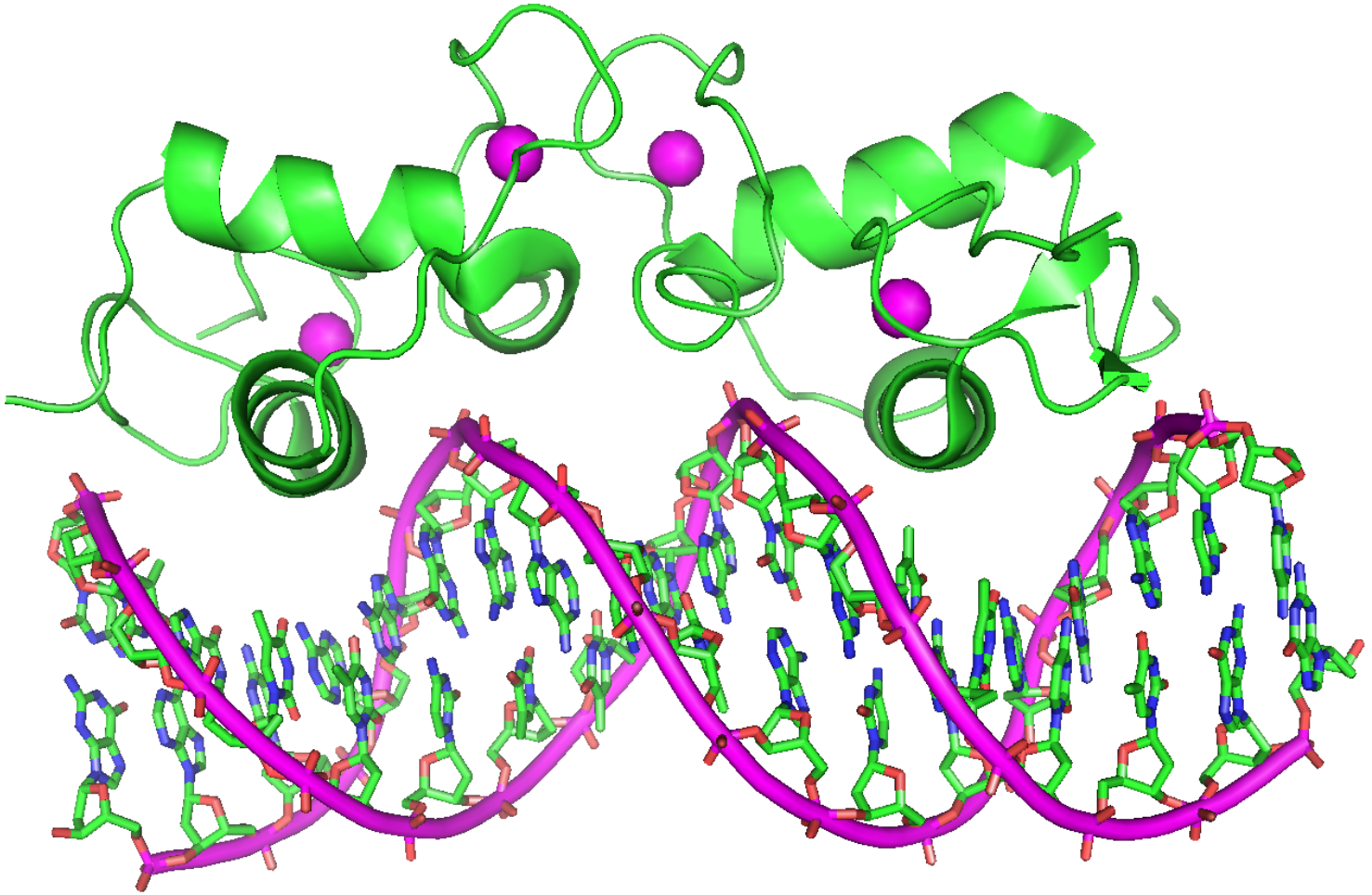
Multiple sequence alignments: method

Applications

Functional sites and their identification

See book chapters 2 and 4

Androgen receptor



Homologues of the androgen receptor identified using BLAST

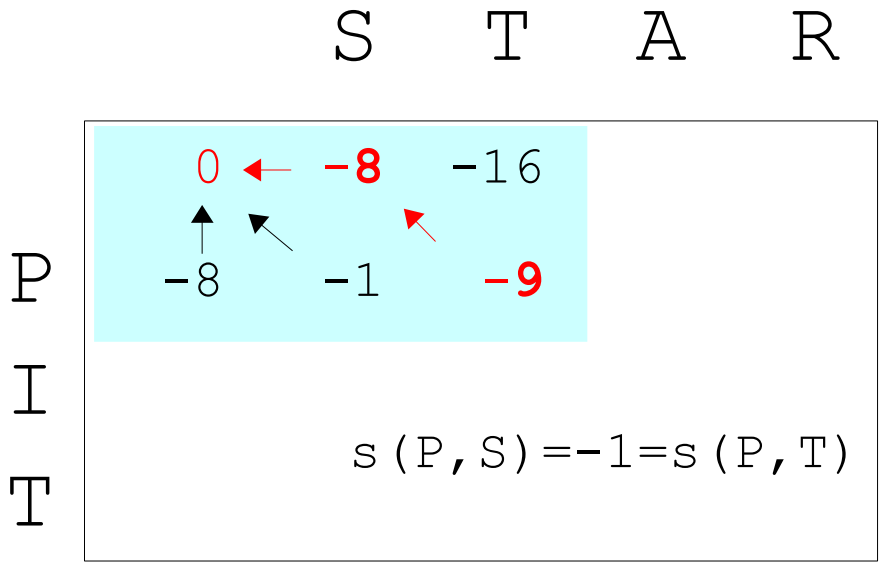
#	ID Swissprot	Name	Description	Score	E	% Identity	Length aligned
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor alpha	98	4e-21	55	72
:	:	:	:	:	:	:	
:	:	:	:	:	:	:	
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47

How can we combine these into a single alignment?

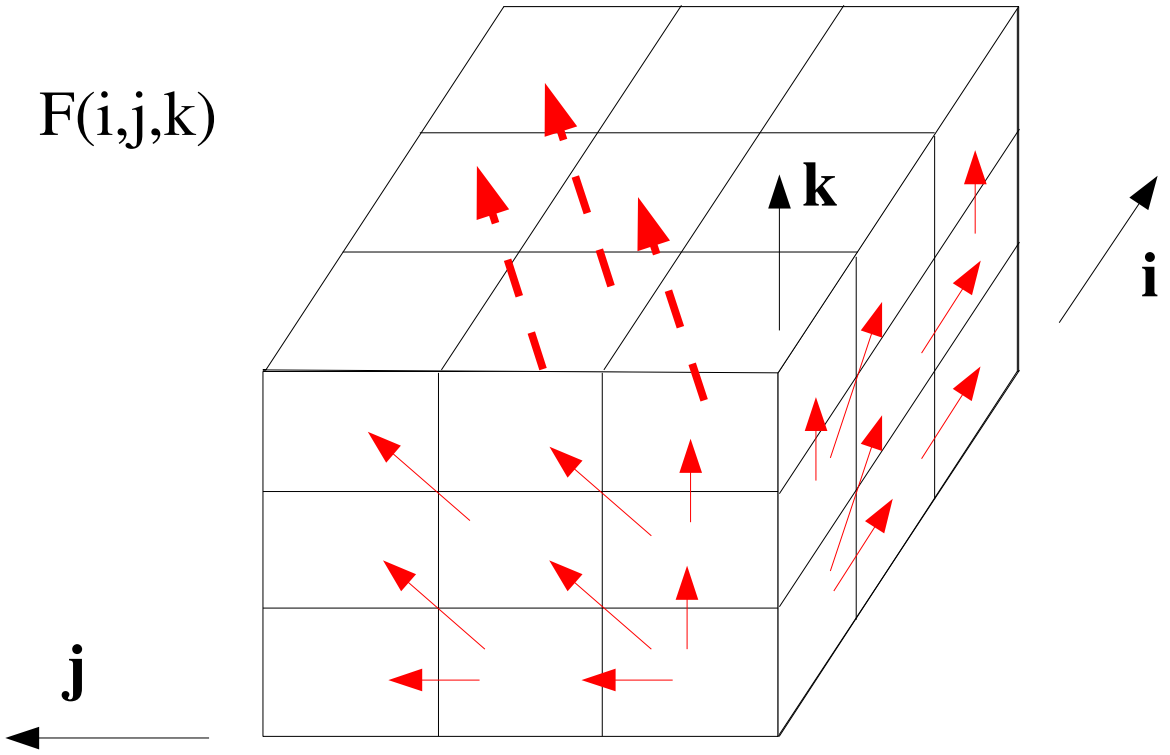
Dynamic programming algorithm for two sequences (seen above)

F(i,j) is computed recursively from smaller alignments

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$



Dynamic programming algorithm for 3 sequences



Too expensive for more than ~10 sequences

$50^{10} = 10^{17}$

We need a more efficient, “heuristic” method

#	ID Swissprot	Name	Description	Score	E	% Identity	Length aligned
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor alpha	98	4e-21	55	72
:	:	:	: : :	:	:	:	
:	:	:	: : :	:	:	:	
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47

How to merge these homologues into a single alignment?

#	ID Swissprot	Name	Description	Score	E	% Identity	Length aligned
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor alpha	98	4e-21	55	72
:	:	:	:	:	:	:	
:	:	:	:	:	:	:	
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47

One method is to infer the alignment from the Blast results: how?

A heuristic method for multiple sequence alignment in three stages

1) Align each pair of sequences

2) ***Classify*** the sequences by similarity: “guide” tree

3) Incorporate the sequences progressively into an alignment, in a sensible order (determined by the tree)

A pairwise alignment corresponds to simple evolutionary model, including

- An assumption of “minimal” divergence from a ***common ancestor***
- An empirical estimation of mutation probabilities
- An assumption of equivalent and independent positions
- A well-defined reference, or “null” model

```
C E C H A T E S T G R A S
L E R - A T E S T P R I S
```

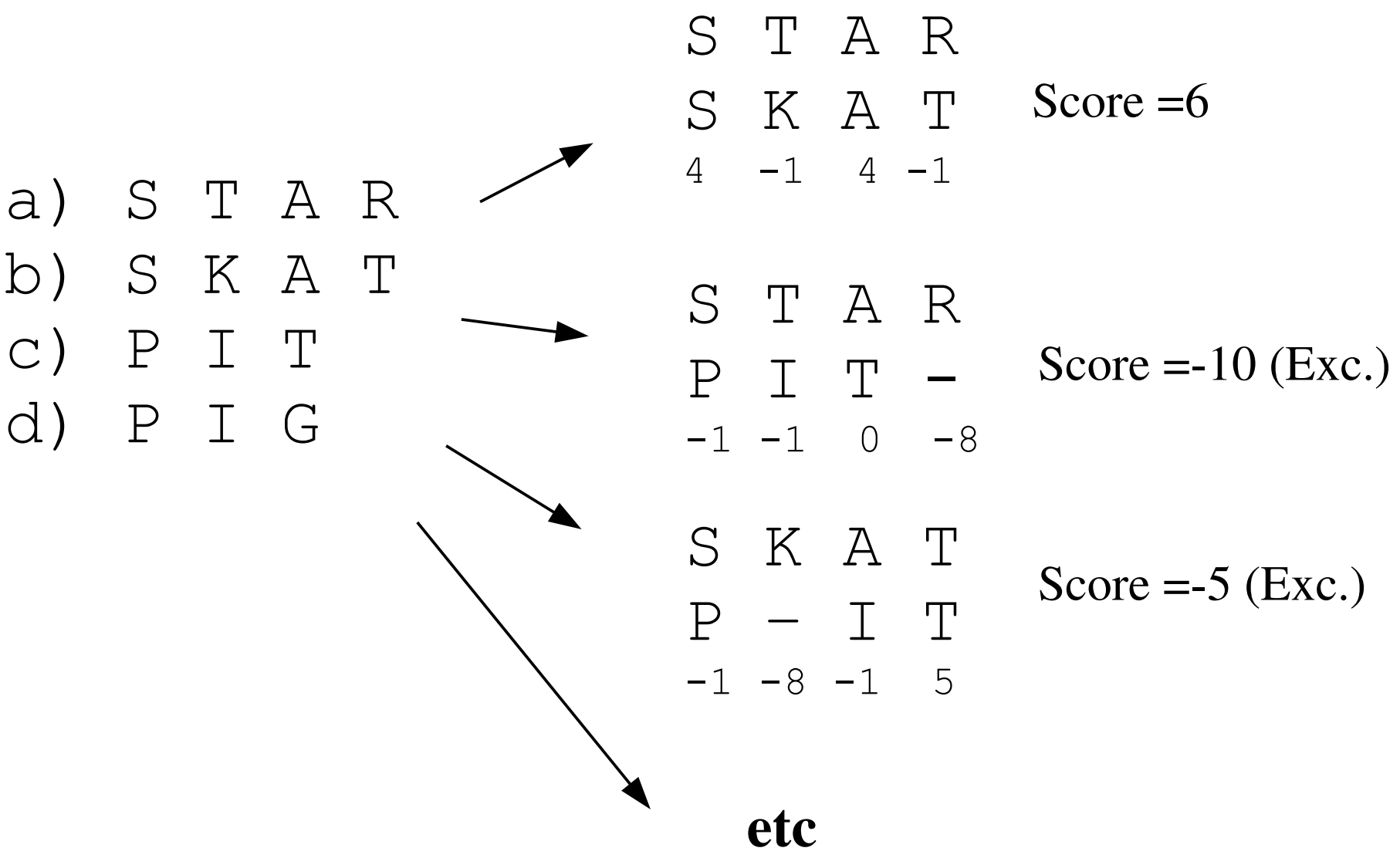
An empirical scoring matrix

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

BLOSUM62

also gap penalties

Stage 1: alignment of all sequence pairs

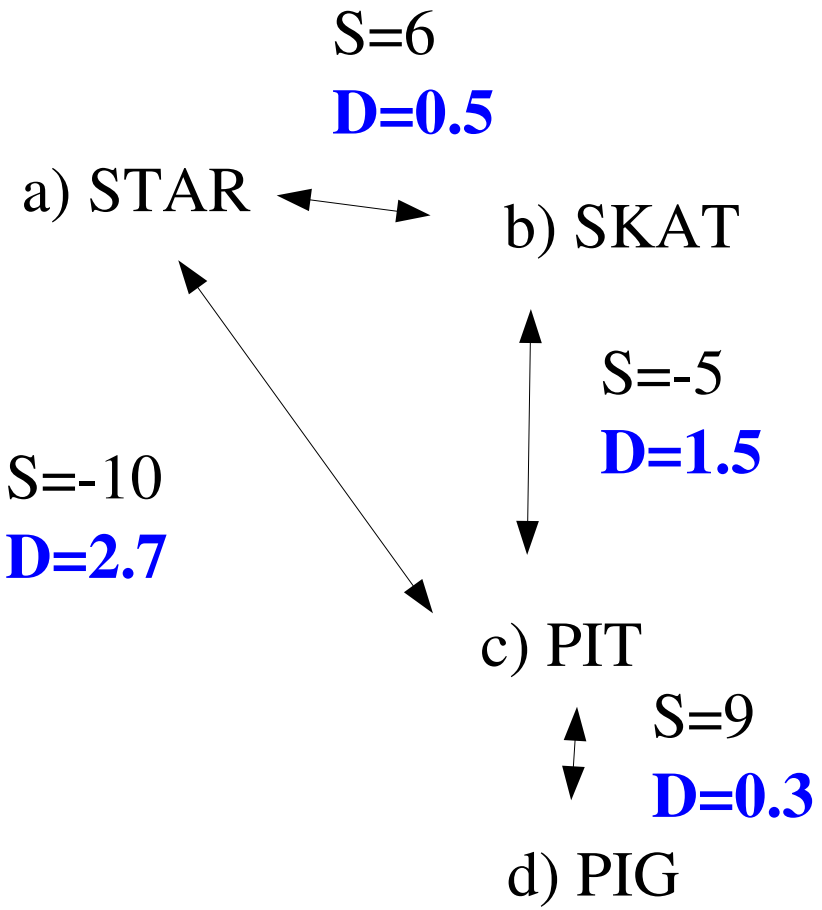


Stage 2: Classify the sequences by similarity: “guide” tree

Stage 3: Incorporate the sequences progressively into the alignment, in a sensible order

Stage 2: compute "distances" between sequences

Classification methods use distances, rather than similarity



What would be your choice of distance?

Stage 2: compute "distances" between sequences

$$D(a,b) = -\log \frac{S(a,b) - \bar{S}_{rand}}{S_{max} - \bar{S}_{rand}}$$

$S_{max} = \max(S(a,a),S(b,b))$

S_{rand} = score to align two random sequences of same length, composition

S=6

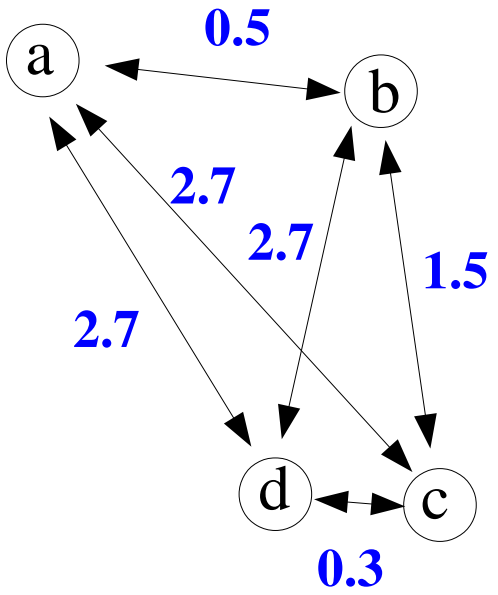
S_{max}=18 [Blosum62]

S_{rand} = -12

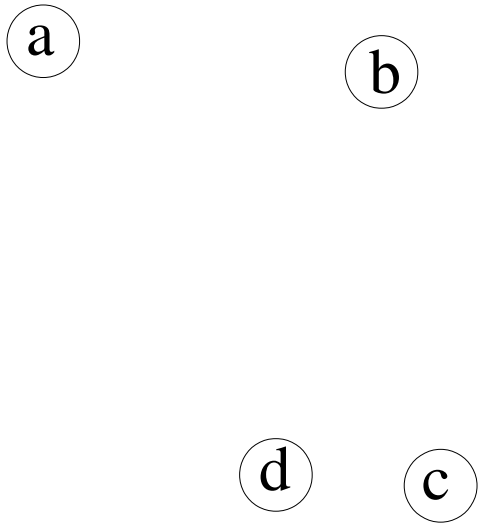
D=0.5

STAR ⟷ SKAT

Stage 2: compute “distances” between sequences

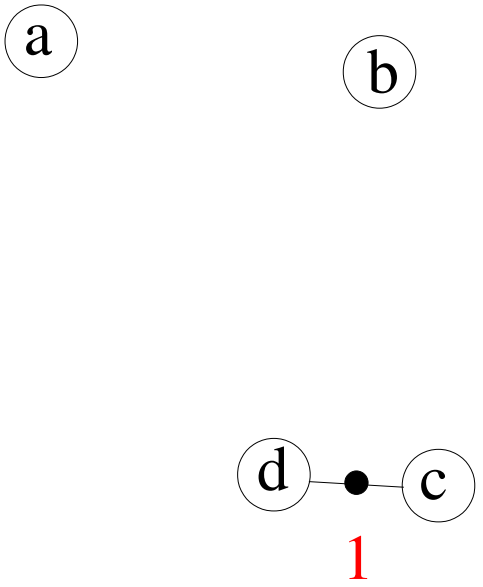


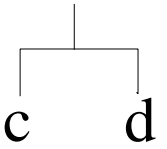
Stage 2: hierarchical classification



“Unweighted Pair Group Joining with Arithmetic Mean”
or UPGMA

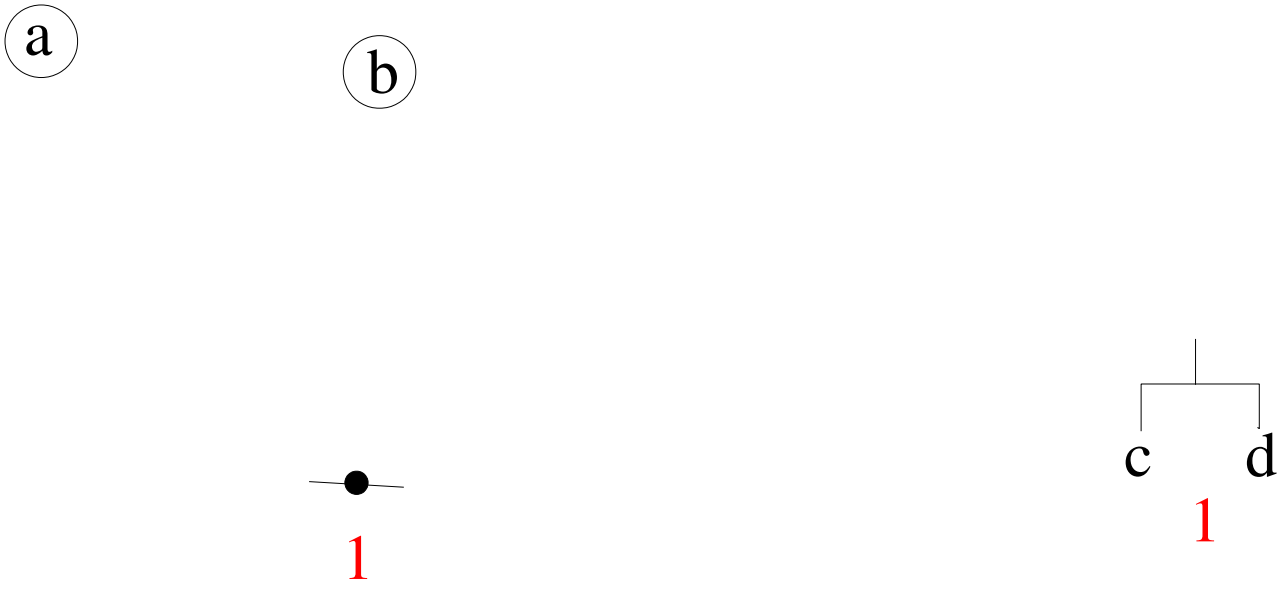
Stage 2: hierarchical classification





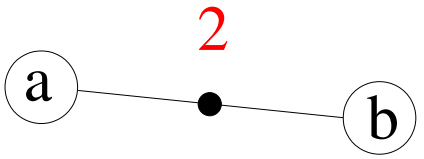
“Unweighted Pair Group Joining with Arithmetic Mean”
or UPGMA

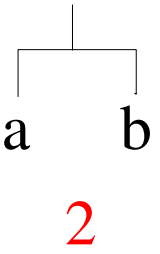
Stage 2: hierarchical classification

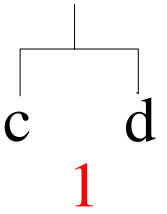


"Unweighted Pair Group Joining with Arithmetic Mean"
or UPGMA

Stage 2: hierarchical classification

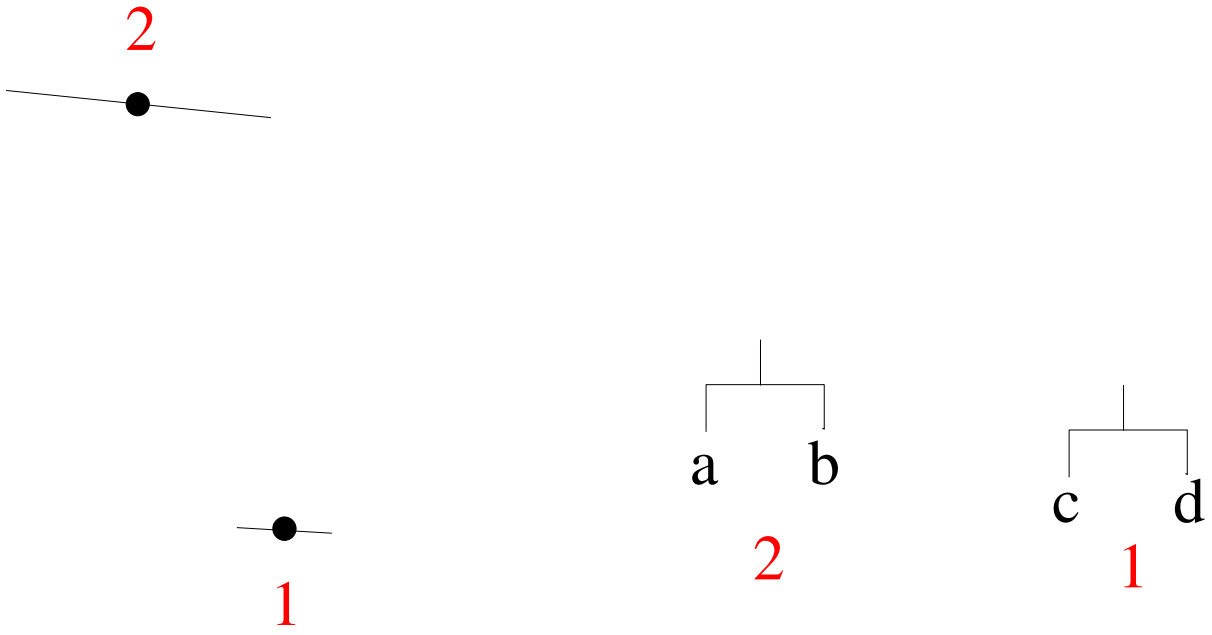






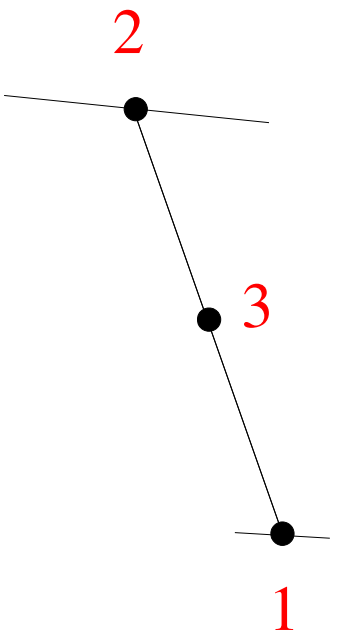
"Unweighted Pair Group Joining with Arithmetic Mean"
or UPGMA

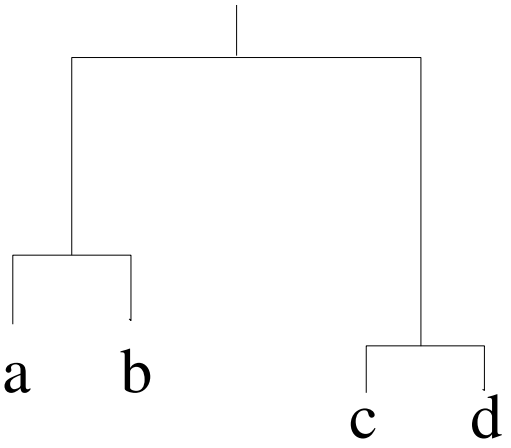
Stage 2: hierarchical classification



"Unweighted Pair Group Joining with Arithmetic Mean"
or UPGMA

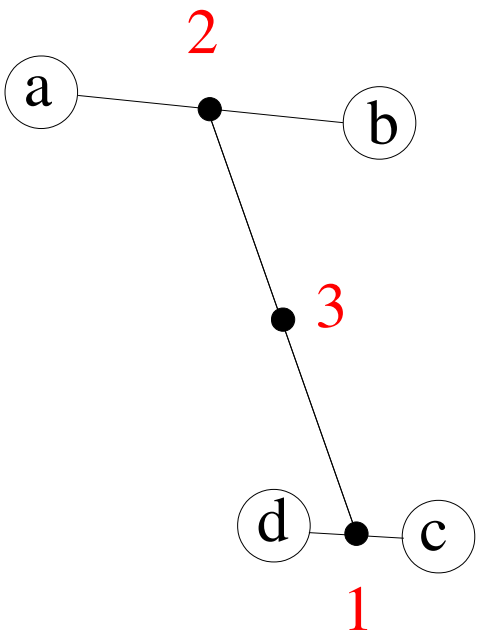
Stage 2: hierarchical classification, or guide tree

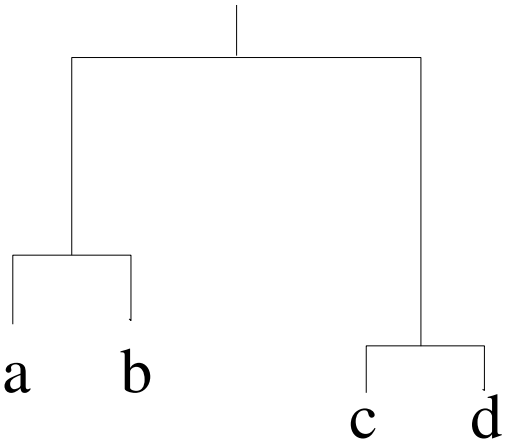




"Unweighted Pair Group Joining with Arithmetic Mean"
or UPGMA

Stage 2: hierarchical classification, or guide tree

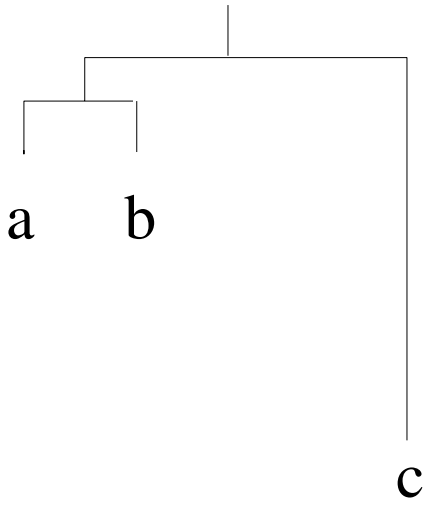




“Unweighted Pair Group Joining with Arithmetic Mean”
or UPGMA

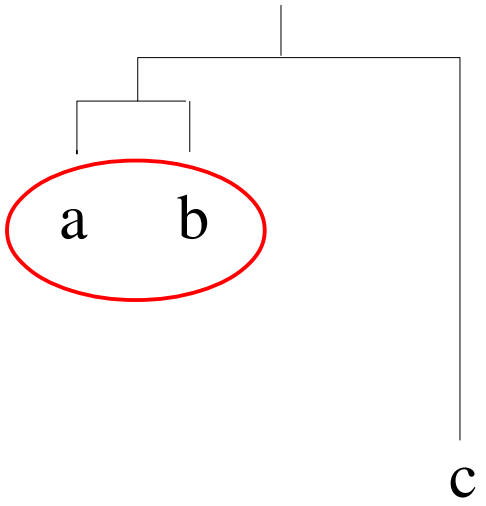
Stage 3: sequence alignment, moving up the tree

Example with 3 sequences



Stage 3: sequence alignment, moving up the tree

Example with 3 sequences

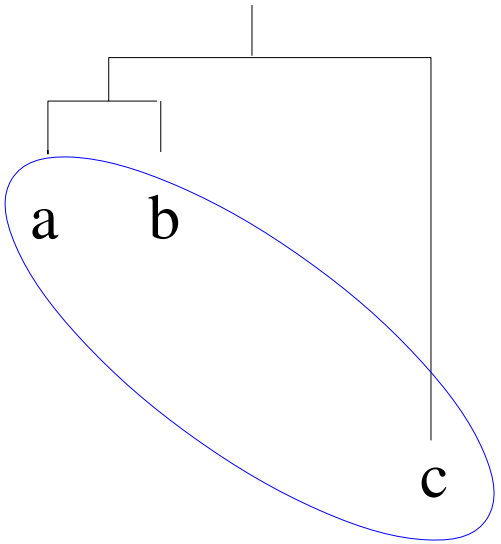


a with b

```
a)  S  T  A  R
b)  S  K  A  T
```

Stage 3: sequence alignment, moving up the tree

Example with 3 sequences



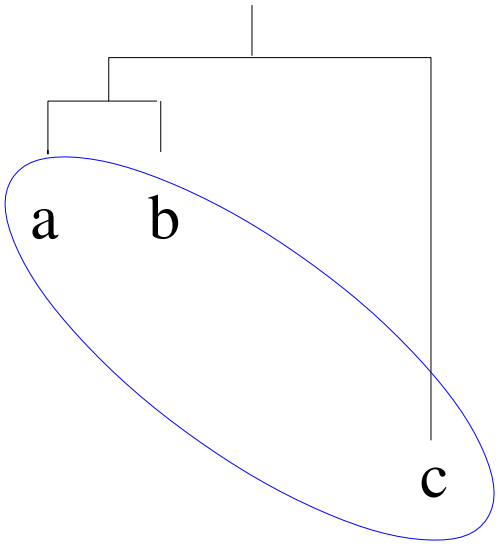
a with b

```
a)  S  T  A  R
b)  S  K  A  T
```

c with {a, b}

Stage 3: sequence alignment, moving up the tree

Example with 3 sequences



a with b

```
a)  S  T  A  R
b)  S  K  A  T
```

c with {a, b}

How to align a sequence with an alignment?

“sequence-profile” alignment

We need to generalize slightly our pairwise alignment method....

Aligning a sequence with an alignment

An alignment of 3 sequences:

```
S   T   A   R
S   T   I   R
S   K   A   T
```

4th sequence y, to be aligned:

```
P   I   T
```

Aligning a sequence with an alignment

An alignment of 3 sequences:

```
S   T   A   R
S   T   I   R
S   K   A   T
```

4^{th} sequence y, to be aligned:

```
P   I   T
```

View the alignment (above) as *a sequence of columns* X_i

Aligning a sequence with an alignment

An alignment of 3 sequences = a *sequence* of columns X_i

```
S   T   A   R
S   T   I   R
S   K   A   T
```

X_2

4th sequence y, to be aligned:

```
P   I   T
```

y_1

Aligning a sequence with an alignment: computing a "mean" score

An alignment of 3 sequences:

```
S  T  A  R
S  T  I  R
S  K  A  T
```

X_2

```
   P
T  -1
T  -1
K  -1
-------
Total -3  =s(y1,X2)
```

4th sequence y, to be aligned:

```
P  I  T
```

y_1

Aligning a sequence with an alignment: computing a "mean" score

An alignment of 3 sequences:

```
S  [T]  A  R
S  [T]  I  R
S  [K]  A  T
```

→

```
     P
T   -1
T   -1
K   -1
total -3
```

4th sequence y, to be aligned:

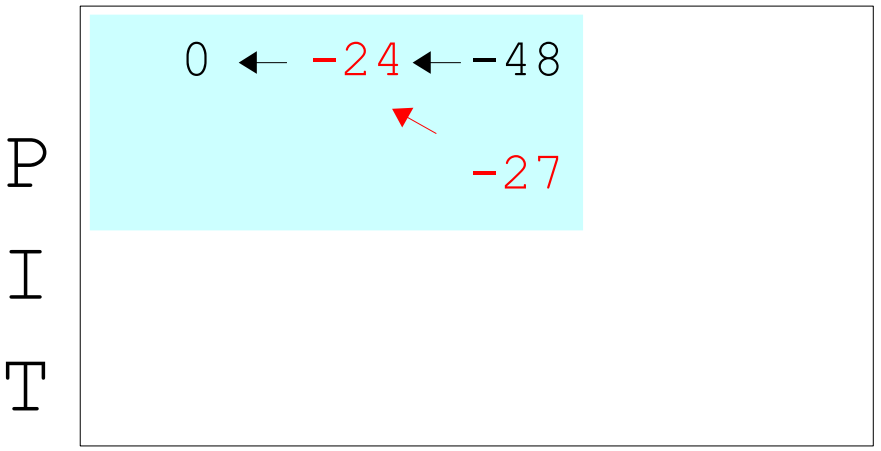
```
[P] I  T
```

$$s(a, \begin{bmatrix} b \\ b' \\ b'' \end{bmatrix}) = s(a,b) + s(a,b') + s(a,b'')$$

Sum over pairs

Sequence-“profile” alignment : dynamic programming, as before

S T A R
S T I R
S K A T

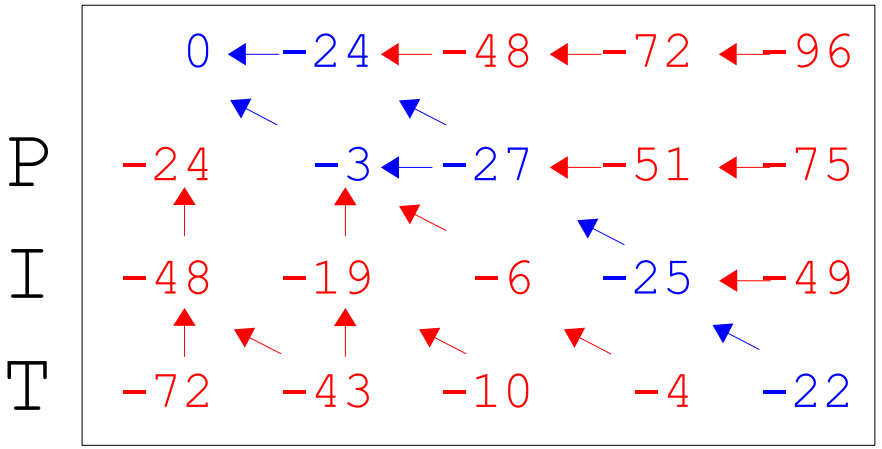


gap penalities are tripled

	P
T	-1
T	-1
K	-1
total	-3

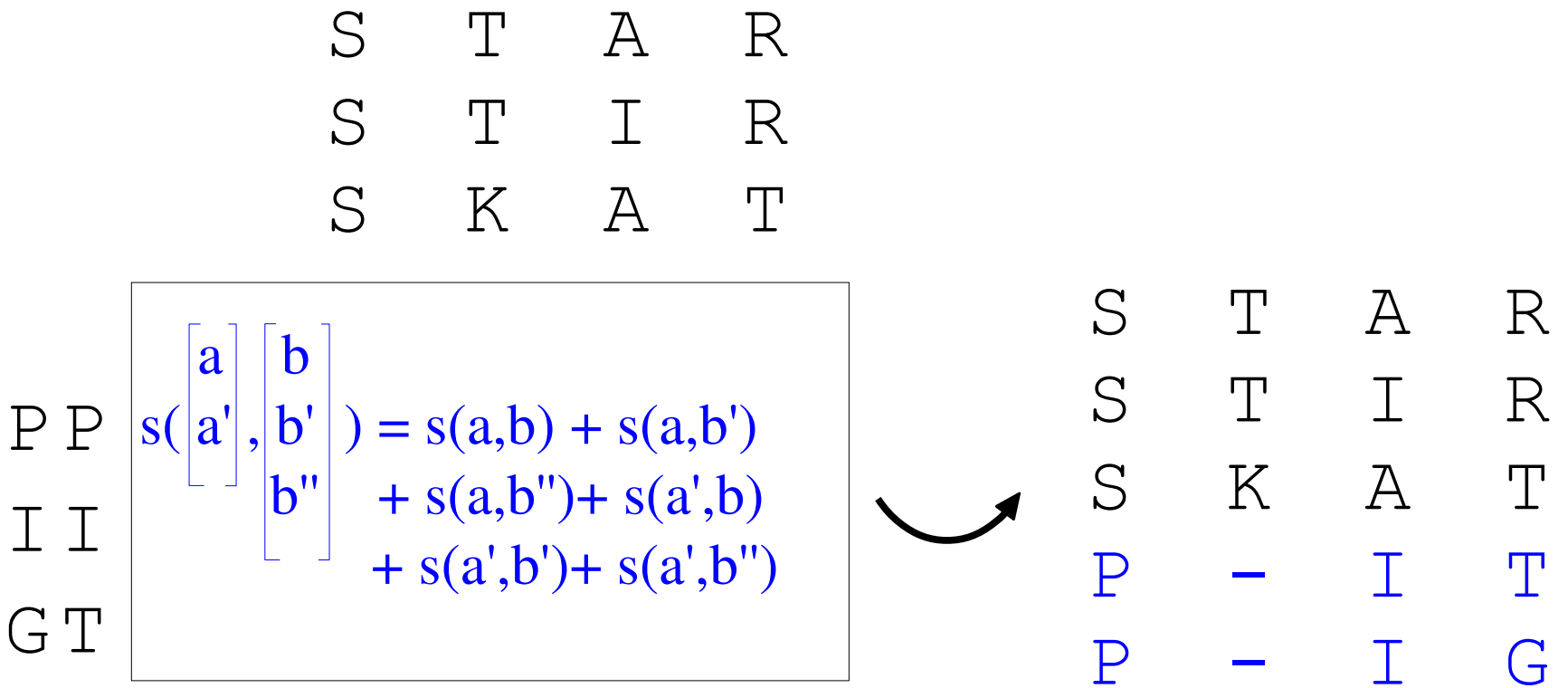
Sequence-“profile” alignment : dynamic programming

S T A R
S T I R
S K A T



S T A R
S T I R
S K I T
P - I T

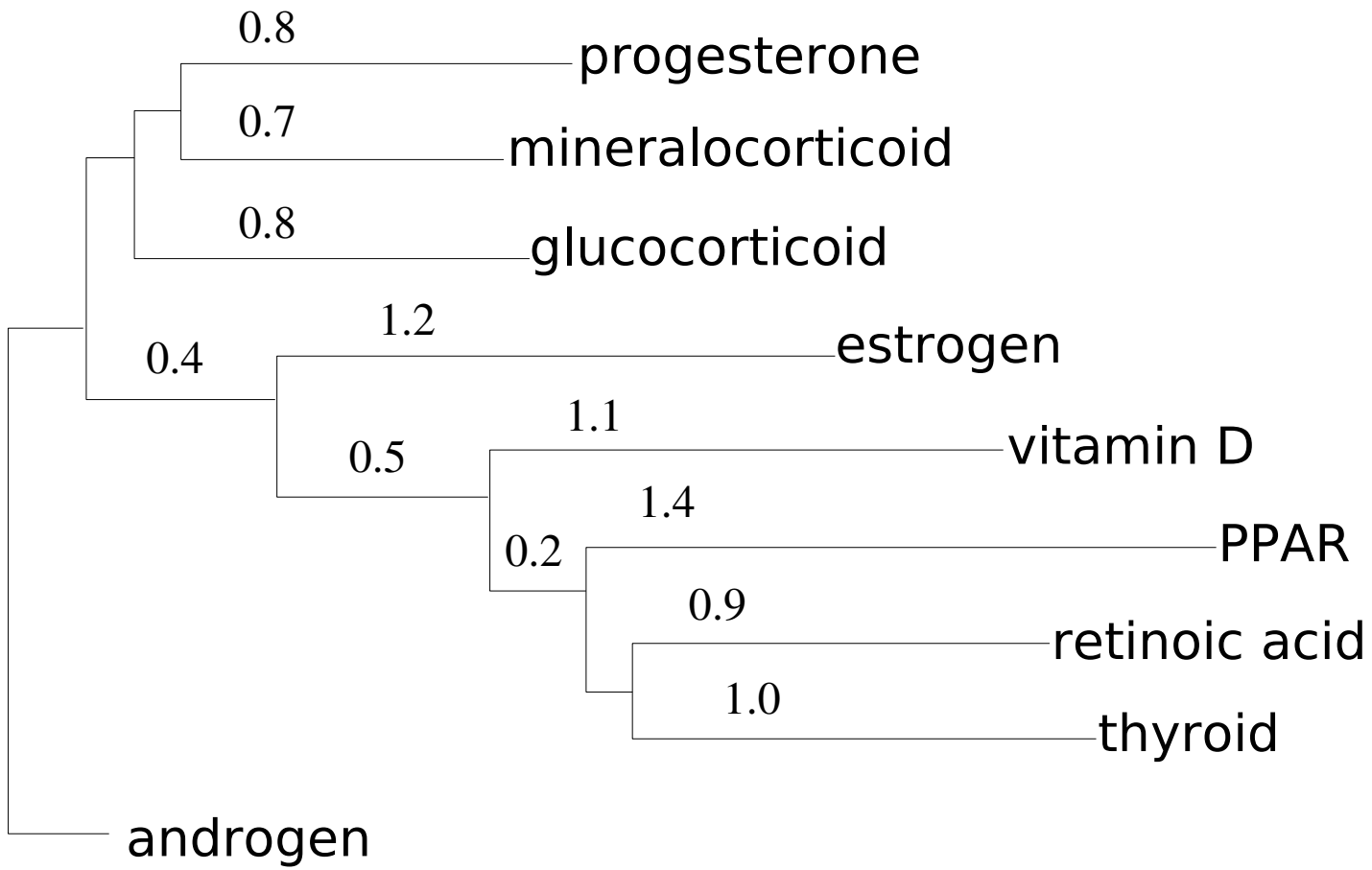
Aligning two alignments: “profile-profile” alignment



Sum over pairs

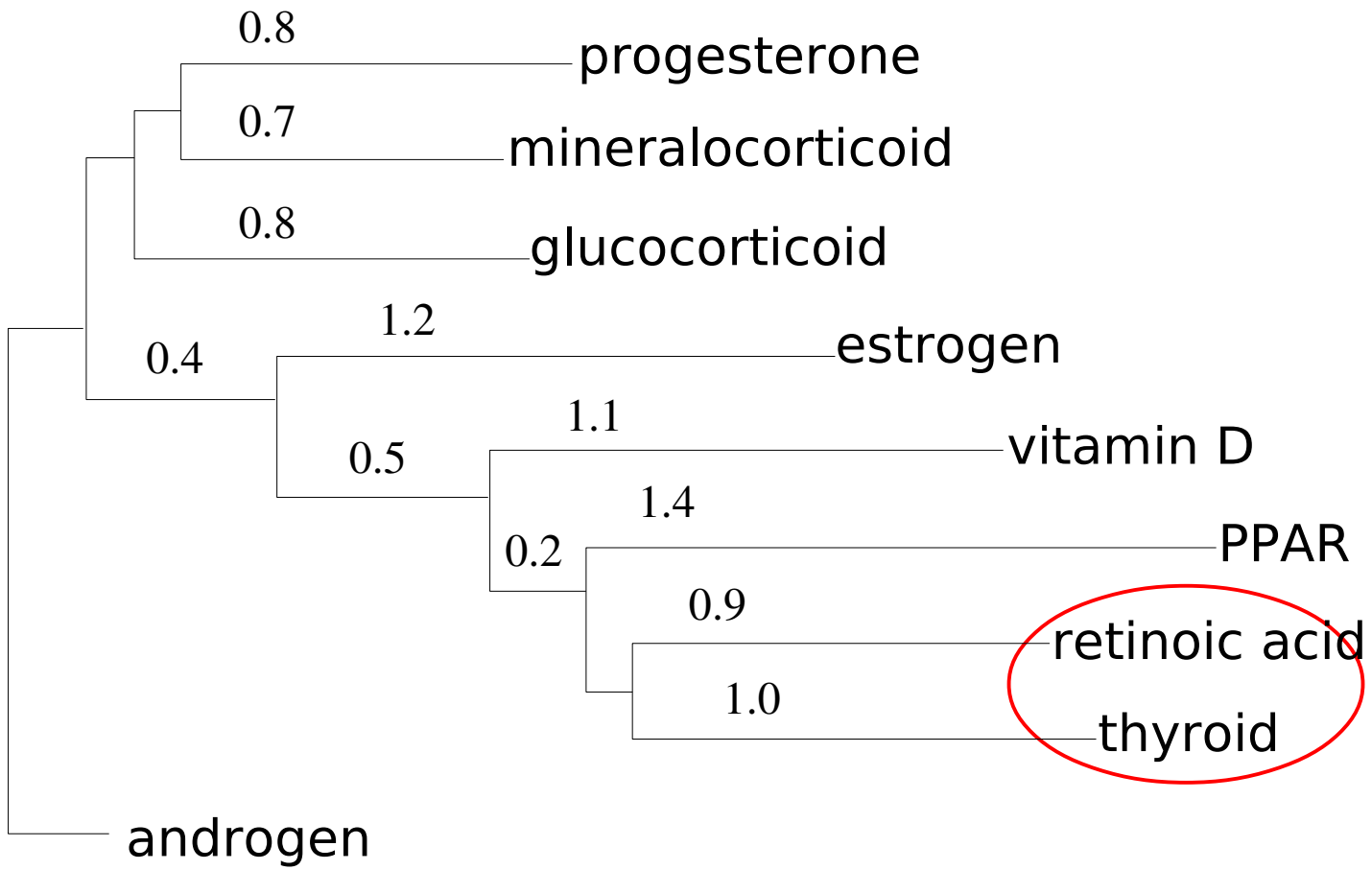
slight generalization of sequence alignment

Stage 3: align the sequences in the order of the guide tree



Branch lengths measure distances between sequences

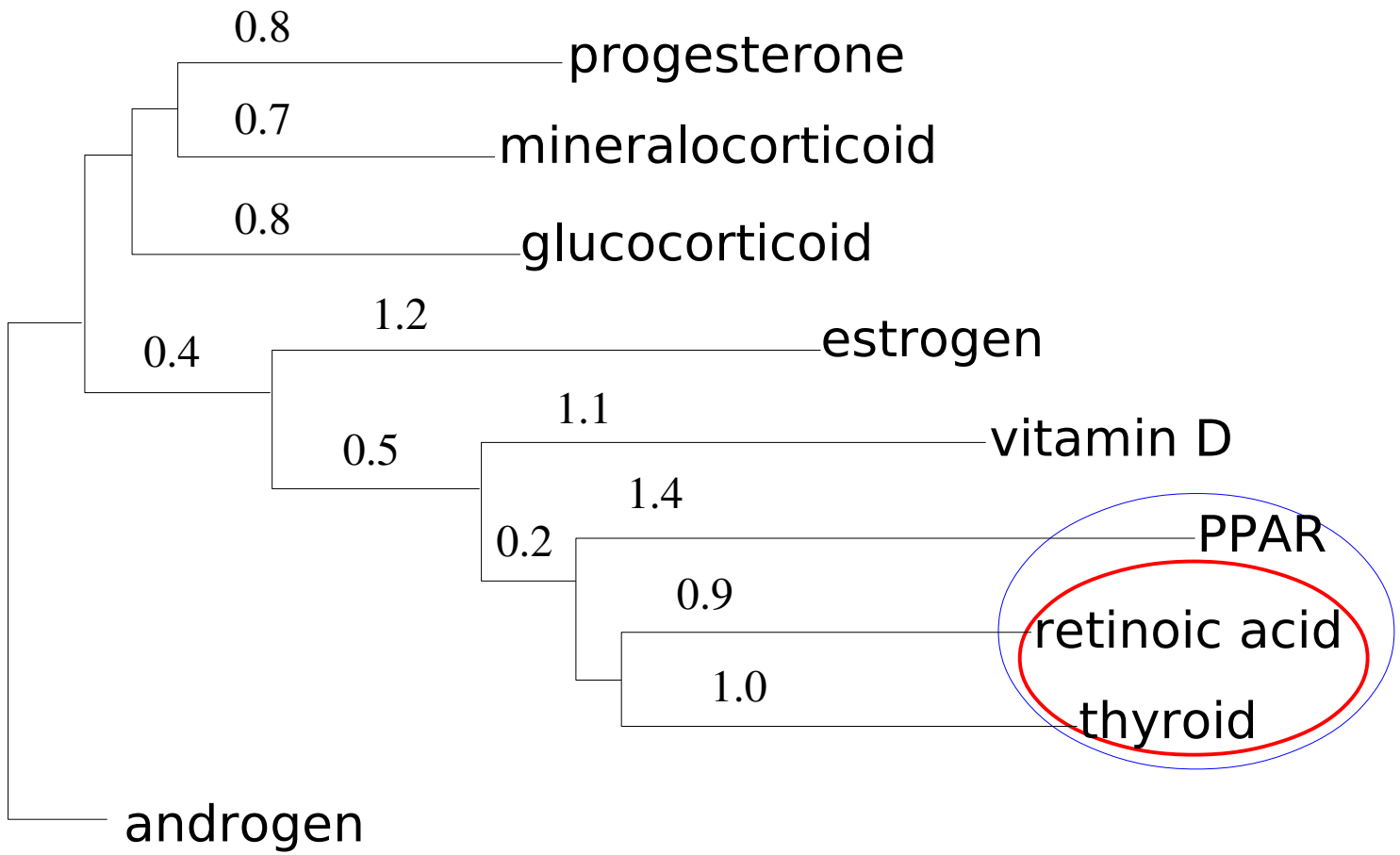
Stage 3: align the sequences in the order of the guide tree



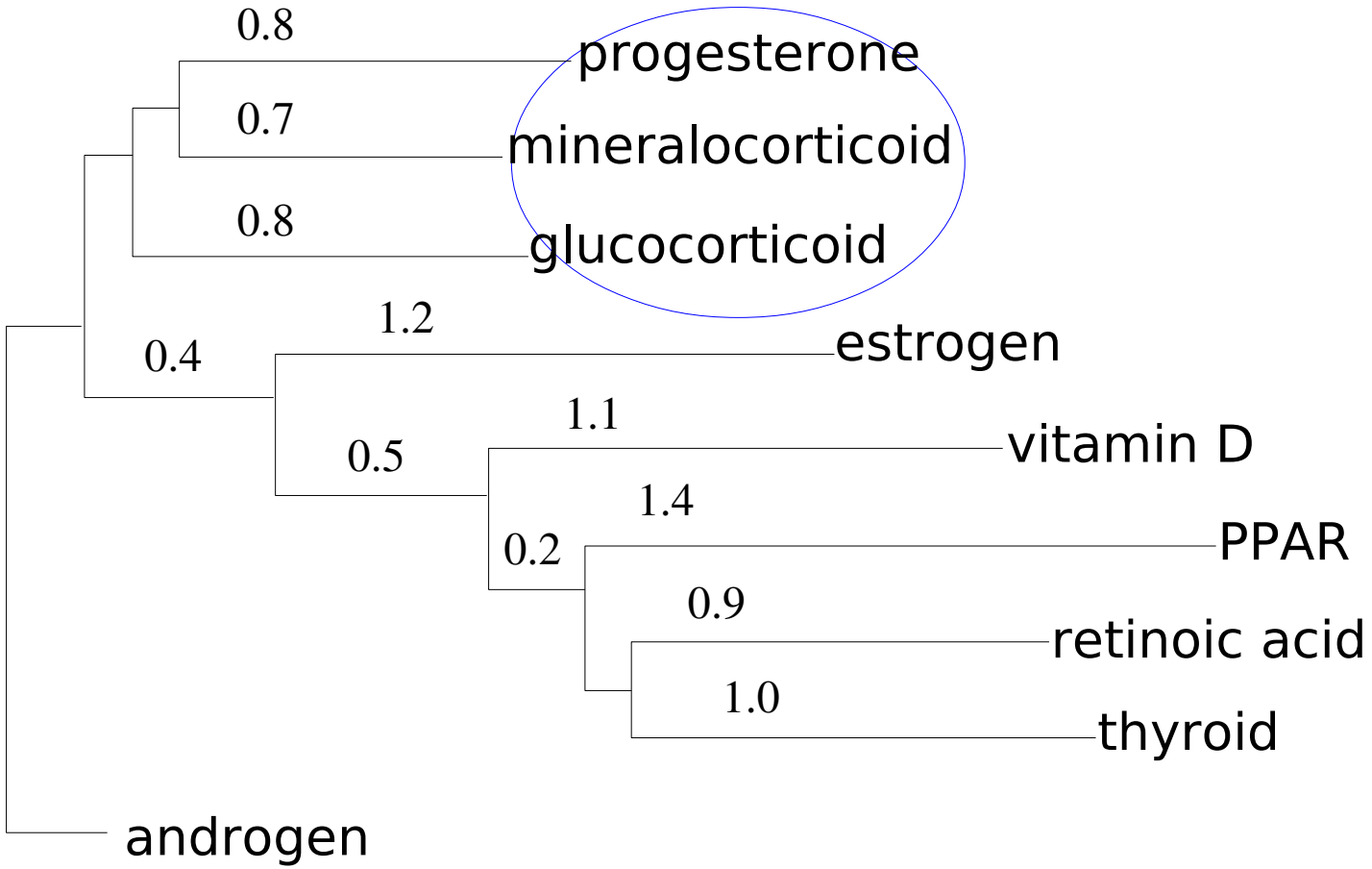
distance
= 0.9 + 1.0

Branch lengths measure distances between sequences

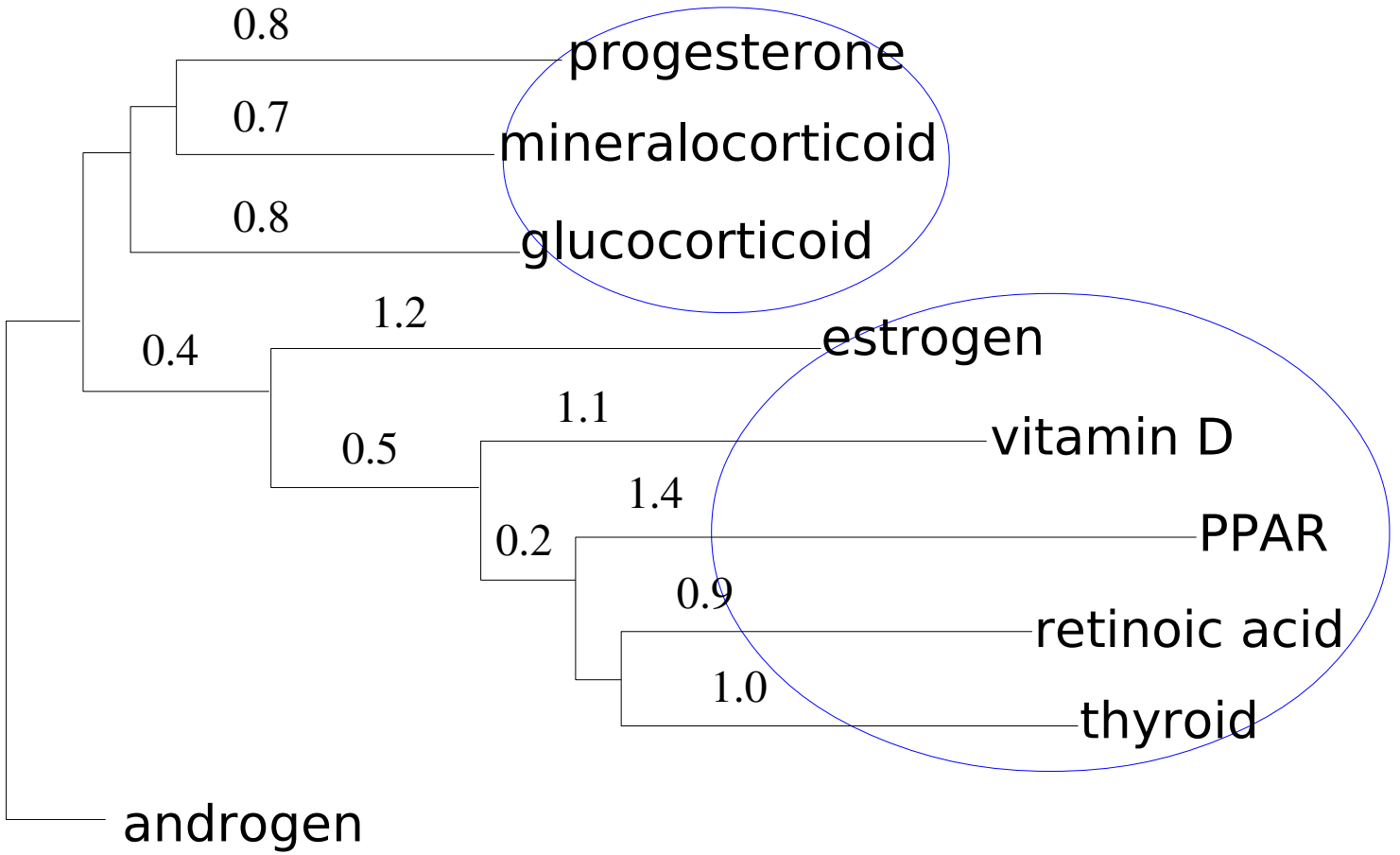
Stage 3: align the sequences in the order of the guide tree



Stage 3: align the sequences in the order of the guide tree

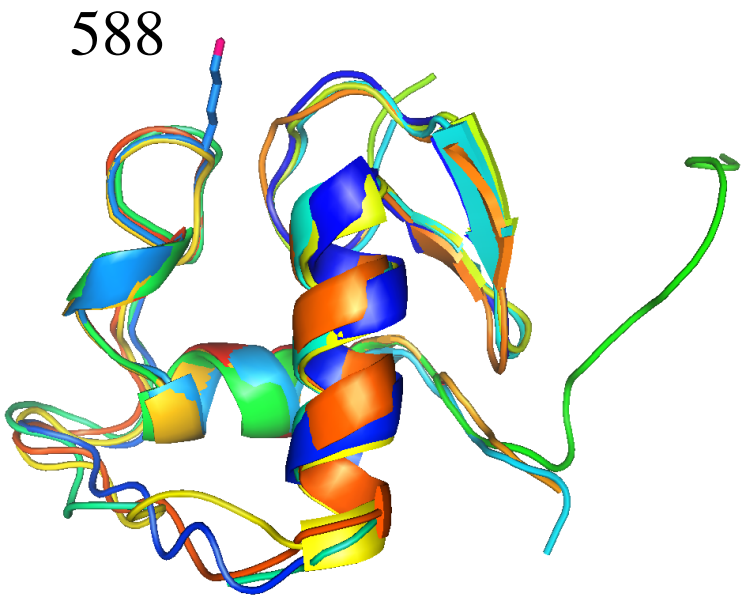


Stage 3: align the sequences in the order of the guide tree

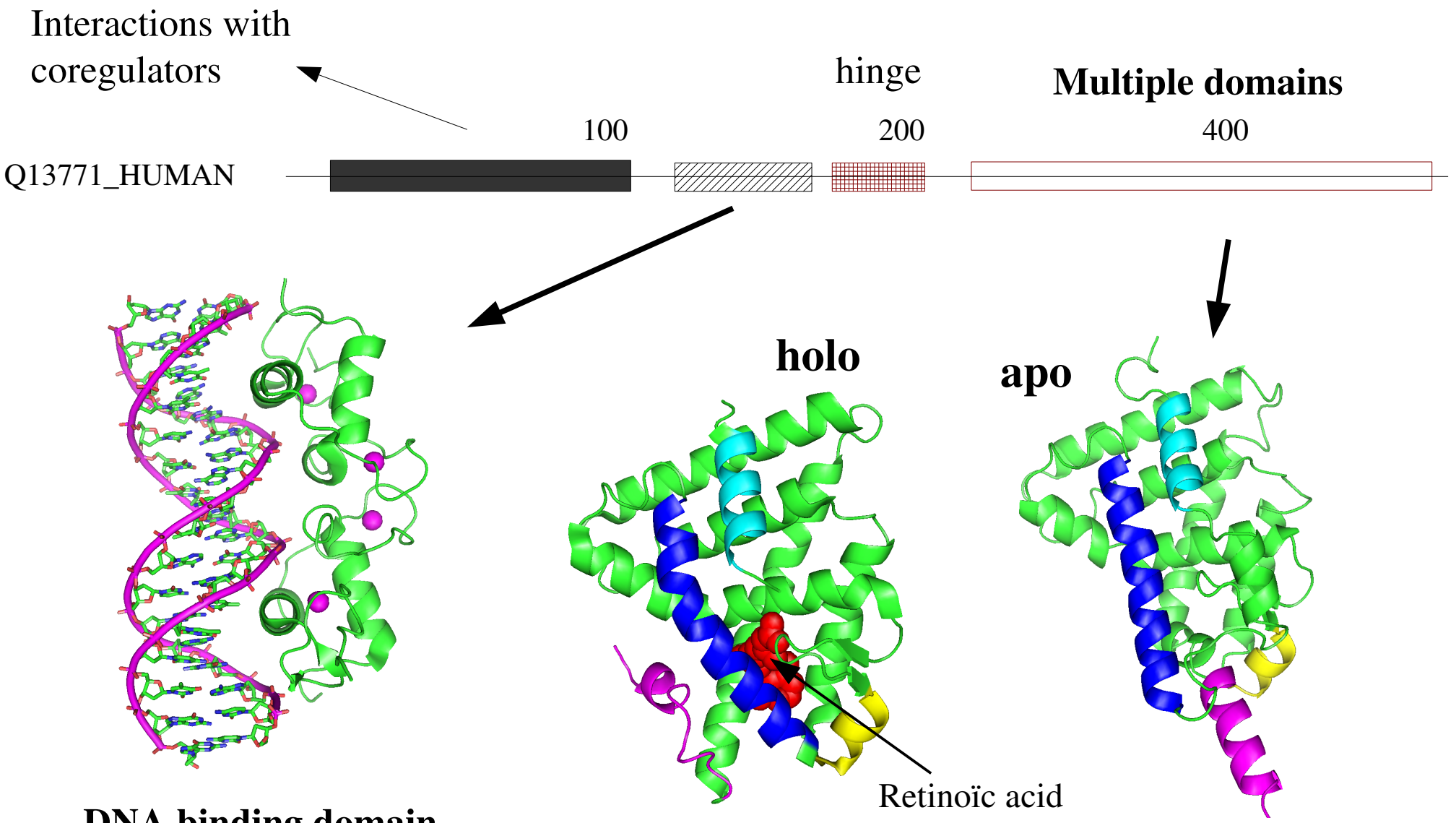


The final result

```
                                                584   588
androgen            VFFKRAAEG--KQKYLCASRNDCTIDKFRRKNCPSCRLRKCY
progesterone        VFFKRAVEG--HHNYLCAGRNDCIVDKIRRKNCPACRLRKCY
Mineralocorticoïd   VFFKRAVEG--QHNYLCAGRNDCIIDKIRRKNCPACRLQKCL
Glucocorticoïd      VFFKRAVEG--QHNYLCAGRNDCIIDKIRRKNCPACRYRKCL
Estrogen            AFFKRSIQG--HNDYMCPATNQCTIDKNRRKSCQACRLRKCY
Retinoic acid       GFFRRSIQK--NMVYTCHRDKNCIINKVTRNRCQYCRLQKCF
Vitamin D3          GFFRRSMKR--KALFTCPFNGDCRITKDNRRHCQACRLKRCV
Thyroid             GFFRRTIQKNLHPTYSCKYDSCCVIDKITRNQCQLCRFKKCL
                     **:*: .   :  :       * : *  *. *  ** ::*
```

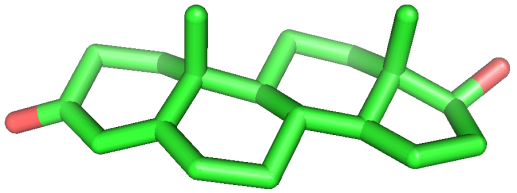


Nuclear receptors

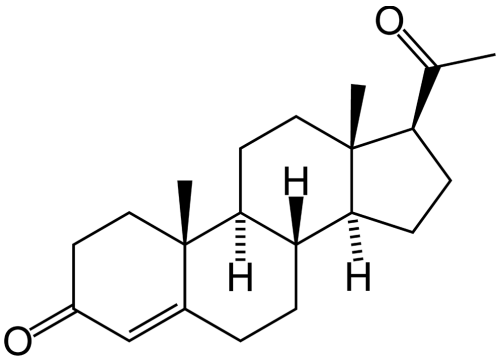


Nature, 2008, 456:350

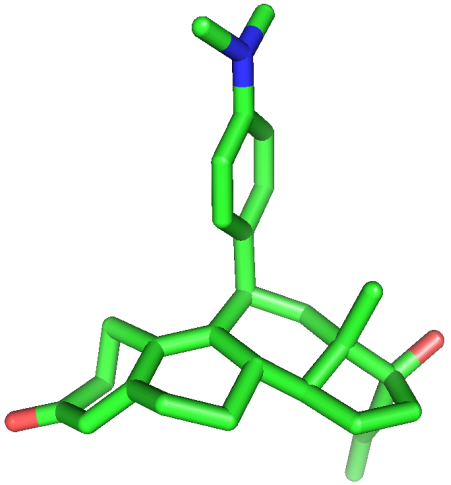
testosterone



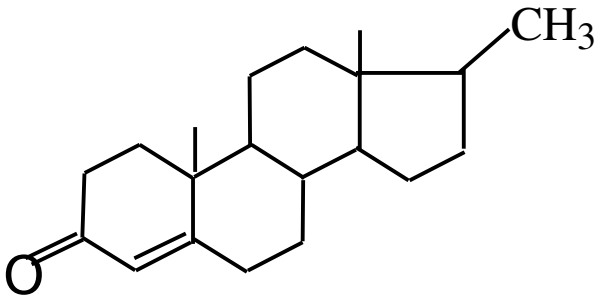
progesterone



mifepristone
or RU486
or Mifegyne™

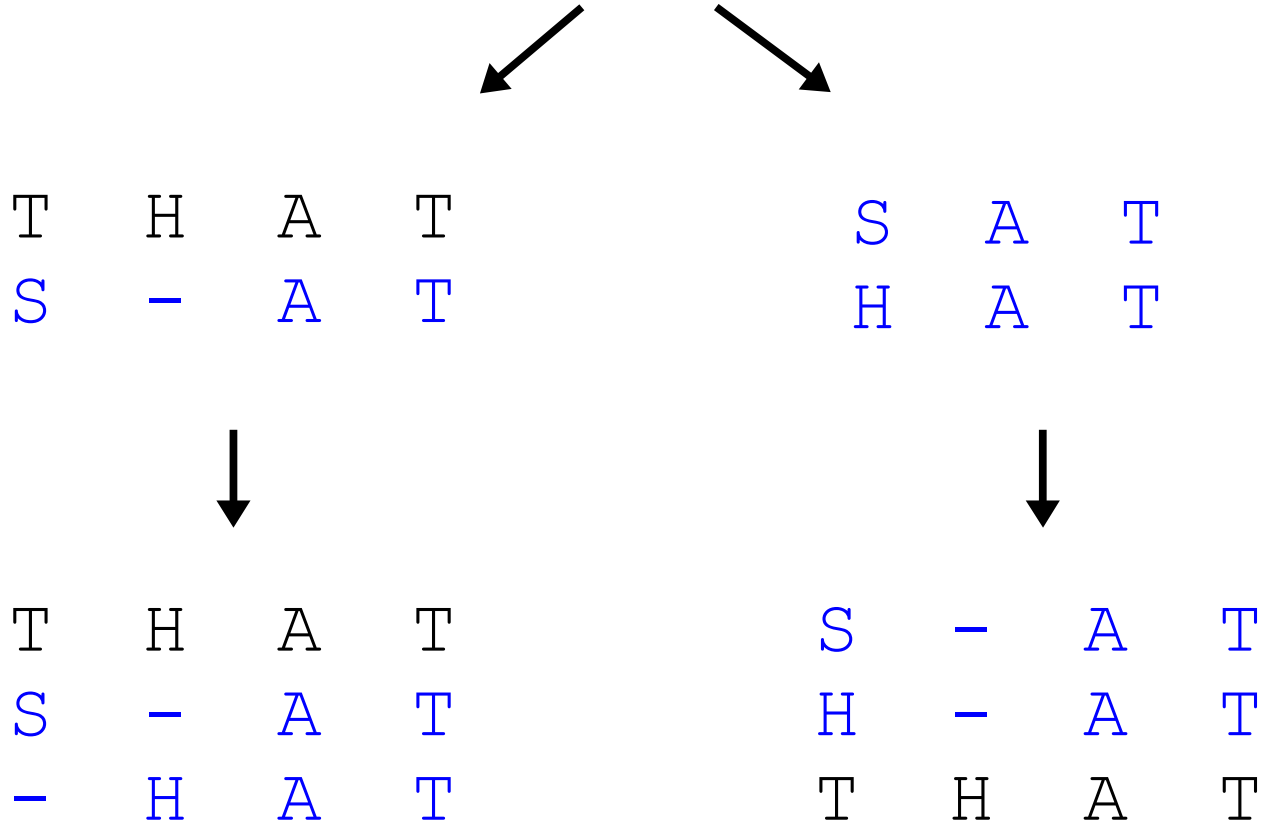


DMT



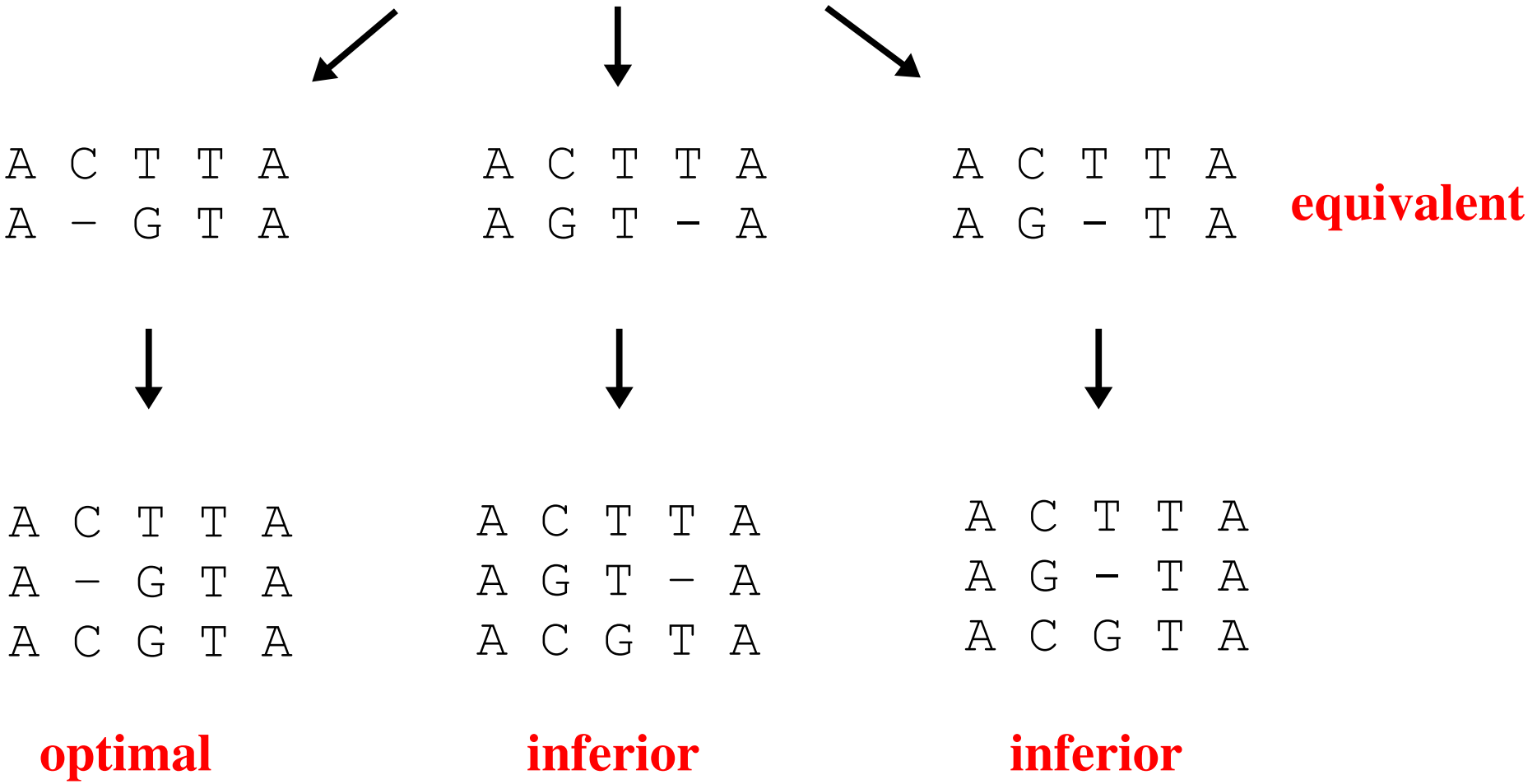
The final alignment depends on its "history"

Example 1: Fixation of gaps in the alignment



The final alignment depends on its "history"

Example 2: Fixation of gaps in the alignment



The pairwise sum approximation for the score has no simple probabilistic interpretation

Probabilistic score, two sequences: $\log [P(x_i,y_j)/q_{xi}\, q_{yj}]$

Probabilistic score, three sequences: $\log [P(x_i,y_j,z_k)/q_{xi}\, q_{yj}\, q_{zk}]$

$\neq \log [P(x_i,y_j)/q_{xi}\, q_{yj}] + \log [P(x_i,z_k)/q_{xi}\, q_{zk}] + \log [P(y_j,z_k)/q_{yj}\, q_{zk}]$

As if x descended from y AND z

There is no single parsimonious ancestor shared by all 3 pairs...

($(x_i U y_j) \cap (x_i U z_k) \cap (y_j U z_k) = \emptyset$ in general)

The pairwise sum approximation: tree-based weighting

Sequence to align:

H A D

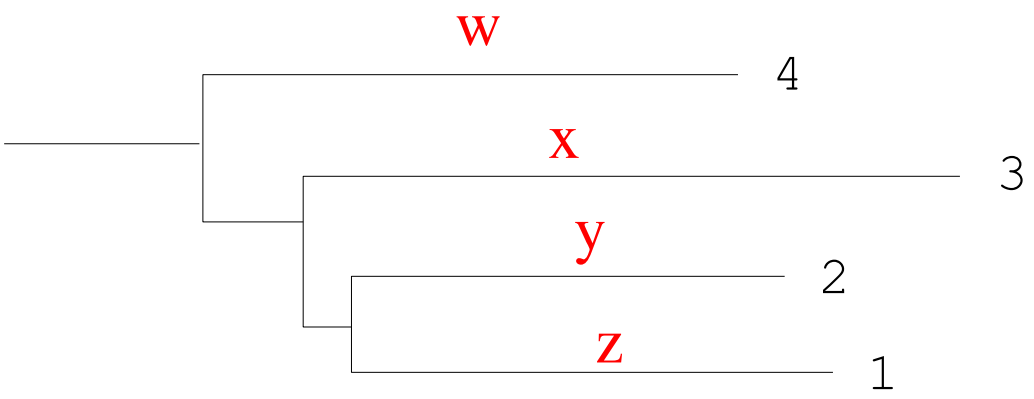
Existing alignment:

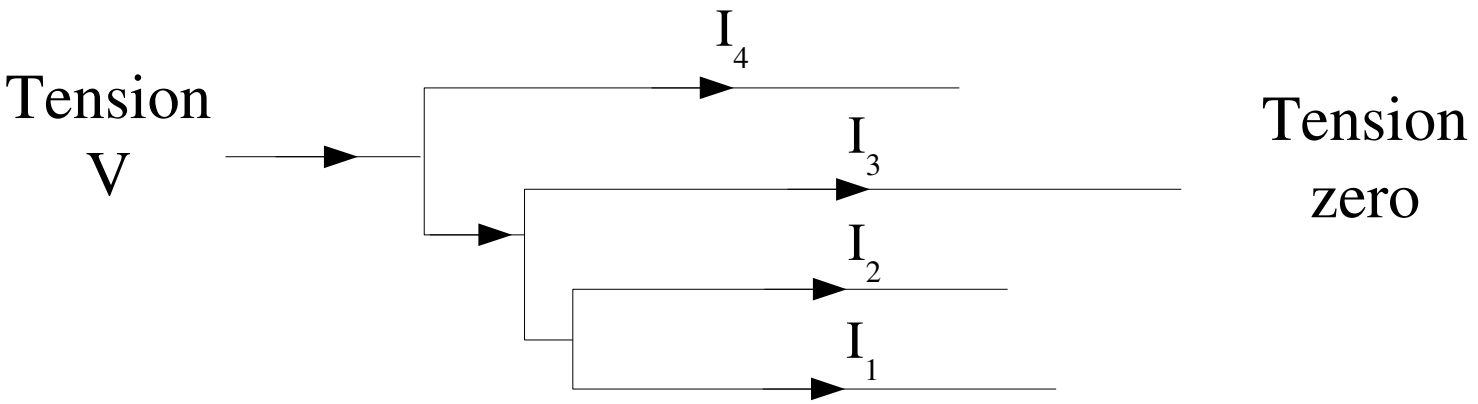
S T A R
S T I R
P - A R

	H	
S	-1	x 0.5
S	-1	x 0.5
P	-2	x 1
total	-3	

The two upper sequences are fairly redundant: reduced weight

Tree-based weighting: an... electric method



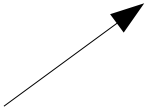


Resistances = branch lengths (w,x,y,z)

Weight of a leaf = electric current

Where should the list of homologues stop?

#	ID Swissprot	Name	Description	Score	E	% Identity	Length aligned
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor alpha	98	4e-21	55	72
:	:	:	: : :	:	:	:	
:	:	:	: : :	:	:	:	
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47



The E-value: see exercise

The procedure above is commonly used, and implemented in a standard program: ClustalW

Main refinements used in ClustalW and elsewhere:

Ideas?

Multiple sequence alignment is still a research topic

The procedure above is commonly used, and implemented in a standard program: ClustalW

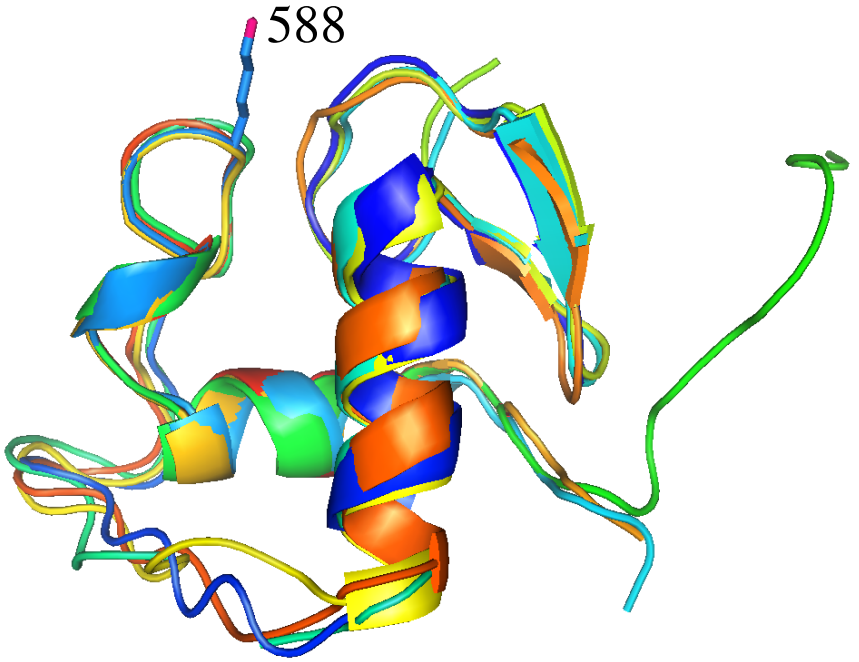
Main refinements used in ClustalW and elsewhere:

- **position-dependent gap scores**
- **use of different scoring matrices in different parts of the tree**
- **schemes to iteratively reshuffle and refine the alignment**

Multiple sequence alignment is still a research topic

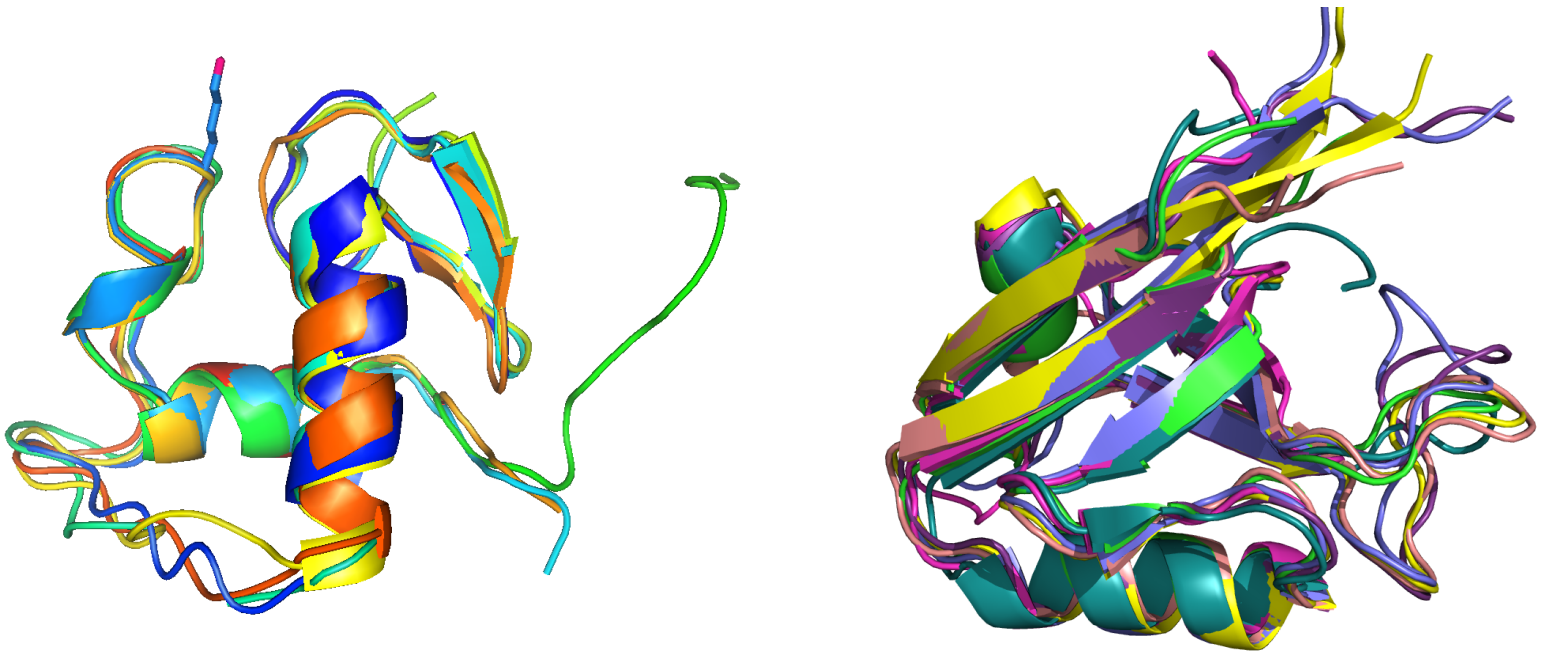
Experimental testing is needed

- Site-directed mutagenesis of conserved residues
- Detecting an interaction with a substrate or inhibitor
- Determination of the 3D-structure!

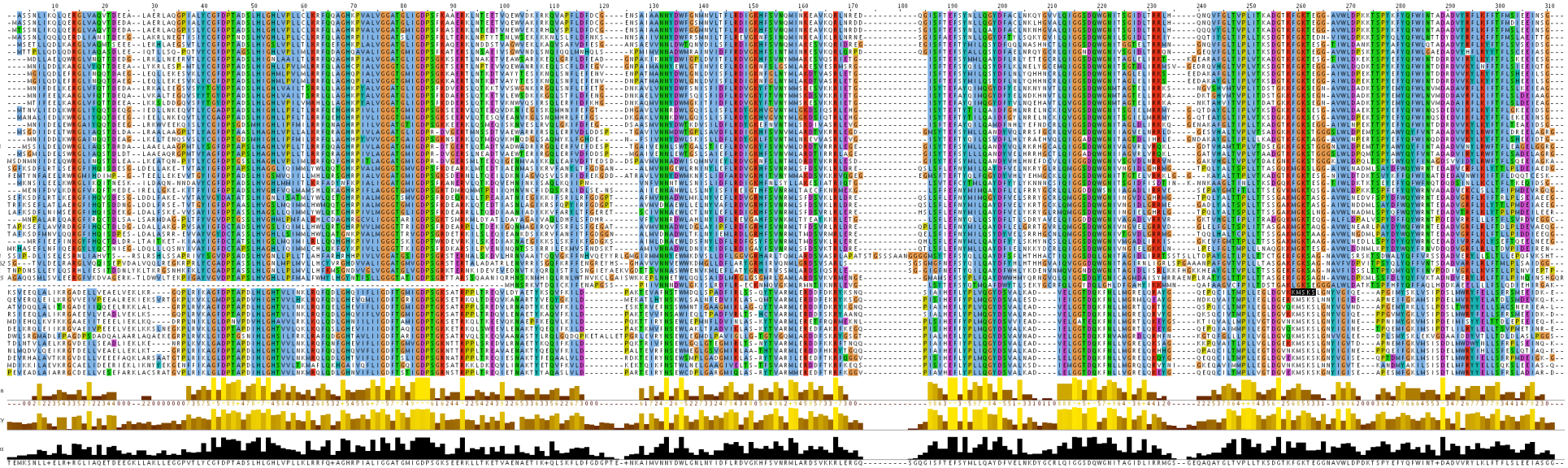


Alignment as a structural model

- Proteins have a specific 3D structure, stabilized by the physical properties of their sequences
- An alignment can be seen as a physical model without explicit evolutionary content: the “structure alignment” problem
- We can adopt this view even if the structures are unknown



Alignment of 30 tyrosyl-tRNA synthetase catalytic domains



Database of protein alignments: one per structural family

http://pfam.sanger.ac.uk

Multiple alignments: some applications

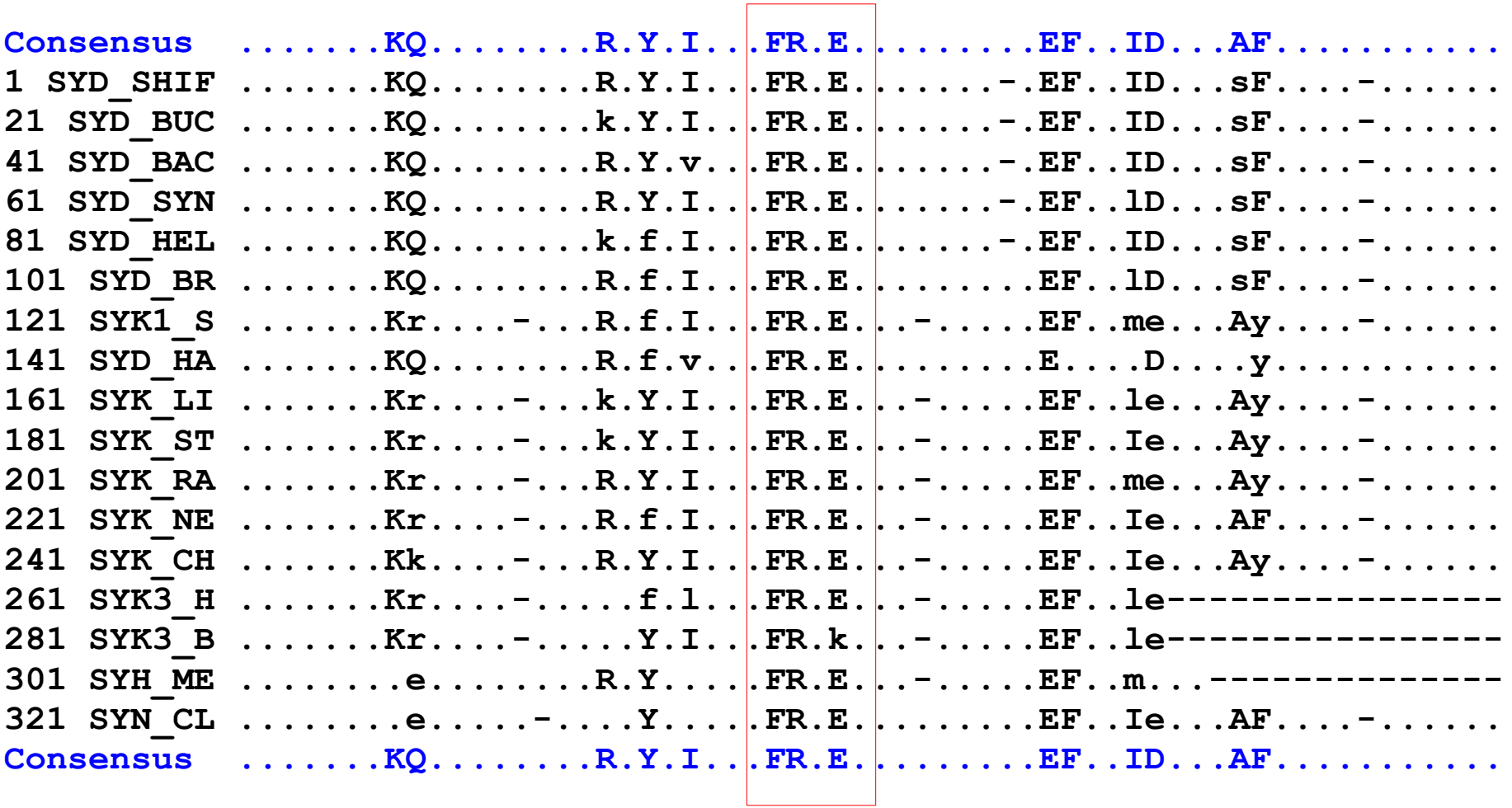
Identifying conserved and thus functional residues

Alignment of 340 aminoacyl-ARNt synthetases,
showing conserved residues only

```
Consensus  .......KQ........R.Y.I...FR.E.........EF..ID...AF..........
1 SYD_SHIF .......KQ........R.Y.I...FR.E.......-.EF..ID...sF....-.....
21 SYD_BUC .......KQ........k.Y.I...FR.E.......-.EF..ID...sF....-.....
41 SYD_BAC .......KQ........R.Y.v...FR.E.......-.EF..ID...sF....-.....
61 SYD_SYN .......KQ........R.Y.I...FR.E.......-.EF..lD...sF....-.....
81 SYD_HEL .......KQ........k.f.I...FR.E.......-.EF..ID...sF....-.....
101 SYD_BR .......KQ........R.f.I...FR.E.........EF..lD...sF....-.....
121 SYK1_S .......Kr....-...R.f.I...FR.E...-.....EF..me...Ay....-.....
141 SYD_HA .......KQ........R.f.v...FR.E.........E....D....y..........
161 SYK_LI .......Kr....-...k.Y.I...FR.E...-.....EF..le...Ay....-.....
181 SYK_ST .......Kr....-...k.Y.I...FR.E...-.....EF..Ie...Ay....-.....
201 SYK_RA .......Kr....-...R.Y.I...FR.E...-.....EF..me...Ay....-.....
221 SYK_NE .......Kr....-...R.f.I...FR.E...-.....EF..Ie...AF....-.....
241 SYK_CH .......Kk....-...R.Y.I...FR.E...-.....EF..Ie...Ay....-.....
261 SYK3_H .......Kr....-.....f.l...FR.E...-.....EF..le---------------
281 SYK3_B .......Kr....-.....Y.I...FR.k...-.....EF..le---------------
301 SYH_ME ........e........R.Y.....FR.E...-.....EF..m...-------------
321 SYN_CL ........e.....-....Y.....FR.E.........EF..Ie...AF....-.....
Consensus  .......KQ........R.Y.I...FR.E.........EF..ID...AF..........
```

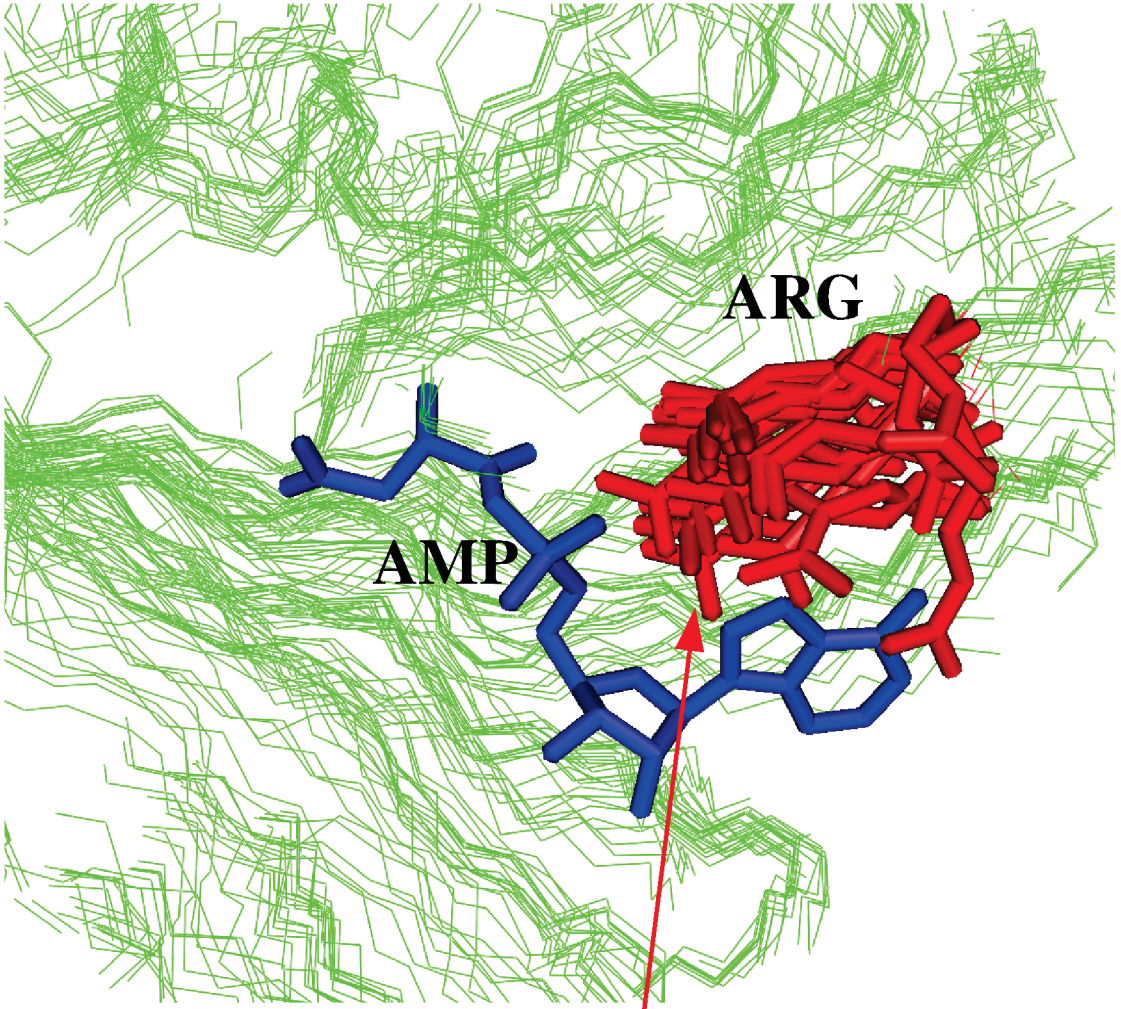
Identifying functional residues

Alignment of 340 aminoacyl-ARNt synthetases,
showing conserved residues only



Identifying functional residues

Structural alignment of 30 aminoacyl-ARNt synthetases

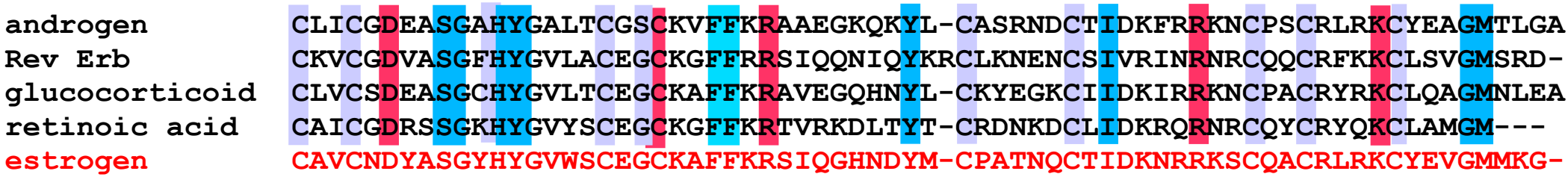


Consensus .....KQ.....R.Y.I...FR.E.....EF..ID...AF.....

Template-based 3D structure prediction

Estrogen receptor example

1) align its sequence with homologues, including at least one of known 3D structure: the template

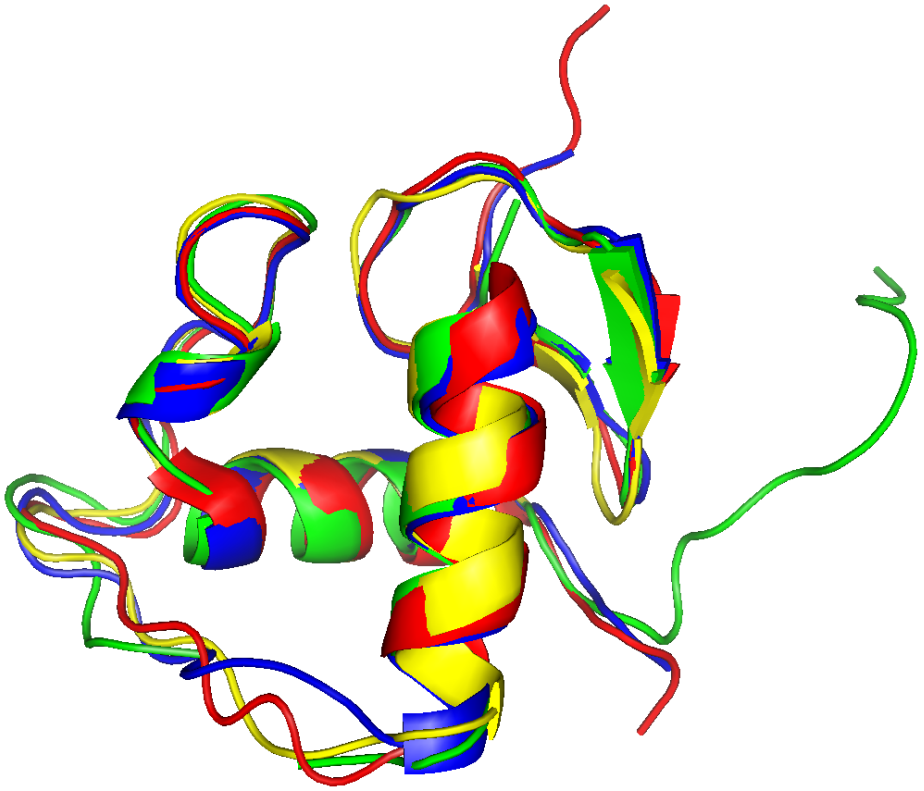


Template-based 3D structure prediction

Estrogen receptor example

2) adopt a sensible or consensus main chain trace

3) build sidechains

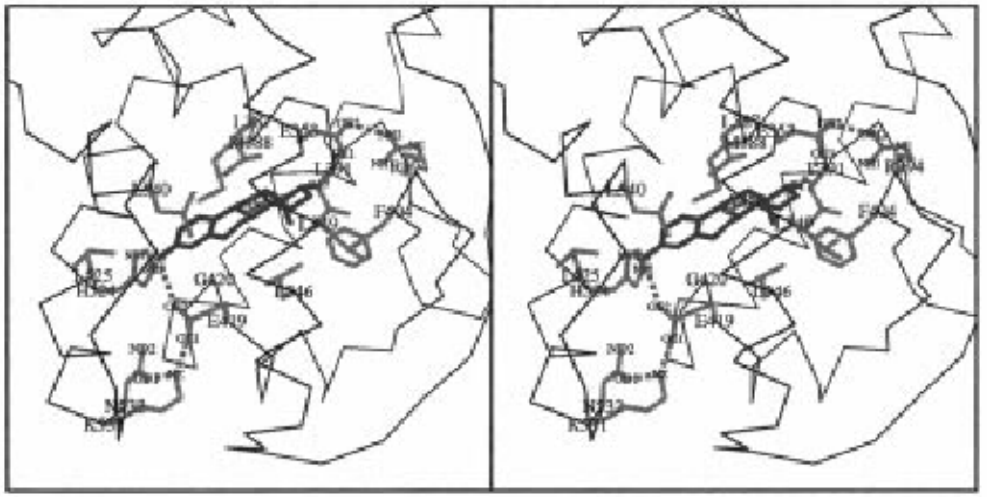


Template-based 3D structure prediction

Estrogen receptor example

2) adopt a consensus main chain trace

3) build sidechains



Wurtz et al (1998)
J Medicinal Chem

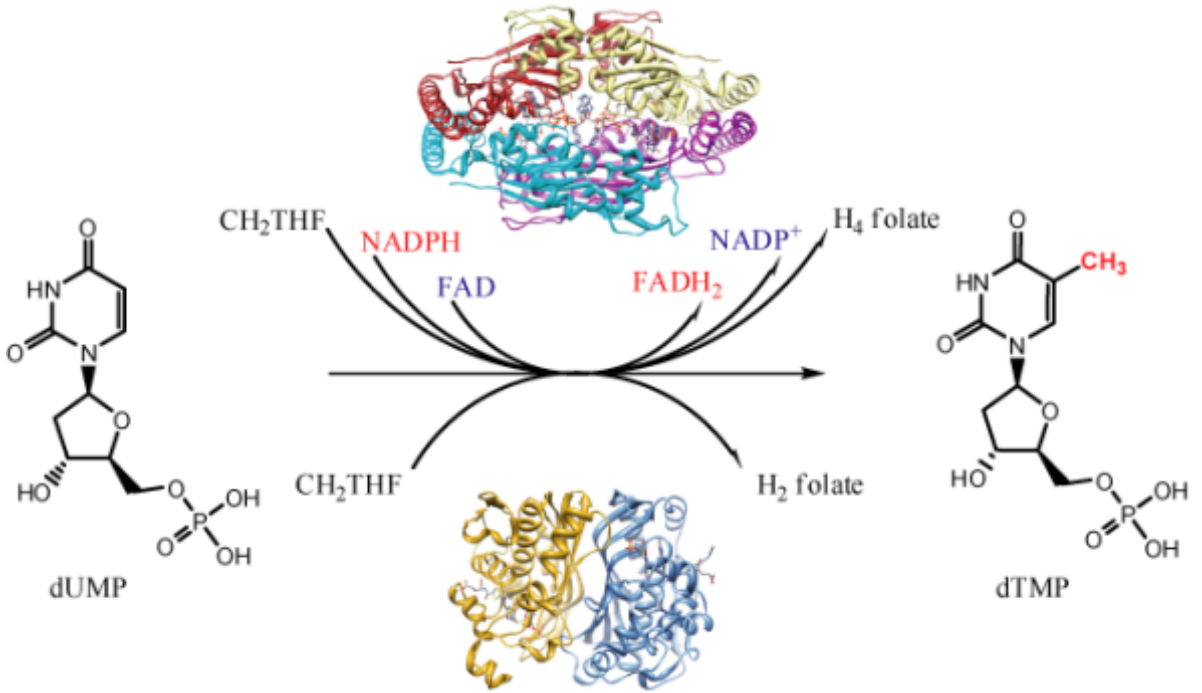
In this example, the model was good enough to explain estrogen binding

Identifying orthologues and possible drug targets

A strategy to attack a pathogen is to identify an essential protein that has no human orthologue, making it a potential drug target.

Identifying orthologues and possible drug targets

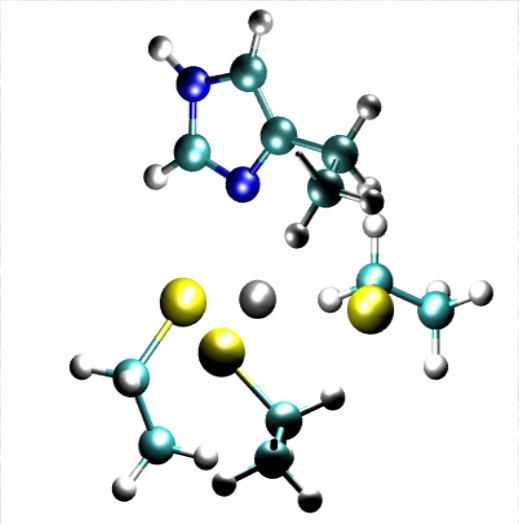
A strategy to attack a pathogen is to identify an essential protein that has no human orthologue, making it a potential drug target.



Distinct structures and chemical reactions in human thymidylate synthase ThyA (top) and the essential analogue ThyX (bottom) present in many pathogenic bacteria, including *Helicobacter pylori*

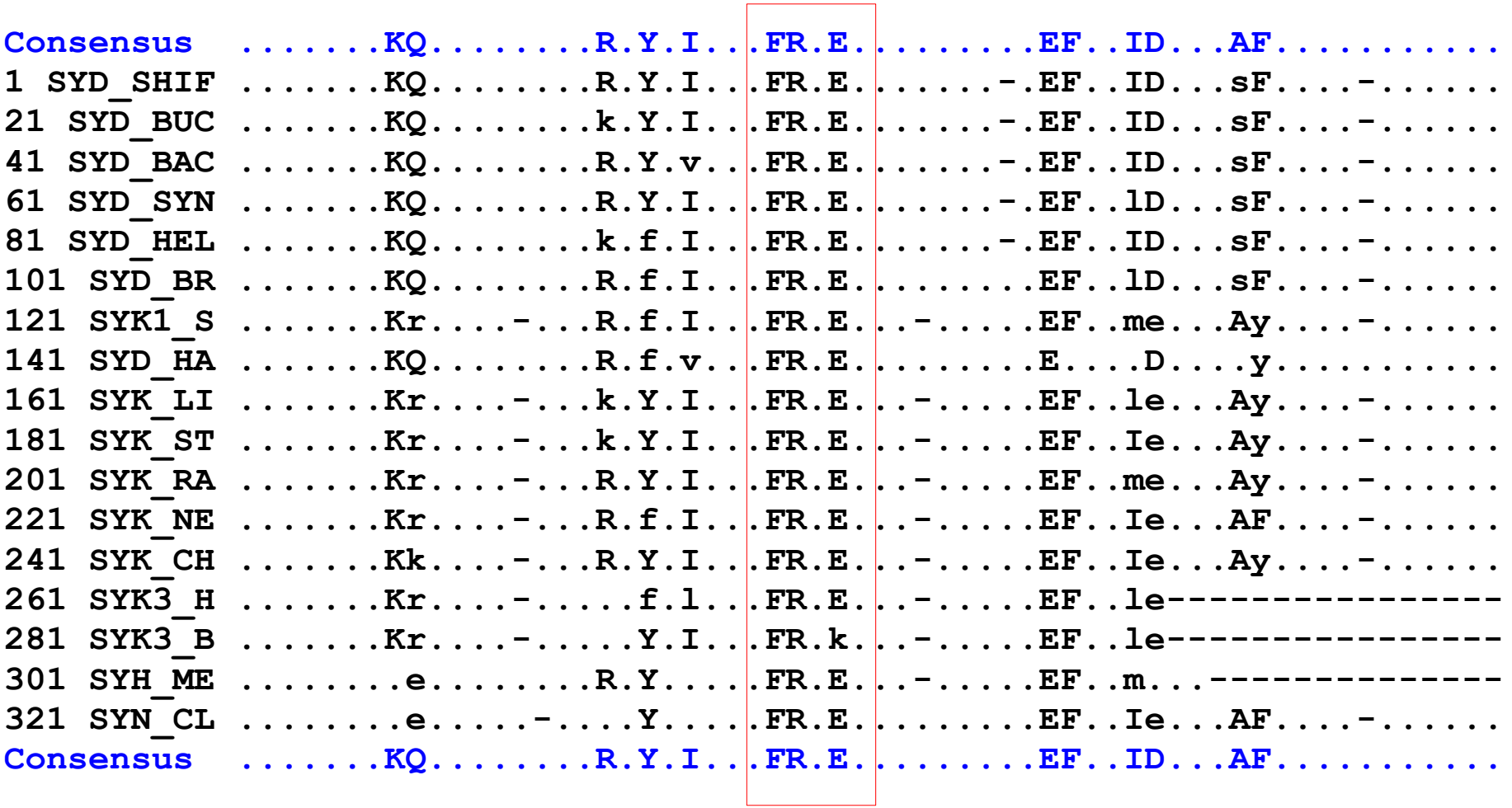
Science (2002) 297:105

Identifying patterns and functional sites



Identifying functional residues

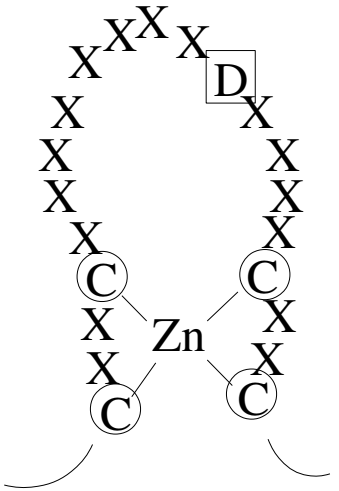
Alignment of 340 aminoacyl-ARNt synthetases,
showing conserved residues only



Divergence within the family may be too great to use a multiple alignment
→ structural, biochemical studies needed to infer functional pattern

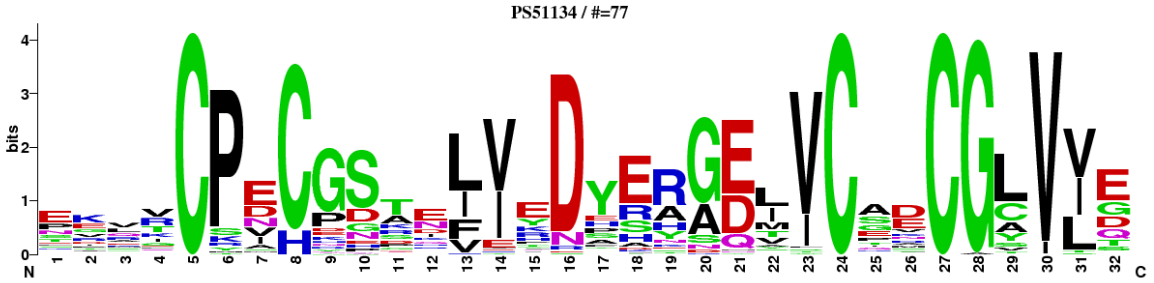
Another protein example:

zinc fingers (TFIIB class)



Database of conserved patterns in proteins: **ProSite**

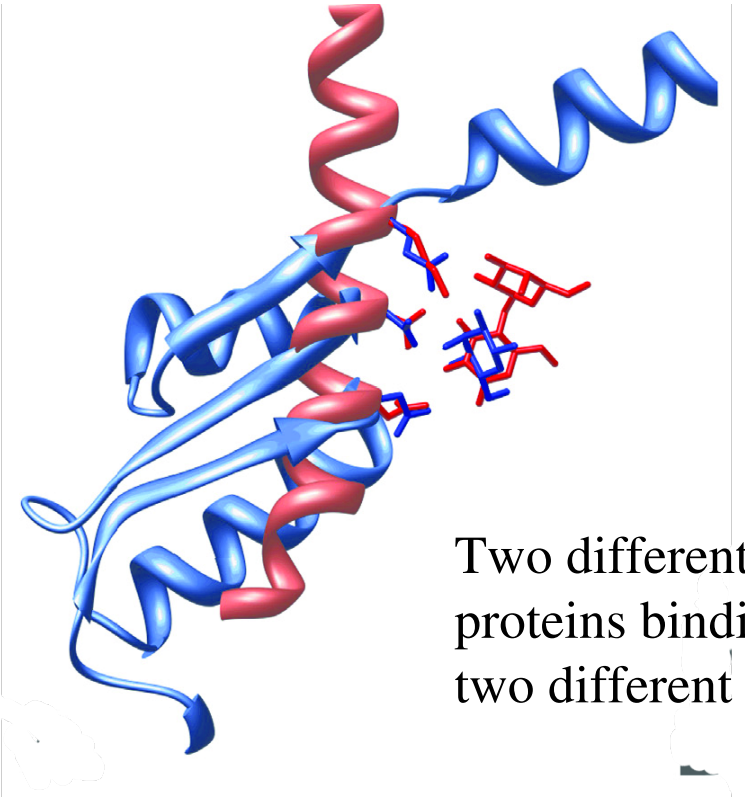
http://prosite.expasy.org



Database of conserved patterns in proteins: **ProSite**
http://prosite.expasy.org

Functional patterns sometimes arise from converent, not divergent evolution... so that a 3D pattern is relevant, not a 1D pattern

BMC Bioinformatics 2009, 10:182



Two different proteins binding two different sugars

Coding, signals and patterns in DNA

- Transcription signals
- Restriction sites
- Splicing sites
- Regulation sites
- Methylation sites

To detect short patterns, special techniques are needed

Restriction sites

Restriction enzymes cut DNA at specific sites

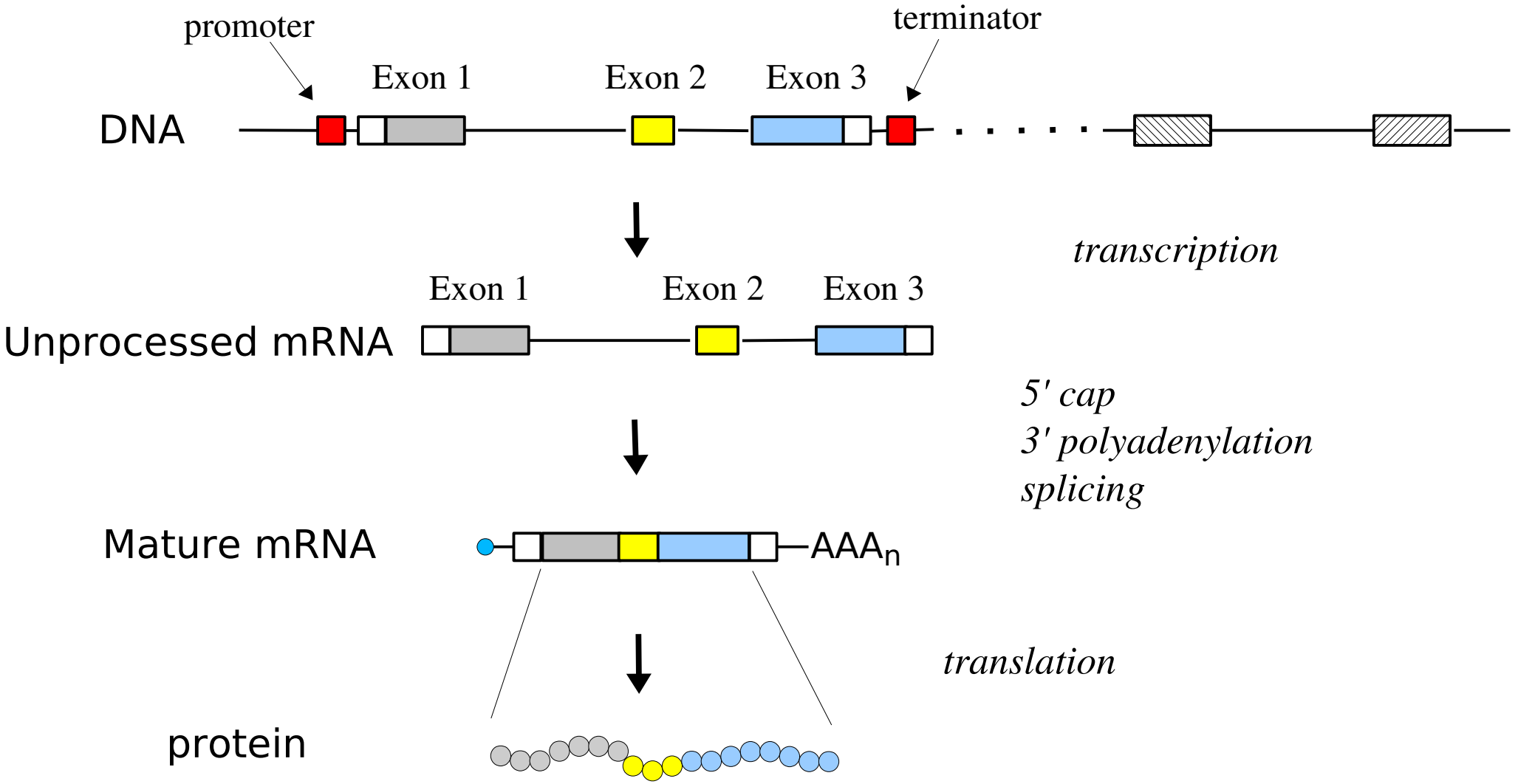
```
5' GAATTC 3'
3' CTTAAG 5'
```
EcoRI

```
5' GTTAAC 3'
3' CAATTG 5'
```
HpaI

```
5' GGATCC 3'
3' CCTAGG 5'
```
BamHI

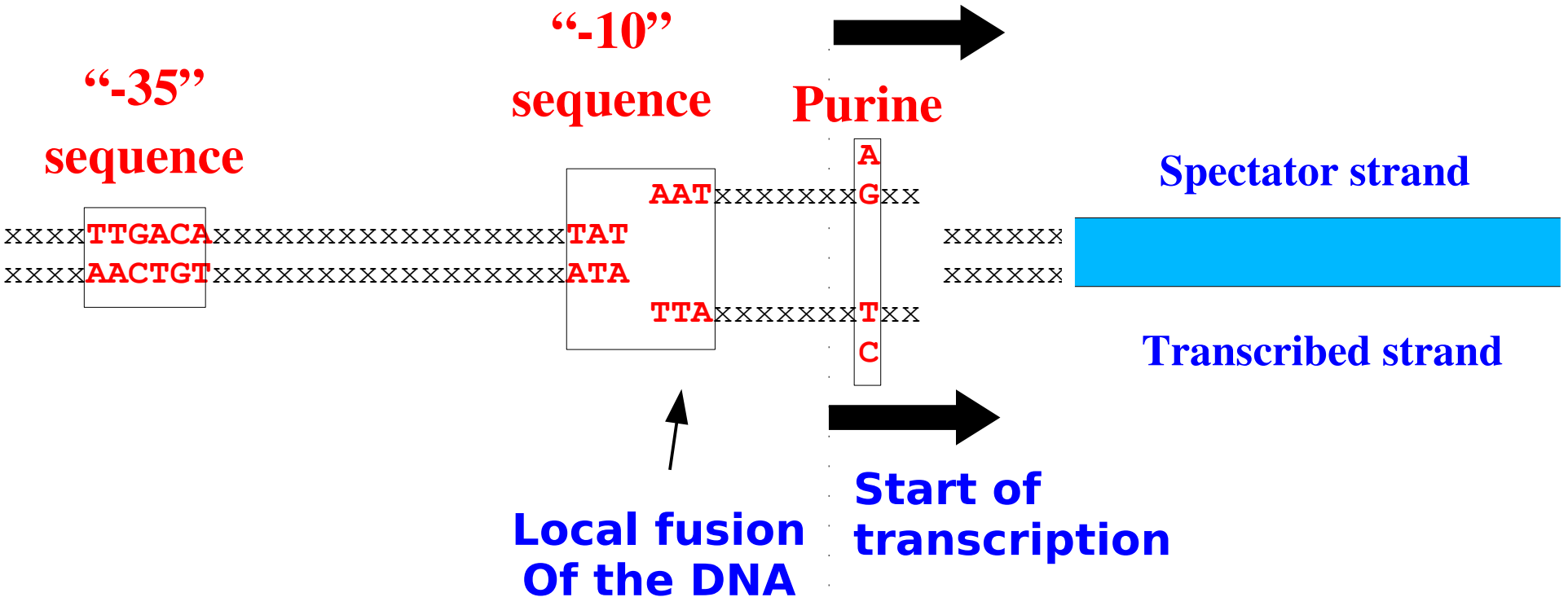
Antiviral defence; a tool for molecular biotechnology

Transcription signals: promoters



Transcription signals: promoters

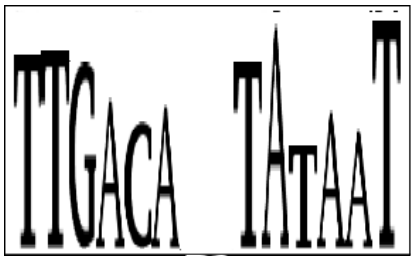
3 regions upstream of the START allow RNA polymerase binding and the initiation of transcription of an ORF (open reading frame): they form the **promoter**.

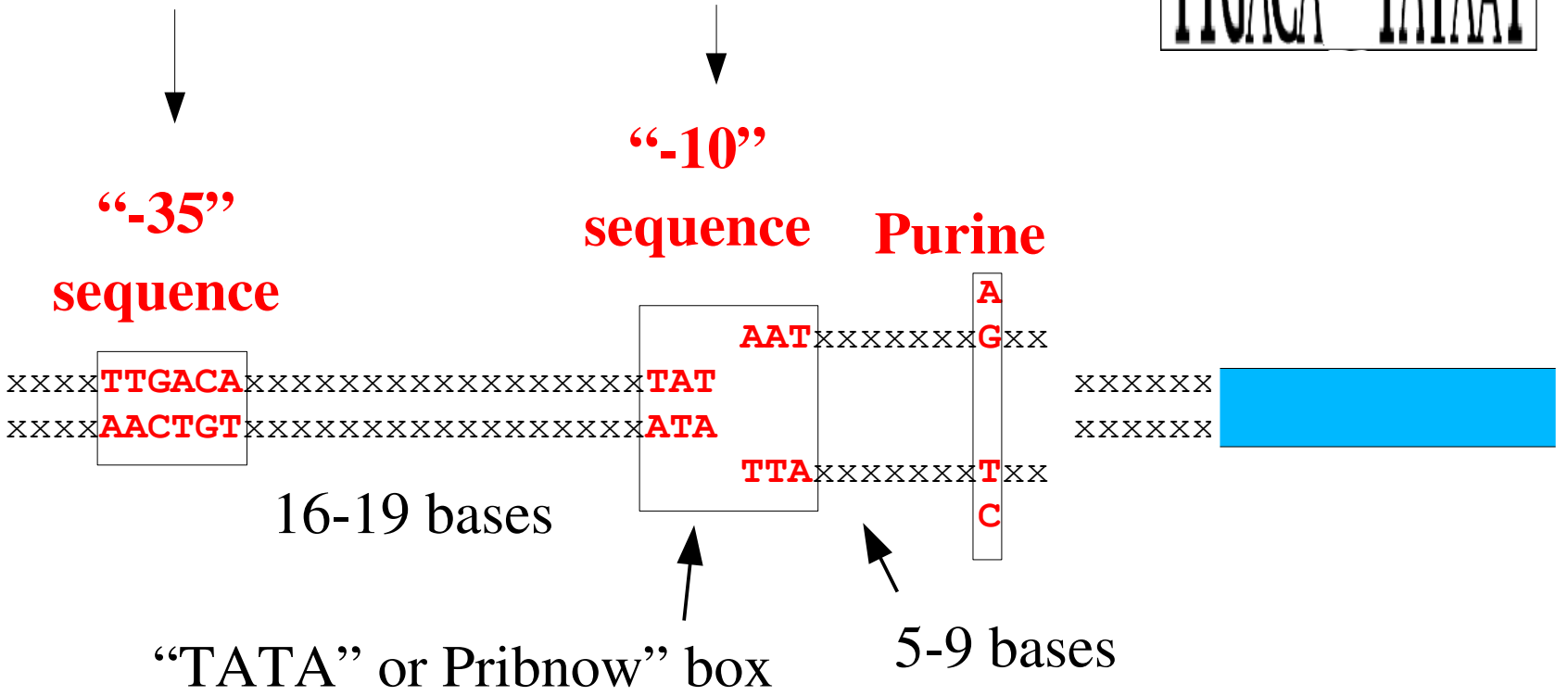


Transcription signals: promoters

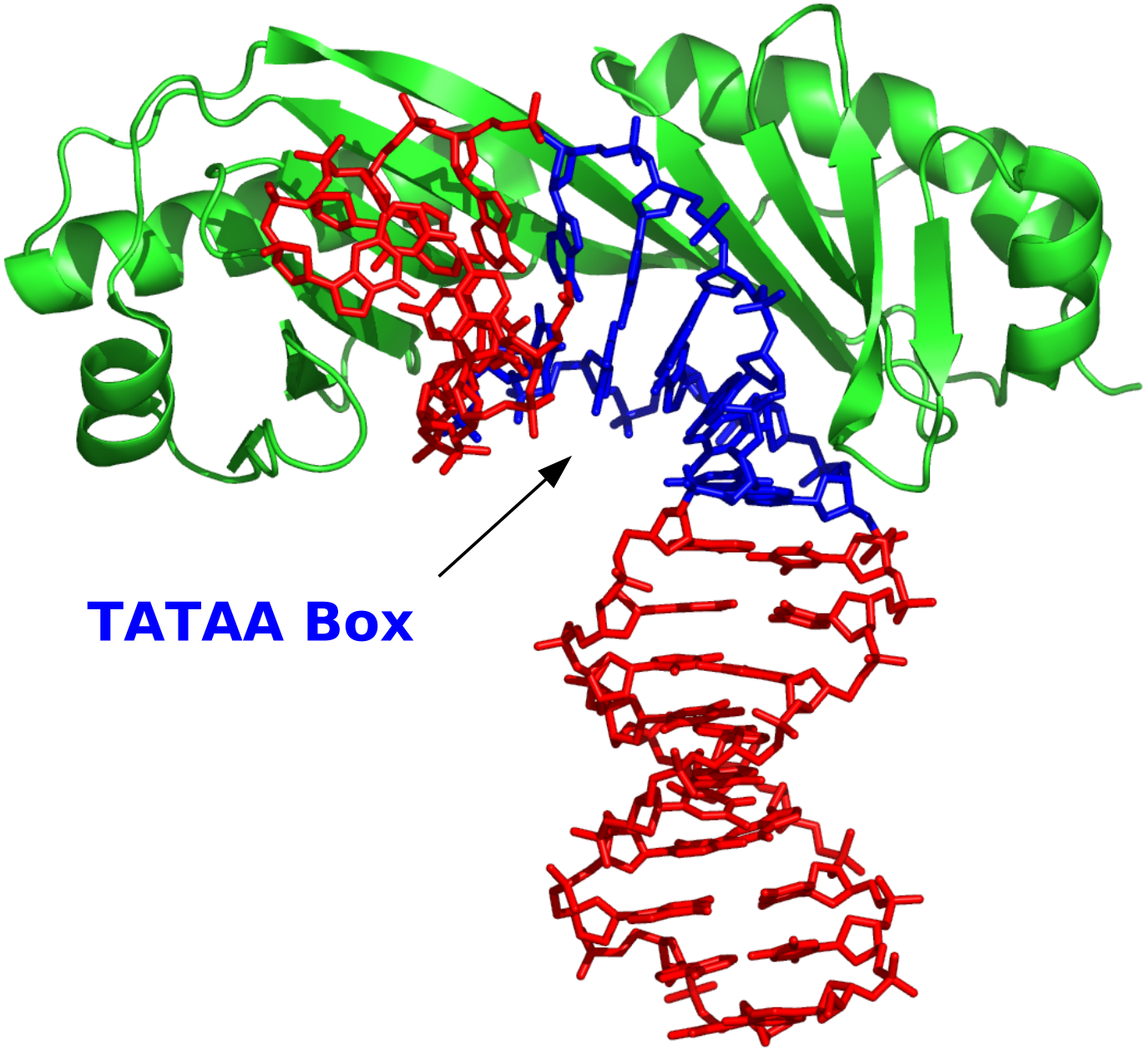
The conservation (%) of the -35 and -10 consensus sequences is:

$T_{82}T_{84}G_{78}A_{65}C_{54}A_{45}$ $T_{80}A_{95}T_{45}A_{60}A_{50}T_{96}$



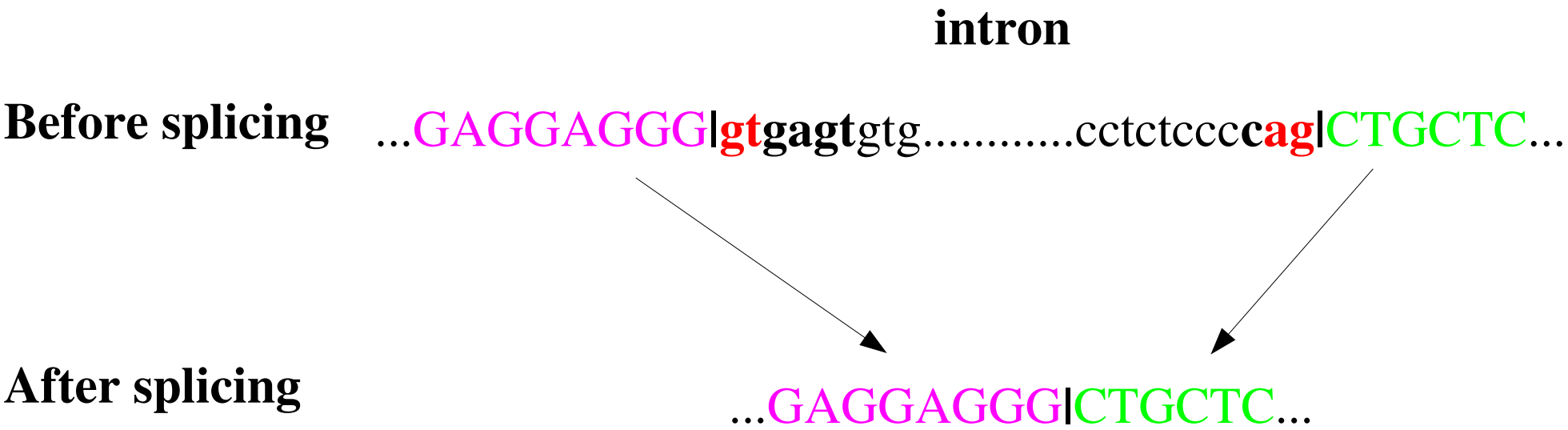


TATAA Binding Protein (TBP)



Nature, 2003, 422:534

Splicing sites



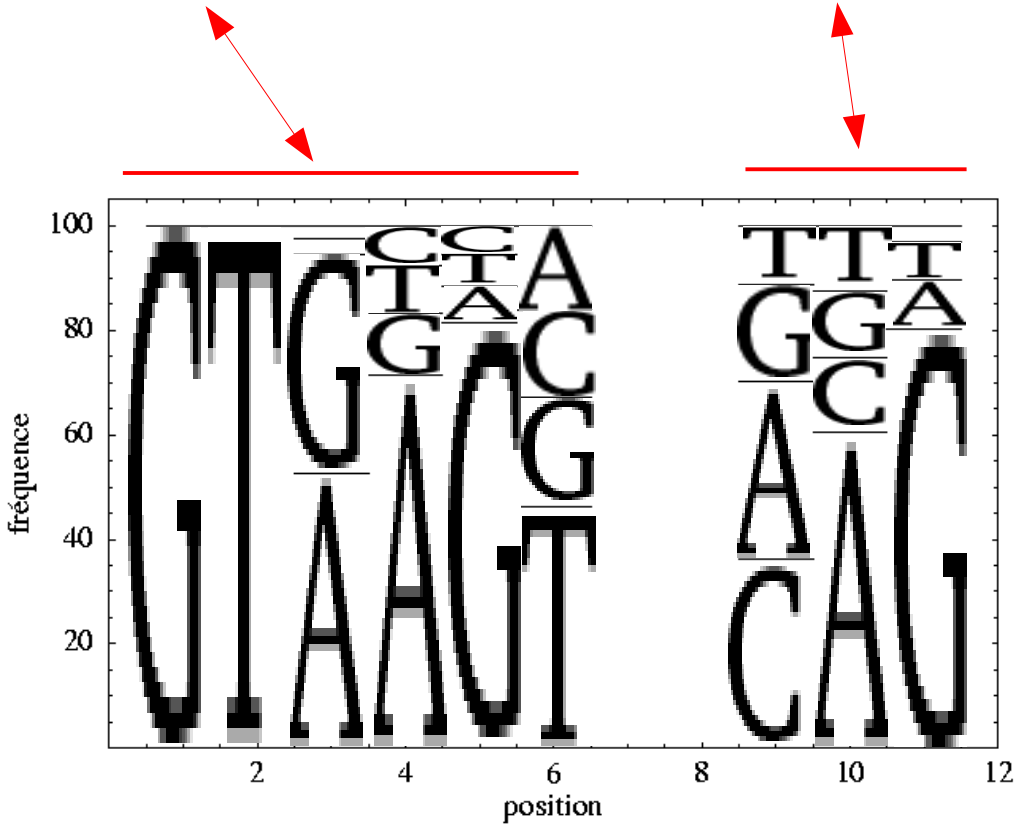
"GT-AG" splicing junction

Splicing sites

intron

...GAGGAGGG|**gtgagt**gtg.....cctctccc**cag**|CTGCTC...

"GT-AG" type consensus sequence

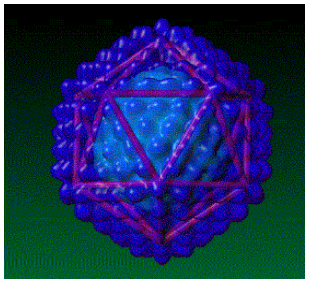


98.7% of known splicing sites

The discovery of splicing

Richard Roberts
New England Biolabs

1977
Adenovirus



Chicken
albumin
1977

Pierre Chambon
Institut de Génétique
et Biologie Moléculaire
et Cellulaire, Strasbourg

Philip Sharp, MIT

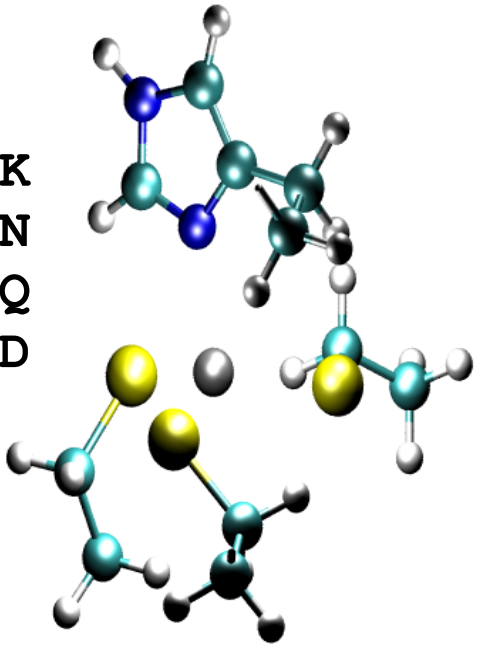
Ovalbumin gene is split in chicken DNA; BREATHNACH, R; MANDEL, JL; CHAMBON, P (1977) NATURE 270:314-319

Prosite database of functional patterns in proteins:
http://www.expasy.org/prosite

Zinc coordination sites in nuclear hormone receptors

```
                  *  *                    *  *
androgen          CLICGDEASGAHYGALTCGSCKVFFKRAAEGK
Rev Erb           CKVCGDVASGFHYGVLACEGCKGFFRRSIQQN
glucocorticoid    CLVCSDEASGCHYGVLTCEGCKAFFKRAVEGQ
retinoic acid     CAICGDRSSGKHYGVYSCEGCKGFFKRTVRKD

consensus         CxxCxDxxxxxHFxxxxCxxCxxFFxR
"PROSITE"               E     NY
pattern
```

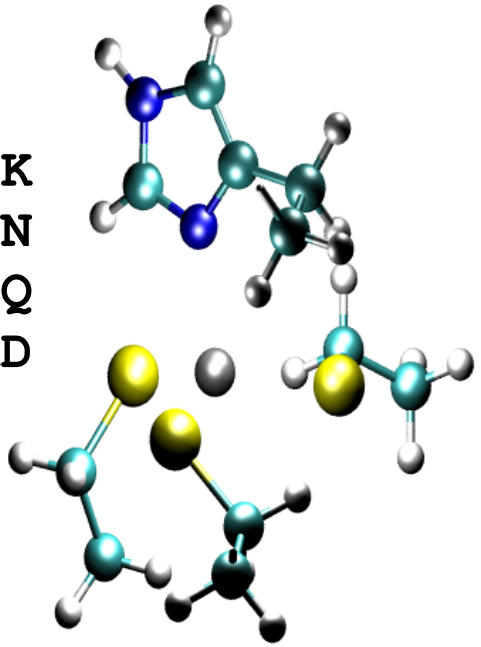


Prosite database of functional patterns in proteins:
http://www.expasy.org/prosite

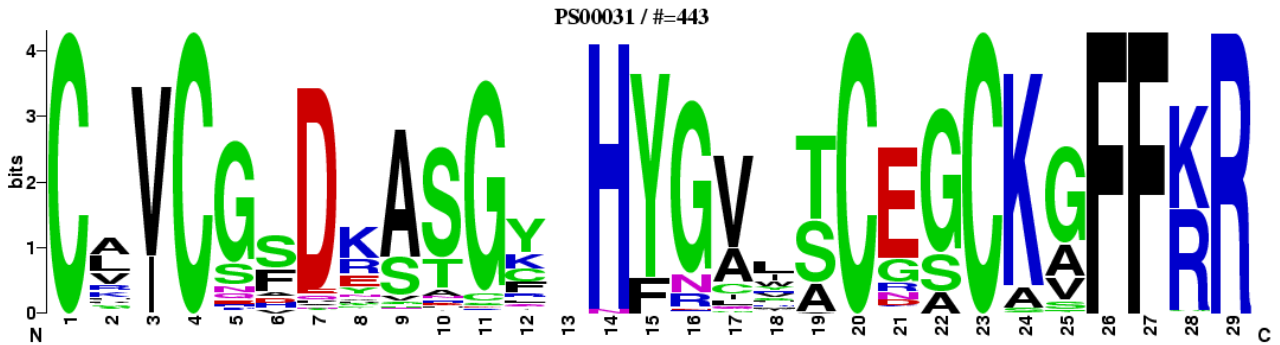
Zinc coordination sites in nuclear hormone receptors

```
                  *  *                 *  *
androgen          CLICGDEASGAHYGALTCGSCKVFFKRAAEGK
Rev Erb           CKVCGDVASGFHYGVLACEGCKGFFRRSIQQN
glucocorticoid    CLVCSDEASGCHYGVLTCEGCKAFFKRAVEGQ
retinoic acid     CAICGDRSSGKHYGVYSCEGCKGFFKRTVRKD

consensus         CxxCxDxxxxxHFxxxxCxxCxxFFxR
"PROSITE"                 E     NY
pattern
```



Sequence profile



Representation by a “regular expression” or search pattern

consensus pattern	**C**xx**C**x**D**xxxxx**HF**xxxx**C**xx**C**xx**FF**x**R** **E** **NY**

C-x(2)-C-x-[DE]-x(5)-[HN]-[FY]-x(4)-C-x(2)-C-x(2)-F-F-x-R

grep format: C.{2}C.[DE].{5}[HN][FY].{4}C.{2}FF.R

perl format: C\w{2}C\w[DE]\w{5}[HN][FY]\w{4}C\w{2}FF\wR

http://www.expasy.org/prosite

Pattern matching: a naïve algorithm

To recognize the polyadenylation motif **AAUAAA** in the sequence below:

AAAUAUUAAAUAUUAGACGAAUAAAAGUAUAUUU

we have the naïve procedure:

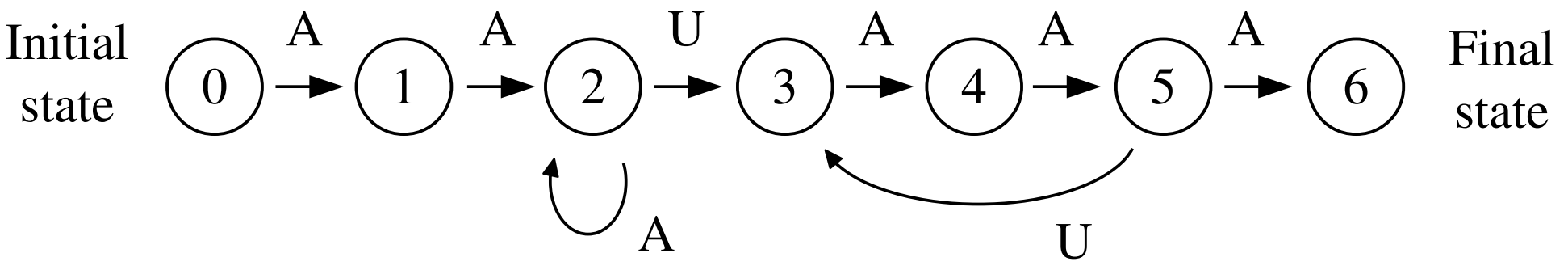
- Find an A; assume it is the beginning of the pattern
- Test the following nucleotides one by one
- If the match fails, start over with the next A

The method is far from optimal (O(nm))

Genome Res. 2000 10:1001

Pattern matching with automata

The following automaton recognizes the polyadenylation motif **AAUAAA**



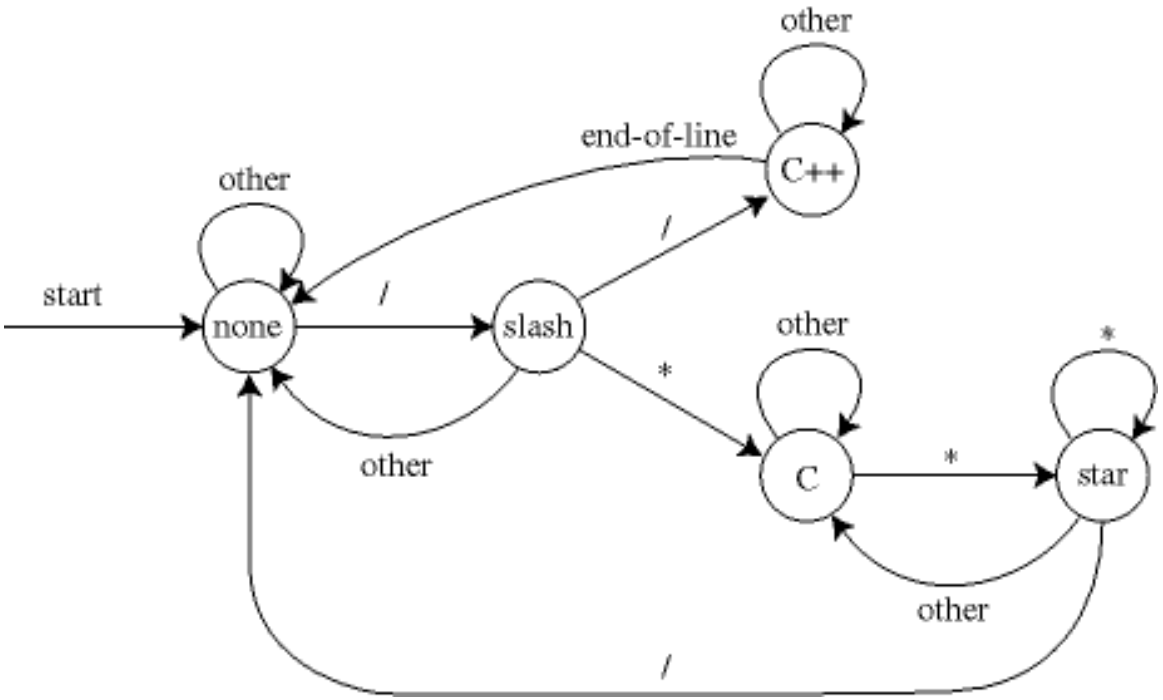
All other transitions go back to 0 (not shown).

Example:

target: AAAUAUUAAAUAUUAGACGAAUAAAAGUAUAUUU

state: 122340012234001010012345**6**

Automata: a vast area in computer science...



An automaton to remove comments from Java programs

Biological Sequence Analysis. Durbin et al, 1998, Cambridge U Press, chapter 9
Algorithms on strings, trees and sequences. D Gusfeld, 1997, Cambridge U Press

Multiple alignments: a 3 stage progressive method

Applications: functional residues, structure prediction, phylogeny

Short patterns and signals: other techniques are needed