Před analýzou

```
>P12345 Yeast chromosome1
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
```

Po částečné analýze

```
         >P12345 Gene_1 – gen kodujici
         protein alkoholdehydrogenazy ...

TATA     TATAAA
         CGATTGACGATGACGAT
start    ATG
exon1    TACAGATTACAGATTACAGATTACAGATGT
intron1  CAGATTACAGATTACAGATTACAGATTACAGATTCA
exon2    AGATTACAGATTACAGATTACAGA
stop     TAA

         >P12346 Protein_1
         MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN
```

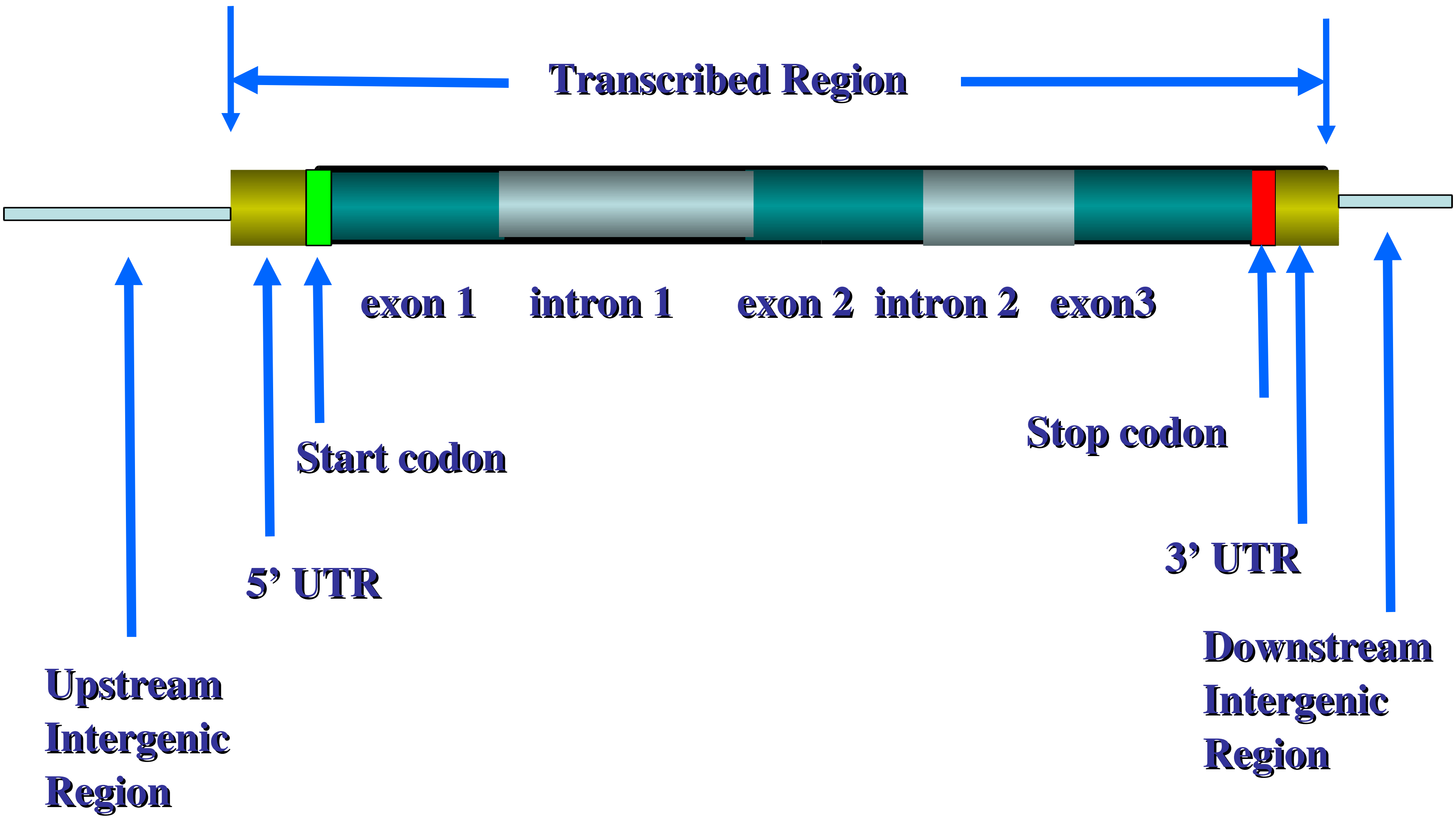
Analýza DNA sekvence

- =~ anotace genomu (sekvence)
  - identifikace signálů a genů
  - anotace genů (jejich kódujících sekvencí)

Anotace genů =~ anotace proteinů

- Identifikace a popis fyzikálně-chemických, funkčních a strukturních vlastností daného genu/proteinu
  - sekvence DNA, AA, pozice v genomu, délka, složení
  - běžné názvy, odkazy na literaturu
  - příslušnost do rodiny, evoluce
  - partneři pro interakci, aktivita, regulační mechanismy
  - struktura, aktivní místa, role v metabolismu buňky

Eukaryotic Gene Structure



Transcribed Region

exon 1

intron 1

exon 2

intron 2

exon3

Start codon

Stop codon

5’ UTR

3’ UTR

Upstream
Intergenic
Region

Downstream
Intergenic
Region

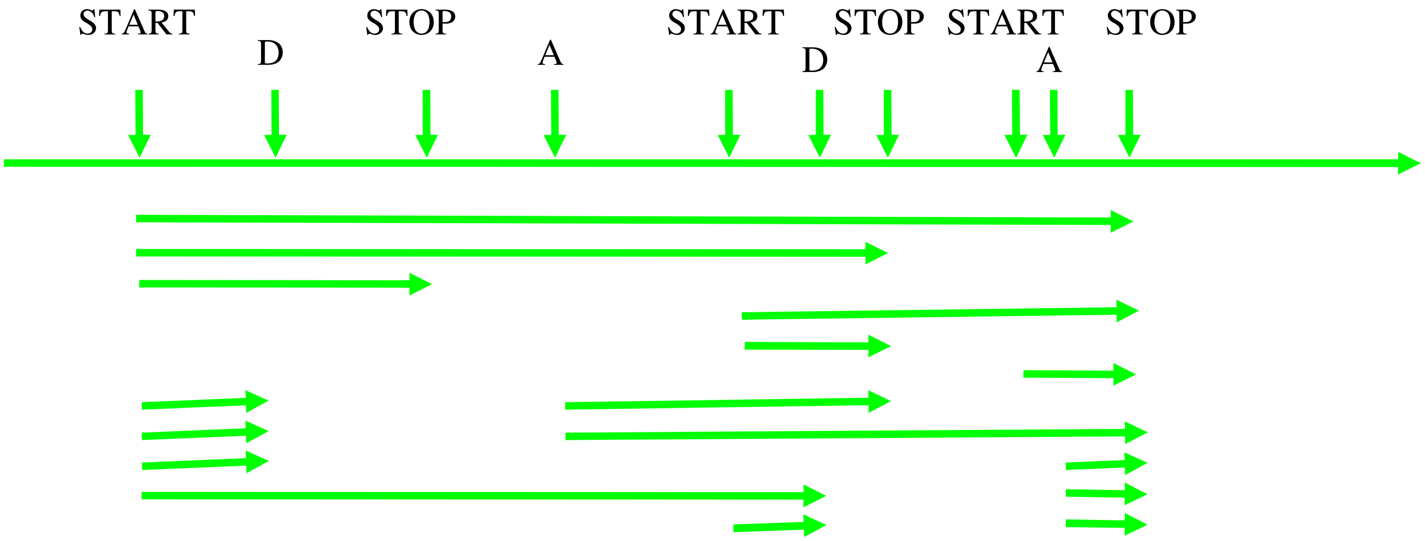
Analýza DNA sekvence

- Statistika
  - frekvence n-gramů a jiných prvků, repetice, kodony
- Signální prvky
  - TATA (promotor), ATG (start), STOP, GT (donor), AG (akceptor) a pod
- Kódující část
  - podobnost kódované sekvence s jinými proteiny
- Kombinované přístupy

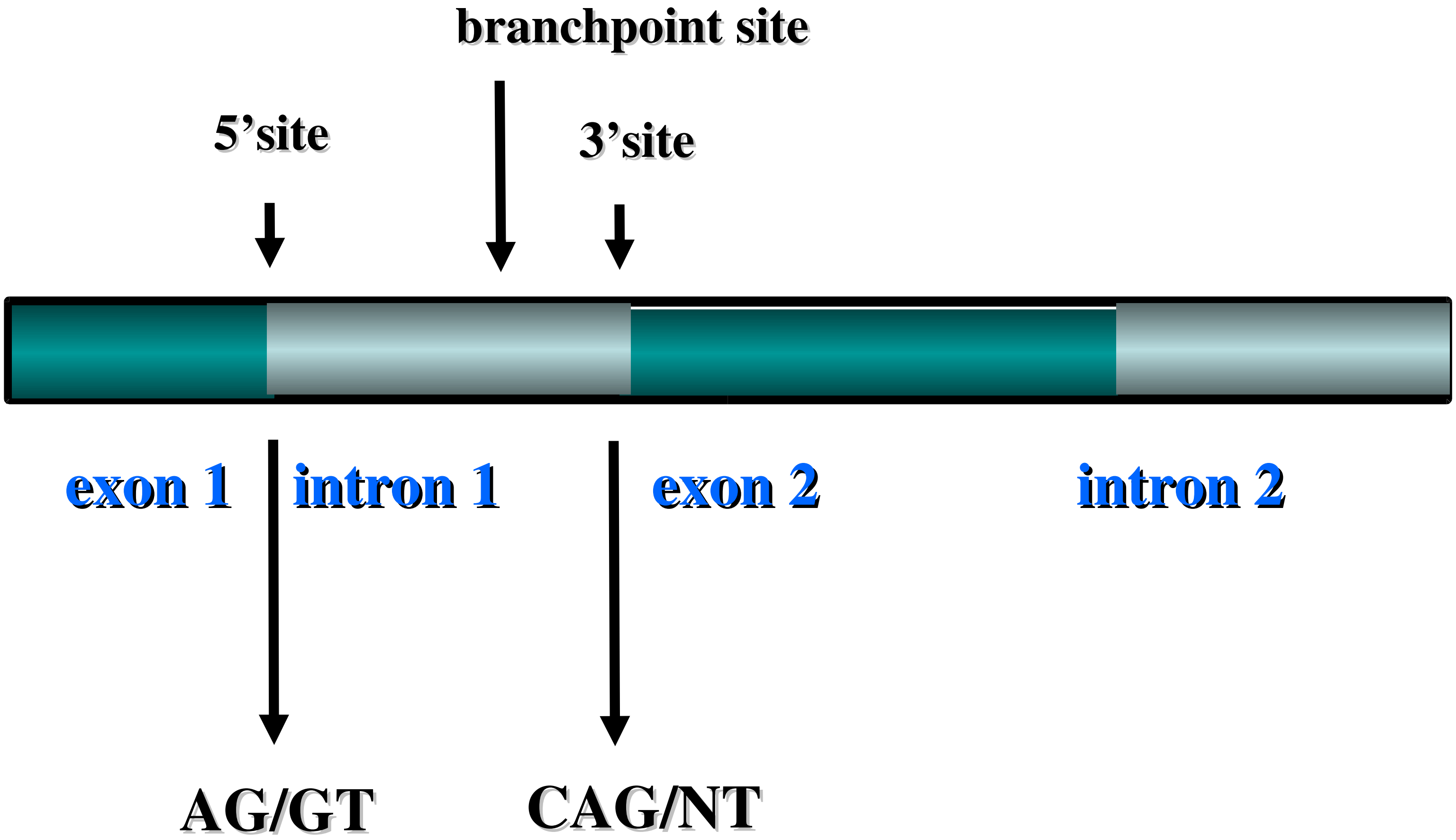
Identifikace genů

- U prokaryotů 95-100% spolehlivost, u složitějších eukaryotů 90% na úrovni bazí, 70% na úrovni exonů/intronů
  - existence intronů
  - větší genomy
  - nízká hustota genů (<30%; 3% u Homo sapiens)
  - alternativní splicing (zhruba u poloviny genů)
  - velké množství repetitivních sekvencí
  - občasný překryv genů

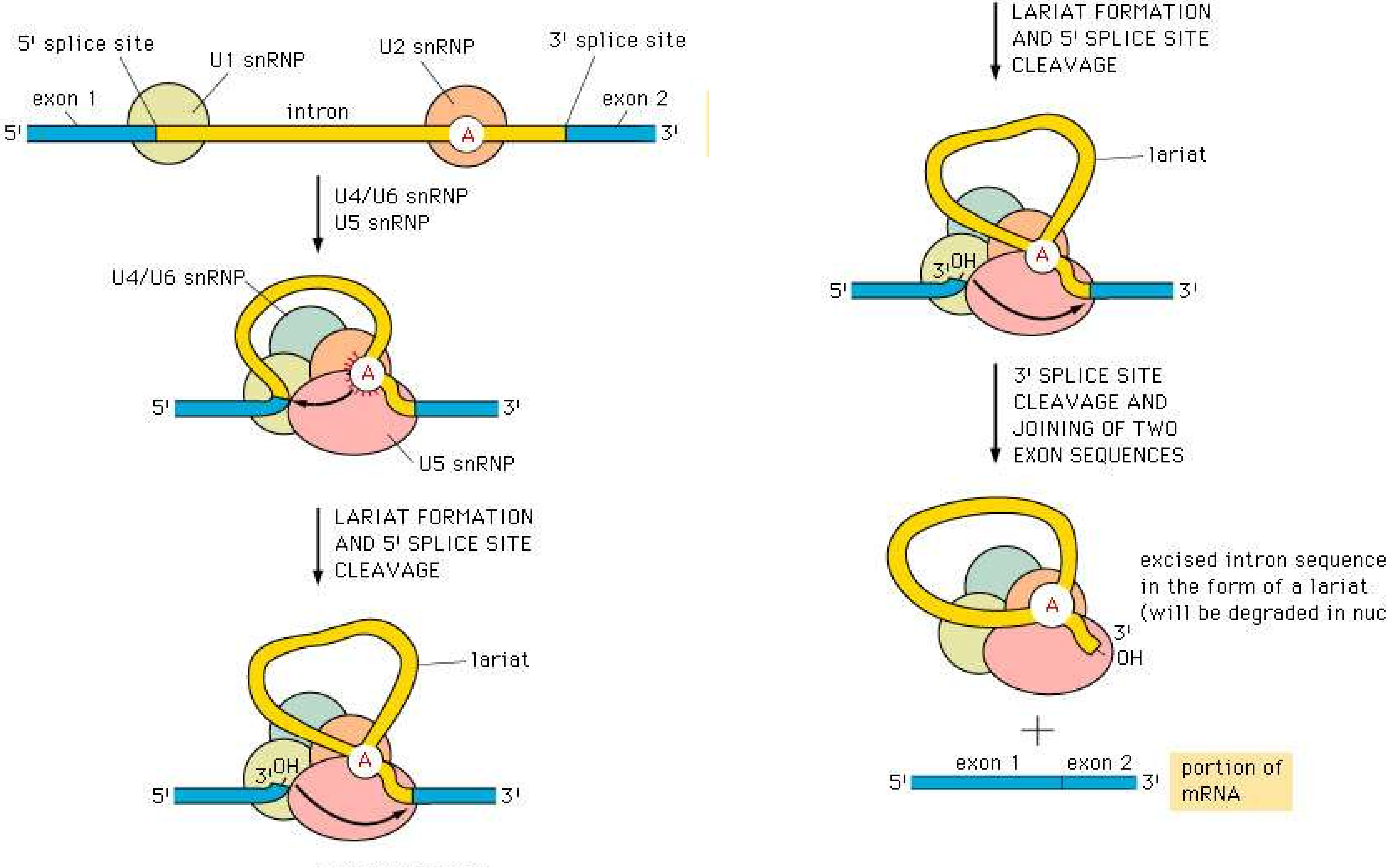
Identifikace genů



Eukaryotic Gene Structure



RNA Splicing



U4/U6 snRNP
U5 snRNP

LARIAT FORMATION
AND 5' SPLICE SITE
CLEAVAGE

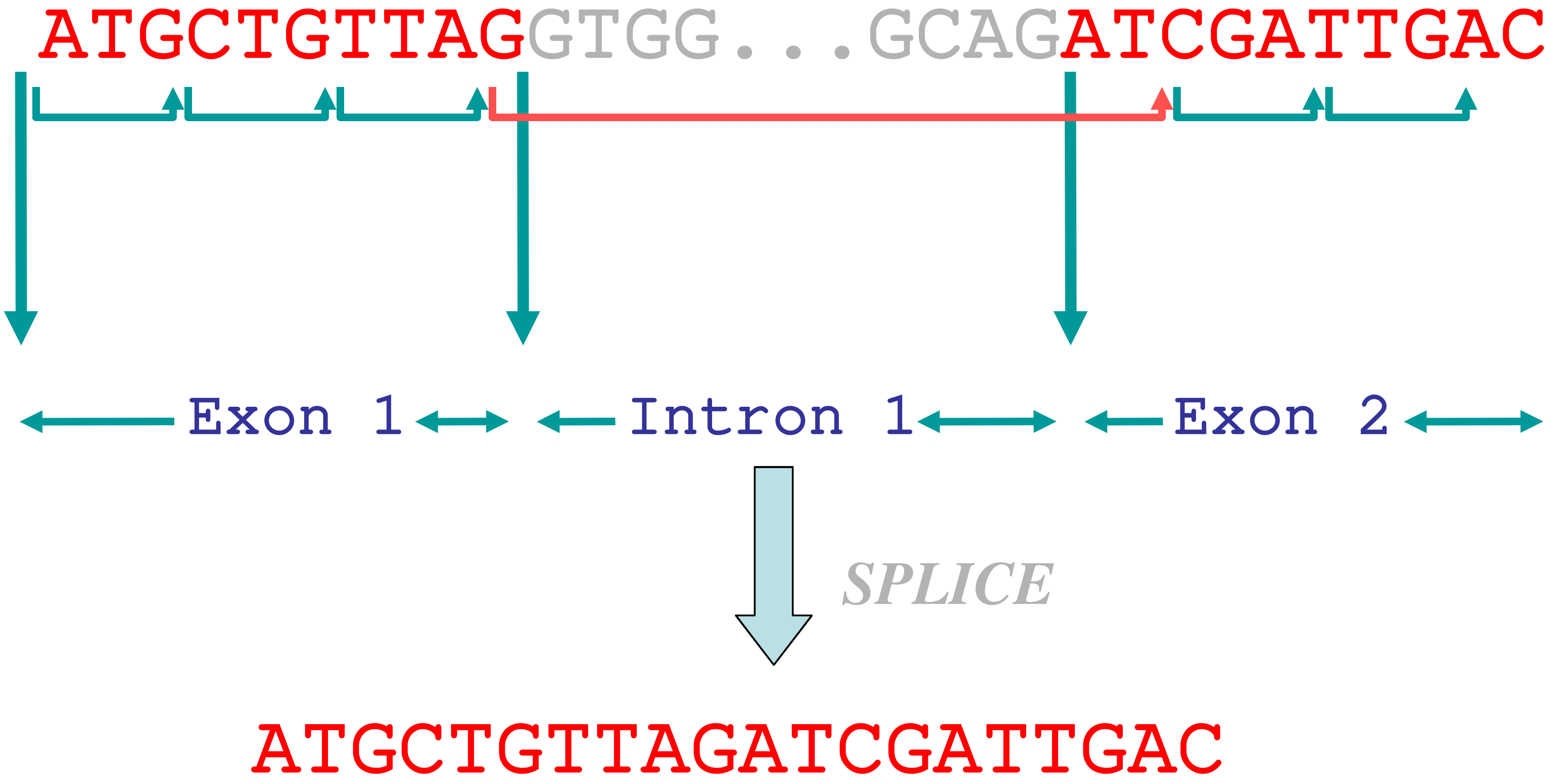
LARIAT FORMATION
AND 5' SPLICE SITE
CLEAVAGE

3' SPLICE SITE
CLEAVAGE AND
JOINING OF TWO
EXON SEQUENCES

excised intron sequence
in the form of a lariat
(will be degraded in nuc

portion of
mRNA

Exon/Intron Structure (Detail)



ATGCTGTTAGGTGG...GCAGATCGATTGAC

Exon 1

Intron 1

Exon 2

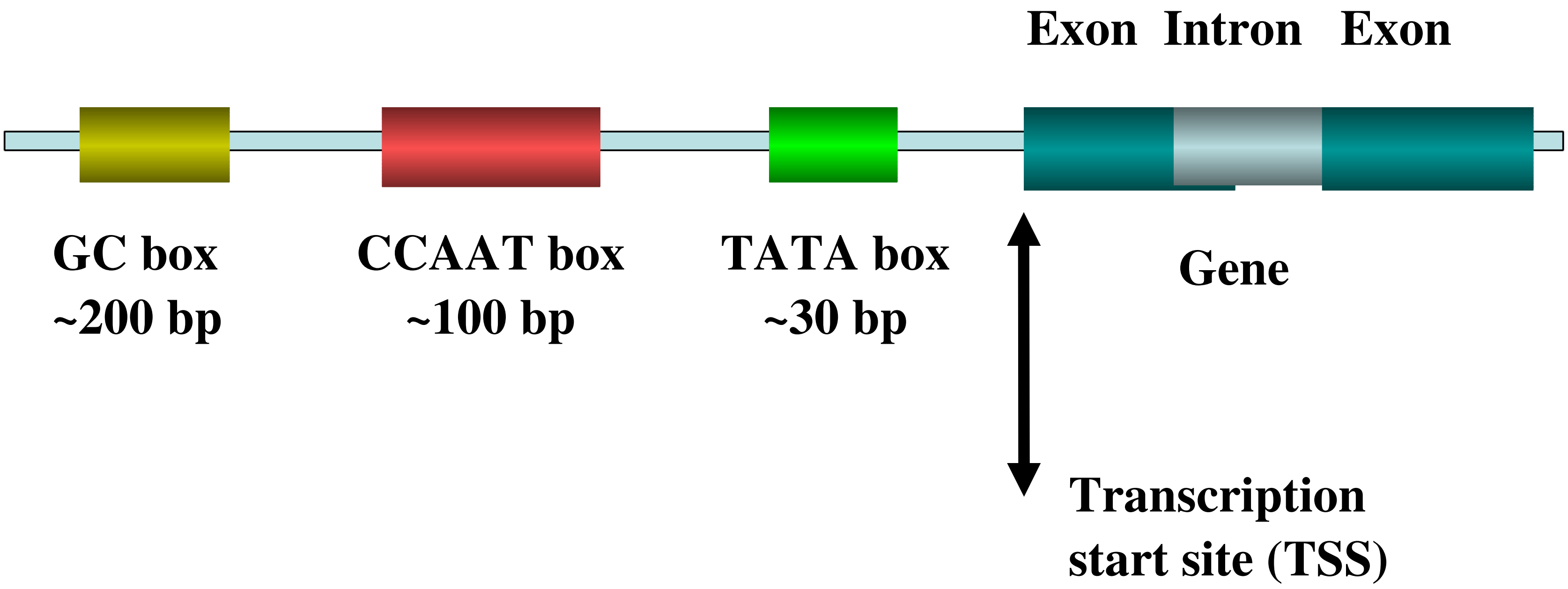
SPLICE

ATGCTGTTAGATCGATTGAC

Typické signály v eukaryotických sekvencích

- Promotorové elementy
  - CAP, CCAAT, GC a TATA
- Kozakova sekvence (rozpoznávána ribozomem = RBS)
- Splicing (donor, acceptor a lariat)
- Terminační signál
- Polyadenylační signál

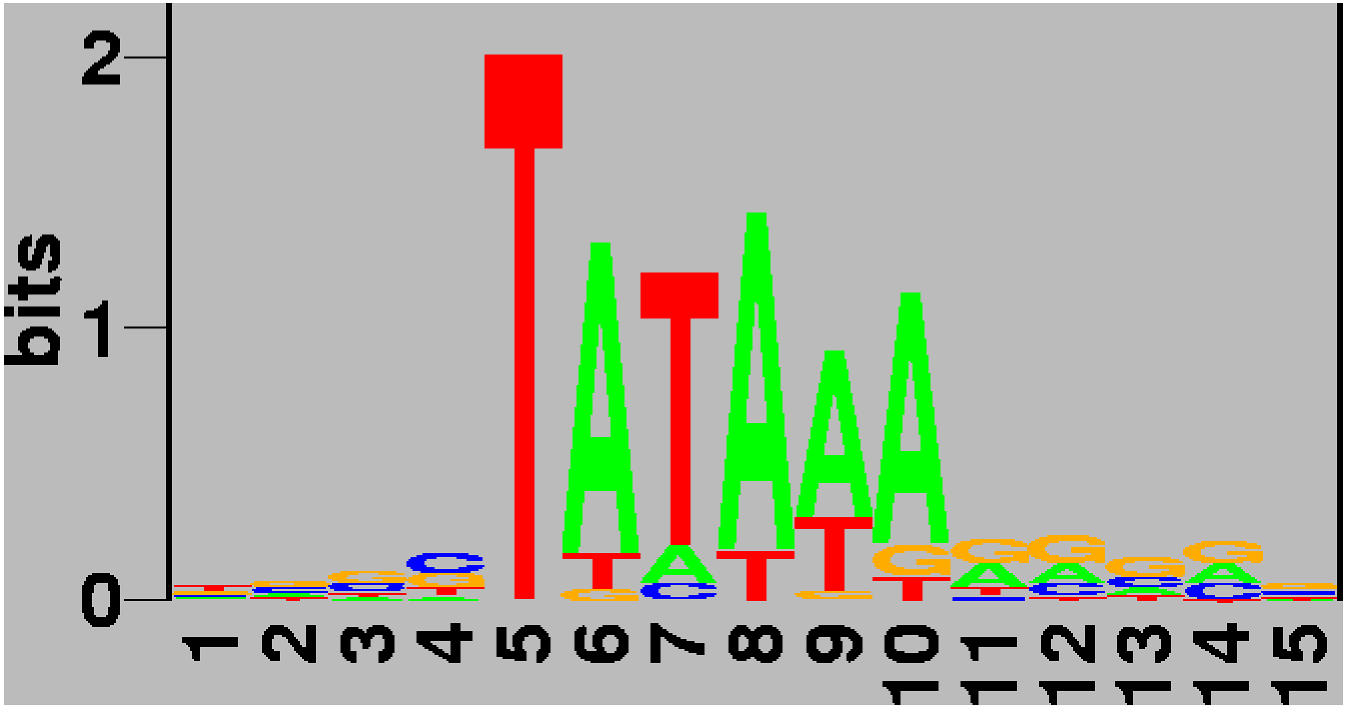
Pol II Promoter Elements



Pol II Promoter Elements

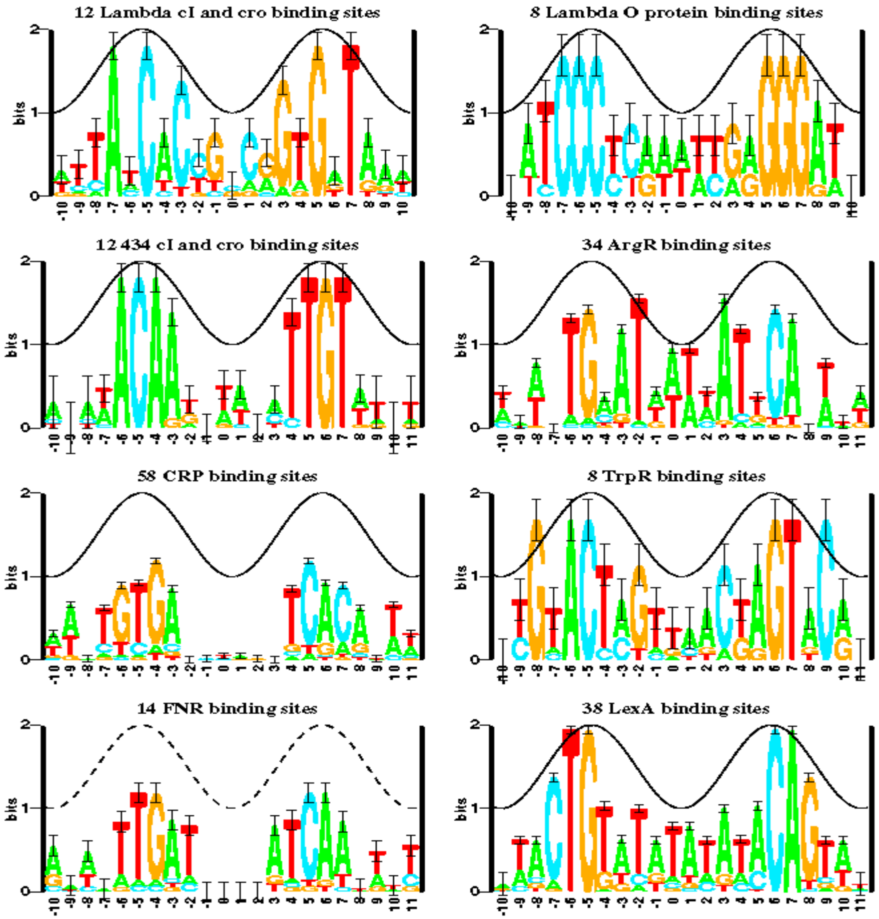
- Cap Region/Signal
  - n C A G T n G
- TATA box (~ 25 bp upstream)
  - T A T A A A n G C C C
- CCAAT box (~100 bp upstream)
  - T A G C C A A T G
- GC box (~200 bp upstream)
  - A T A G G C G nGA

Pol II Promoter Elements



TATA box is found in ~70% of promoters

WebLogos



12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

8 TrpR binding sites

14 FNR binding sites

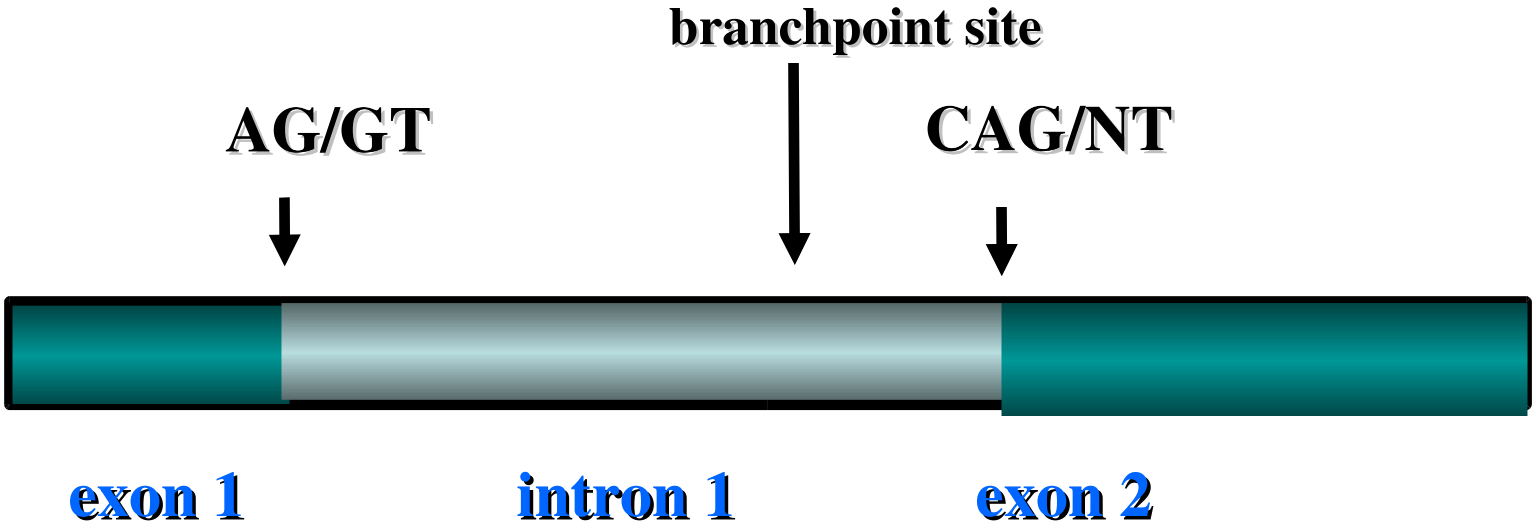
38 LexA binding sites

http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi

Kozak (RBS) Sequence

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

Splice Signals



branchpoint site

AG/GT

CAG/NT

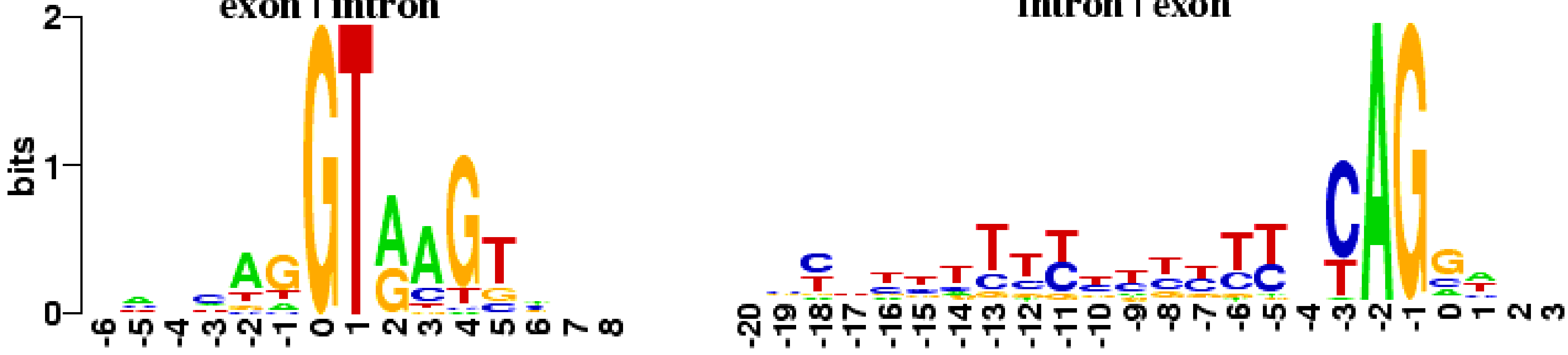
exon 1

intron 1

exon 2

exon | intron

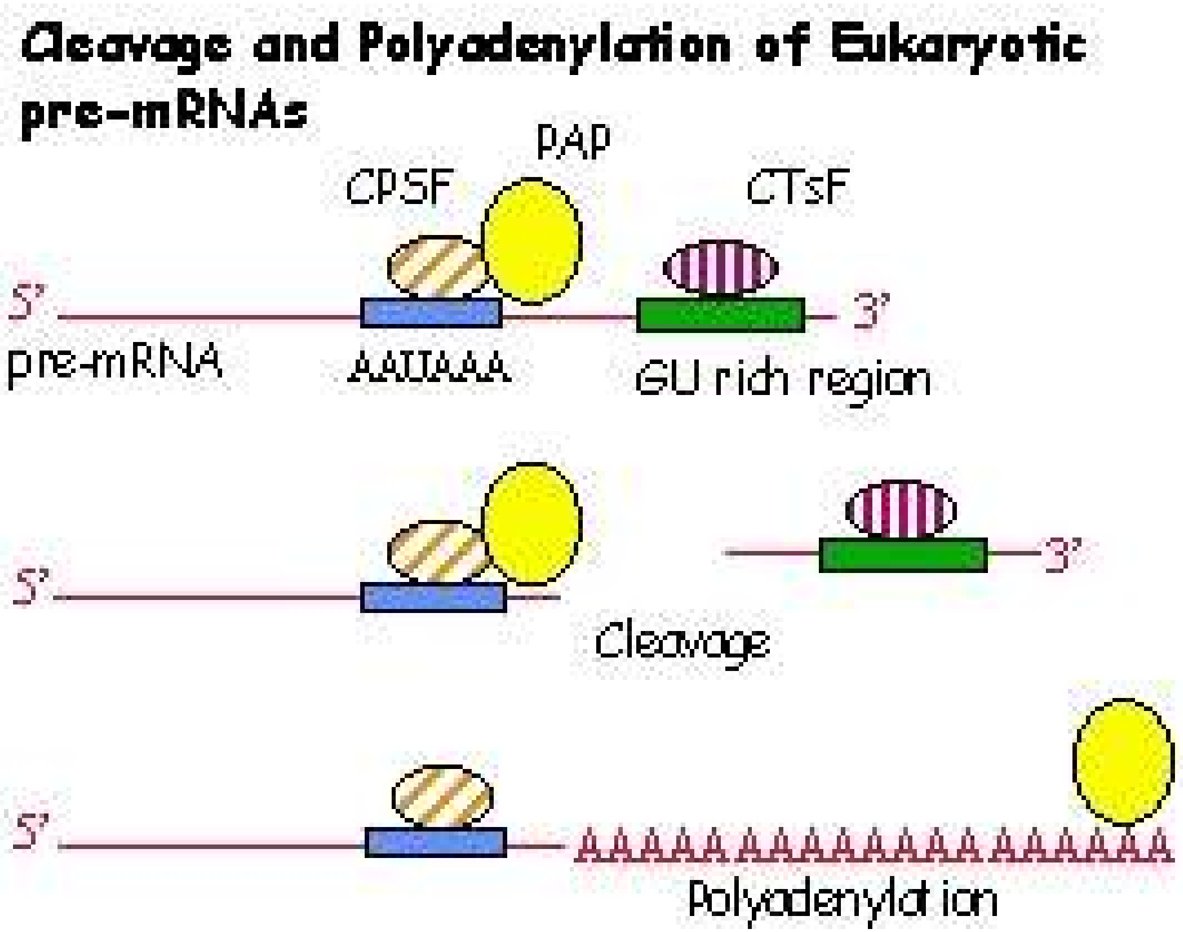
intron | exon



Miscellaneous Signals

- Polyadenylation signal
  - A A T A A A or A T T A A A
  - Located 20 bp upstream of poly-A cleavage site
- Termination Signal
  - A G T G T T C A
  - Located ~30 bp downstream of poly-A cleavage site

Polyadenylation



CPSF – Cleavage & Polyadenylation Specificity Factor

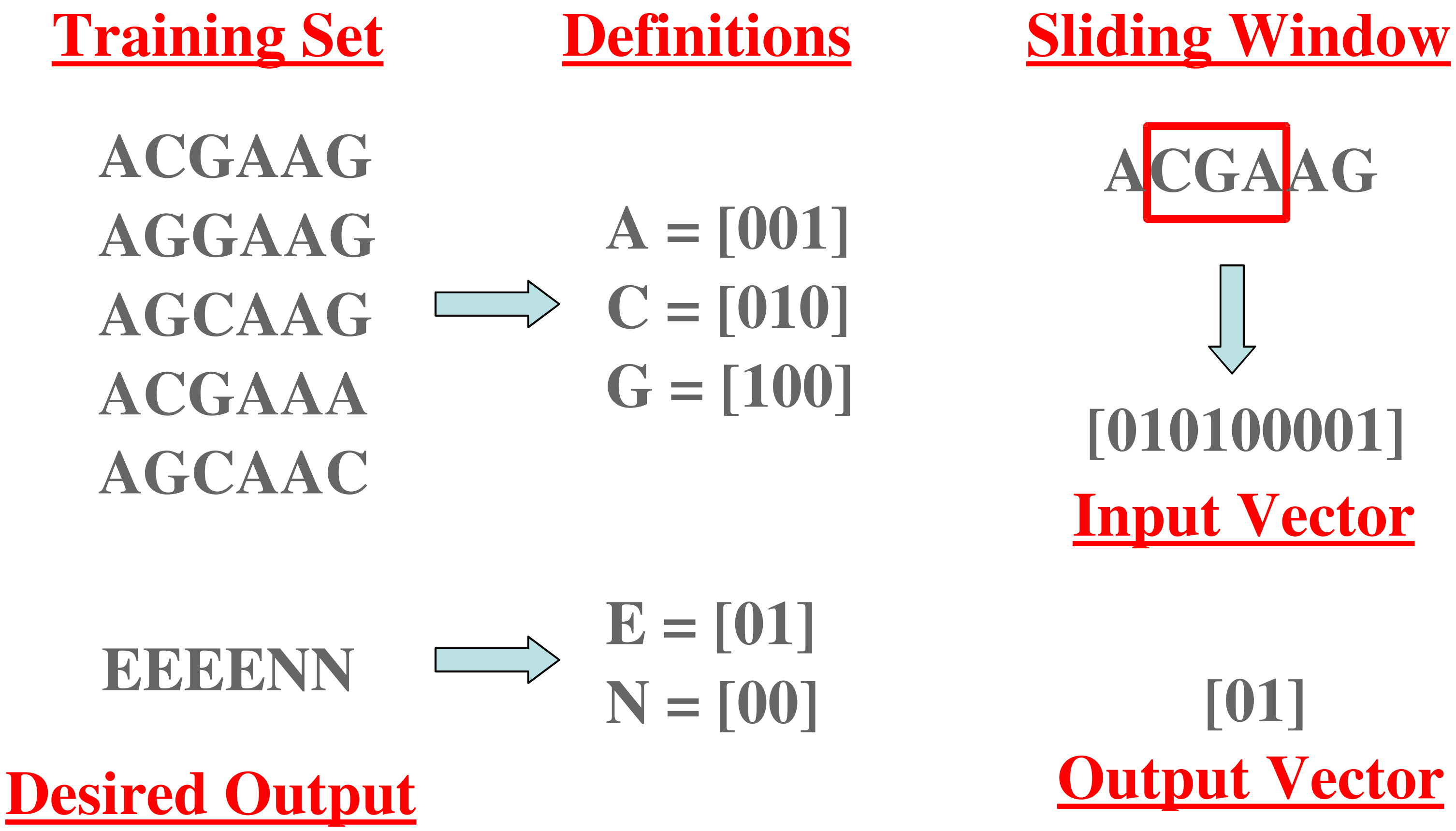
PAP – Poly-A Polymerase

CTsF – Cleavage Stimulation Factor

Analýza genomu – kombinované metody

- Neurónové sítě
  - Grail, GeneParser
- Lineární diskriminační analýza
  - GeneFinder, GeneID, MZEF
- Lingvistická
  - GeneLang
- Markovovy řetězce
  - Genie, GeneMark, GenScan, VEIL
- Podobnosti
  - Procrustes, AAT
- Rozhodovací stromy

Neural Network



Training Set

ACGAAG
AGGAAG
AGCAAG
ACGAAA
AGCAAC

Definitions

A = [001]
C = [010]
G = [100]

Sliding Window

ACGAAG

[010100001]

Input Vector

EEEENN

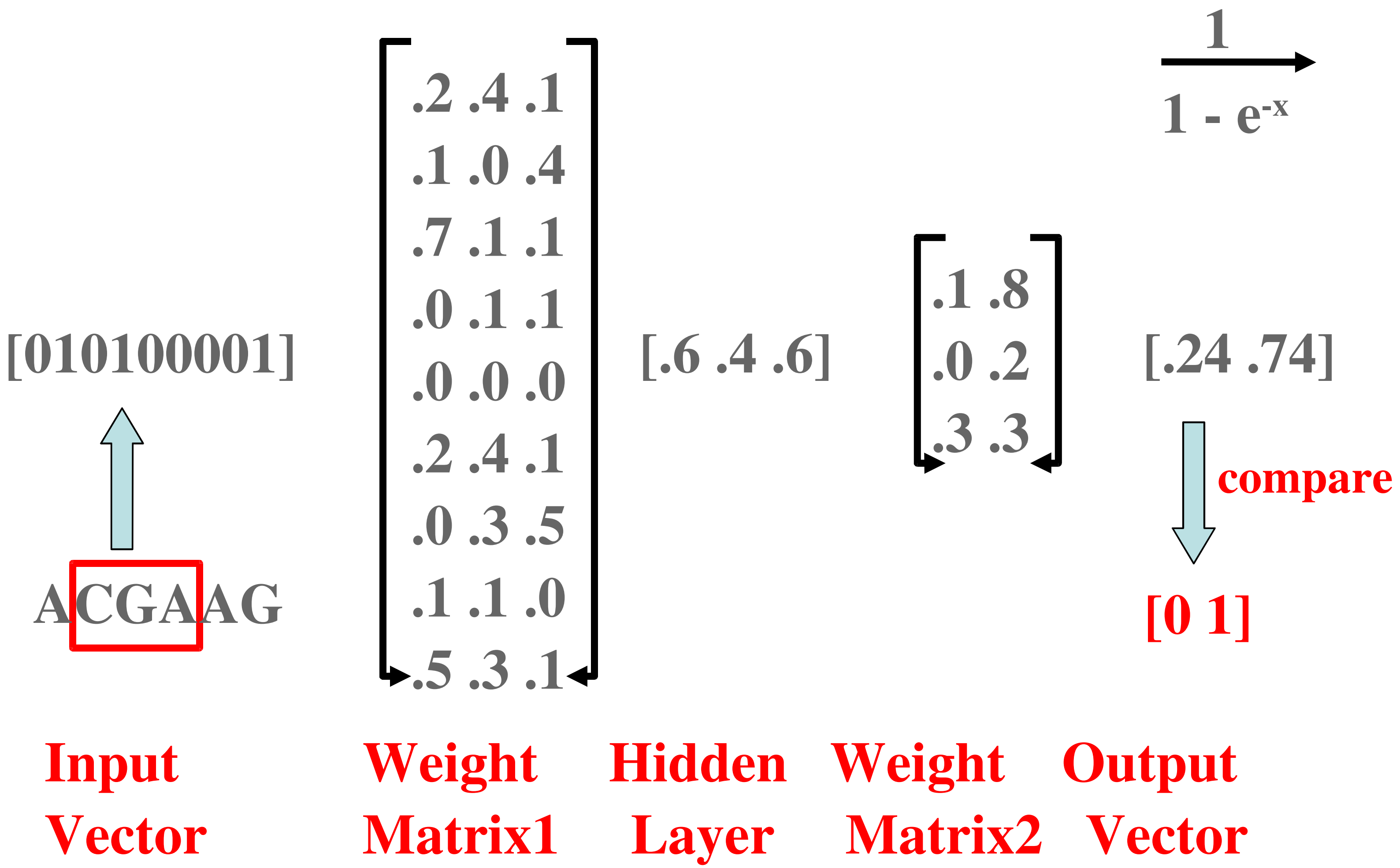
E = [01]
N = [00]

Desired Output

[01]

Output Vector

Neural Network Training



$$\frac{1}{1 - e^{-x}}$$

ACGAAG

[010100001]

$$\begin{bmatrix} .2 & .4 & .1 \\ .1 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .4 & .1 \\ .0 & .3 & .5 \\ .1 & .1 & .0 \\ .5 & .3 & .1 \end{bmatrix}$$

[.6 .4 .6]

$$\begin{bmatrix} .1 & .8 \\ .0 & .2 \\ .3 & .3 \end{bmatrix}$$

[.24 .74]

compare

[0 1]

Input Vector

Weight Matrix1

Hidden Layer

Weight Matrix2

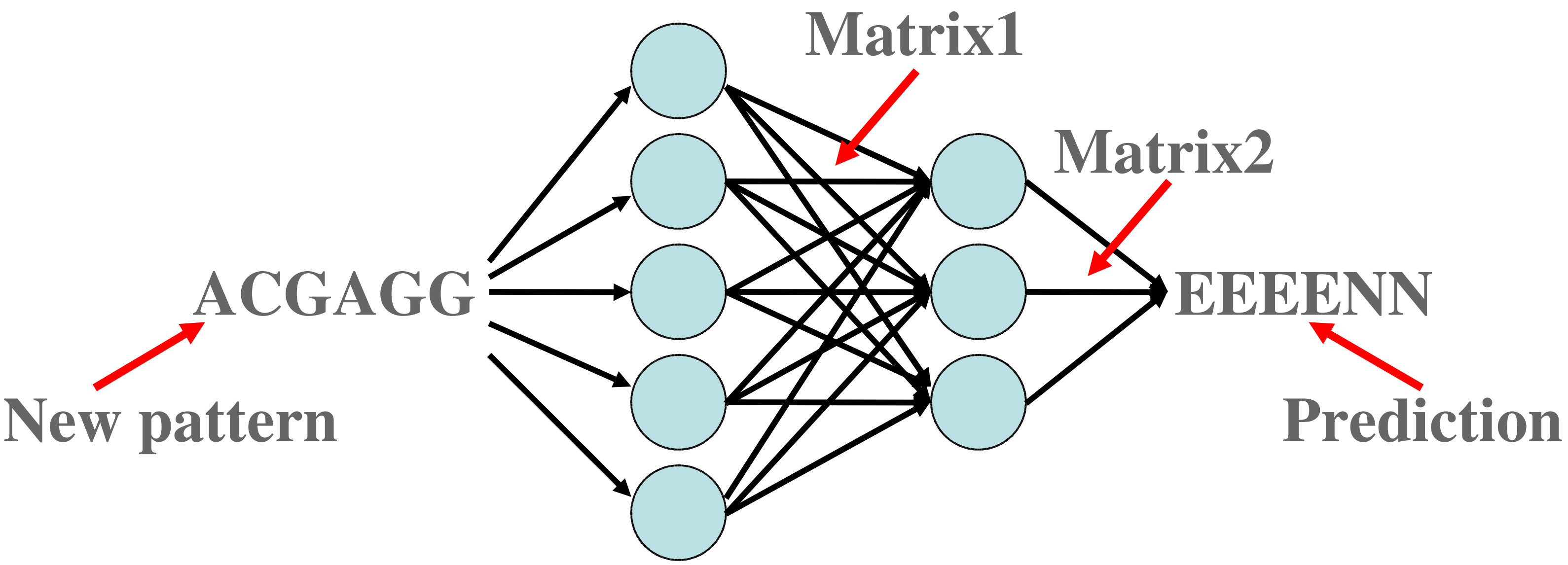
Output Vector

After Many Iterations….

$$\begin{bmatrix} .13 & .08 & .12 \\ .24 & .01 & .45 \\ .76 & .01 & .31 \\ .06 & .32 & .14 \\ .03 & .11 & .23 \\ .21 & .21 & .51 \\ .10 & .33 & .85 \\ .12 & .34 & .09 \\ .51 & .31 & .33 \end{bmatrix} \qquad \begin{bmatrix} .03 & .93 \\ .01 & .24 \\ .12 & .23 \end{bmatrix}$$

Two “Generalized” Weight Matrices

Neural Networks



Matrix1

Matrix2

ACGAGG

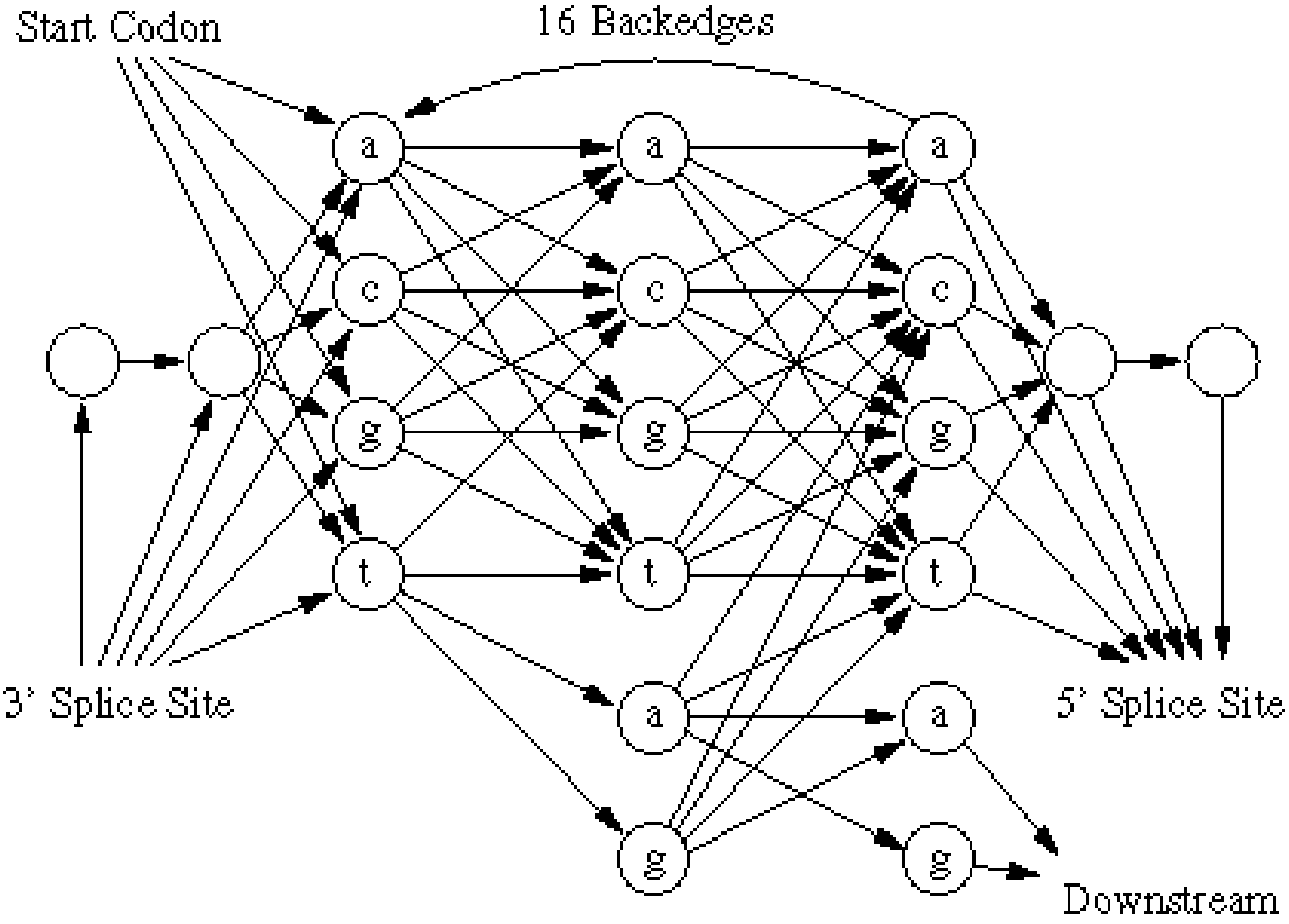
EEEENN

New pattern

Prediction

Input Layer 1 Hidden Layer Output

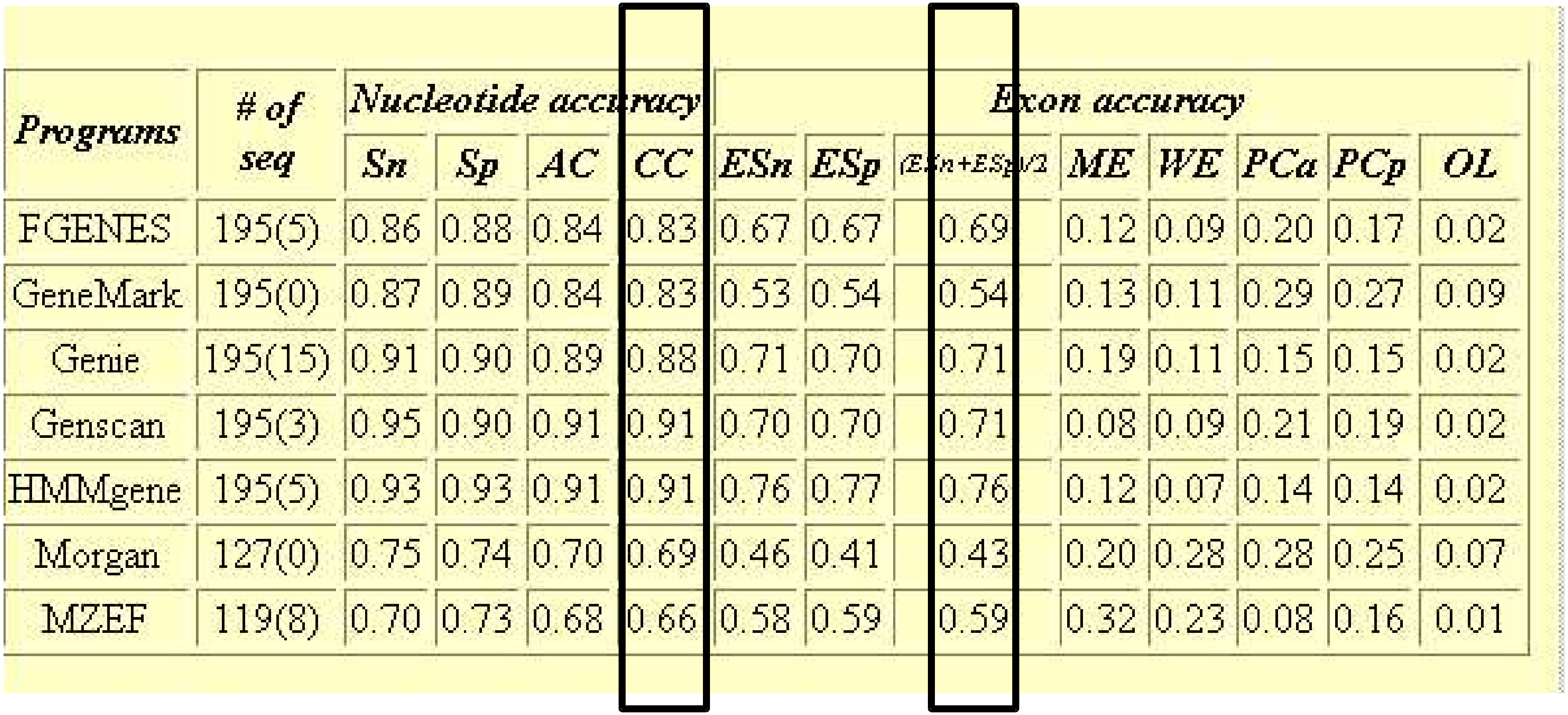
HMM for Gene Finding



Combined Methods

- Bring 2 or more methods together (usually site detection + composition)
- GRAIL (http://compbio.ornl.gov/Grail-1.3/)
- FGENEH (http://genomic.sanger.ac.uk/gf/gf.shtml)
- HMMgene (http://www.cbs.dtu.dk/services/HMMgene/)
- GENSCAN(http://genes.mit.edu/GENSCAN.html)
- Gene Parser (http://beagle.colorado.edu/~eesnyder/GeneParser.html)
- GRPL (GeneTool/BioTools)

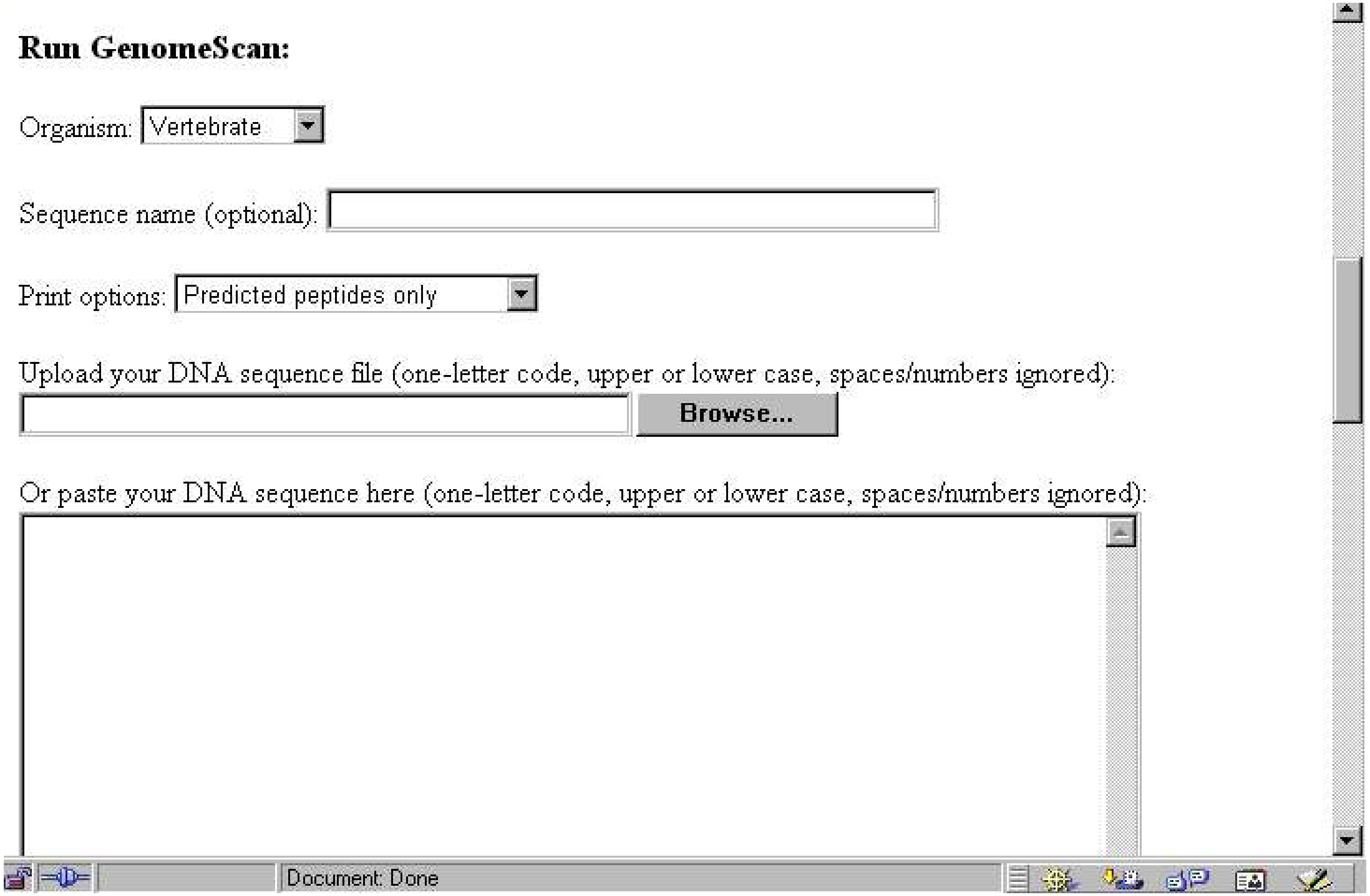
How Well Do They Do?



Programs	# of seq	Nucleotide accuracy				Exon accuracy								
		Sn	Sp	AC	CC	ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195(5)	0.86	0.88	0.84	0.83	0.67	0.67	0.69	0.12	0.09	0.20	0.17	0.02
GeneMark	195(0)	0.87	0.89	0.84	0.83	0.53	0.54	0.54	0.13	0.11	0.29	0.27	0.09
Genie	195(15)	0.91	0.90	0.89	0.88	0.71	0.70	0.71	0.19	0.11	0.15	0.15	0.02
Genscan	195(3)	0.95	0.90	0.91	0.91	0.70	0.70	0.71	0.08	0.09	0.21	0.19	0.02
HMMgene	195(5)	0.93	0.93	0.91	0.91	0.76	0.77	0.76	0.12	0.07	0.14	0.14	0.02
Morgan	127(0)	0.75	0.74	0.70	0.69	0.46	0.41	0.43	0.20	0.28	0.28	0.25	0.07
MZEF	119(8)	0.70	0.73	0.68	0.66	0.58	0.59	0.59	0.32	0.23	0.08	0.16	0.01

"Evaluation of gene finding programs" S. Rogic, A. K. Mackworth and B. F. F. Ouellette. Genome Research, 11: 817-832 (2001).

GenomeScan - http://genes.mit.edu/genomescan.html



Run GenomeScan:

Organism: Vertebrate

Sequence name (optional):

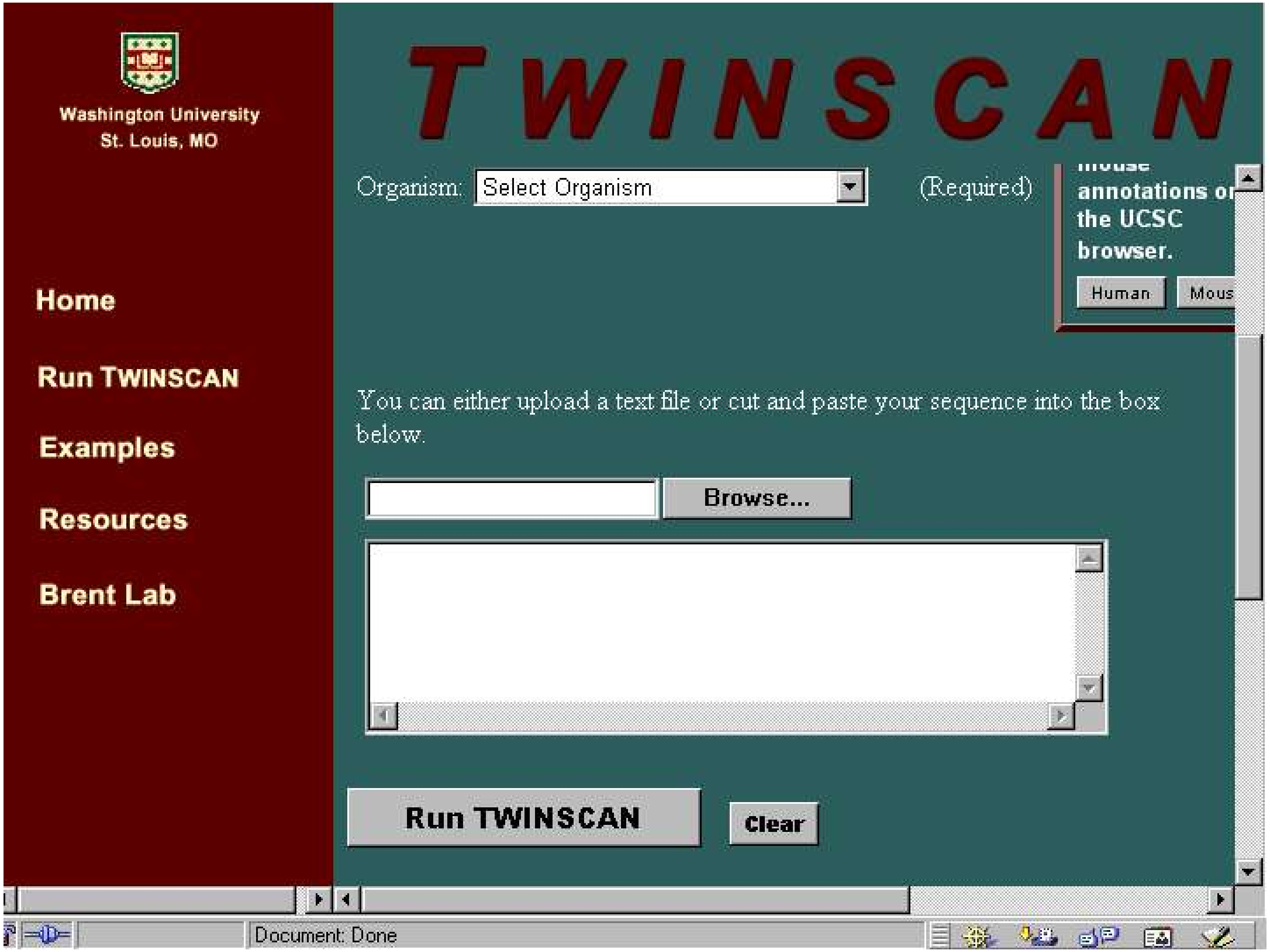
Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

Document: Done

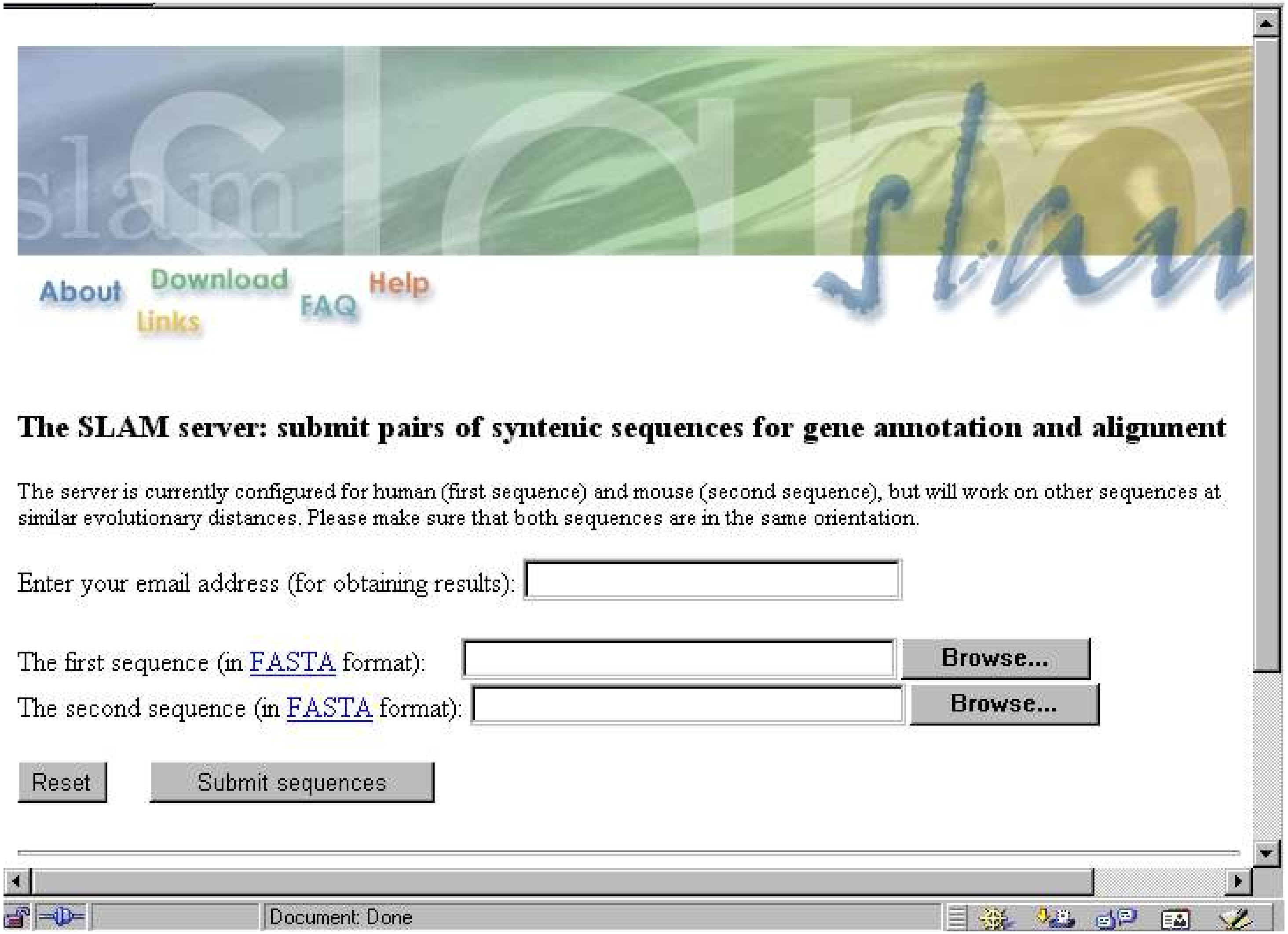
TwinScan - http://genes.cs.wustl.edu/



Organism: Select Organism (Required)

You can either upload a text file or cut and paste your sequence into the box below.

SLAM - http://baboon.math.berkeley.edu/~syntenic/slam.html



The SLAM server: submit pairs of syntenic sequences for gene annotation and alignment

The server is currently configured for human (first sequence) and mouse (second sequence), but will work on other sequences at similar evolutionary distances. Please make sure that both sequences are in the same orientation.

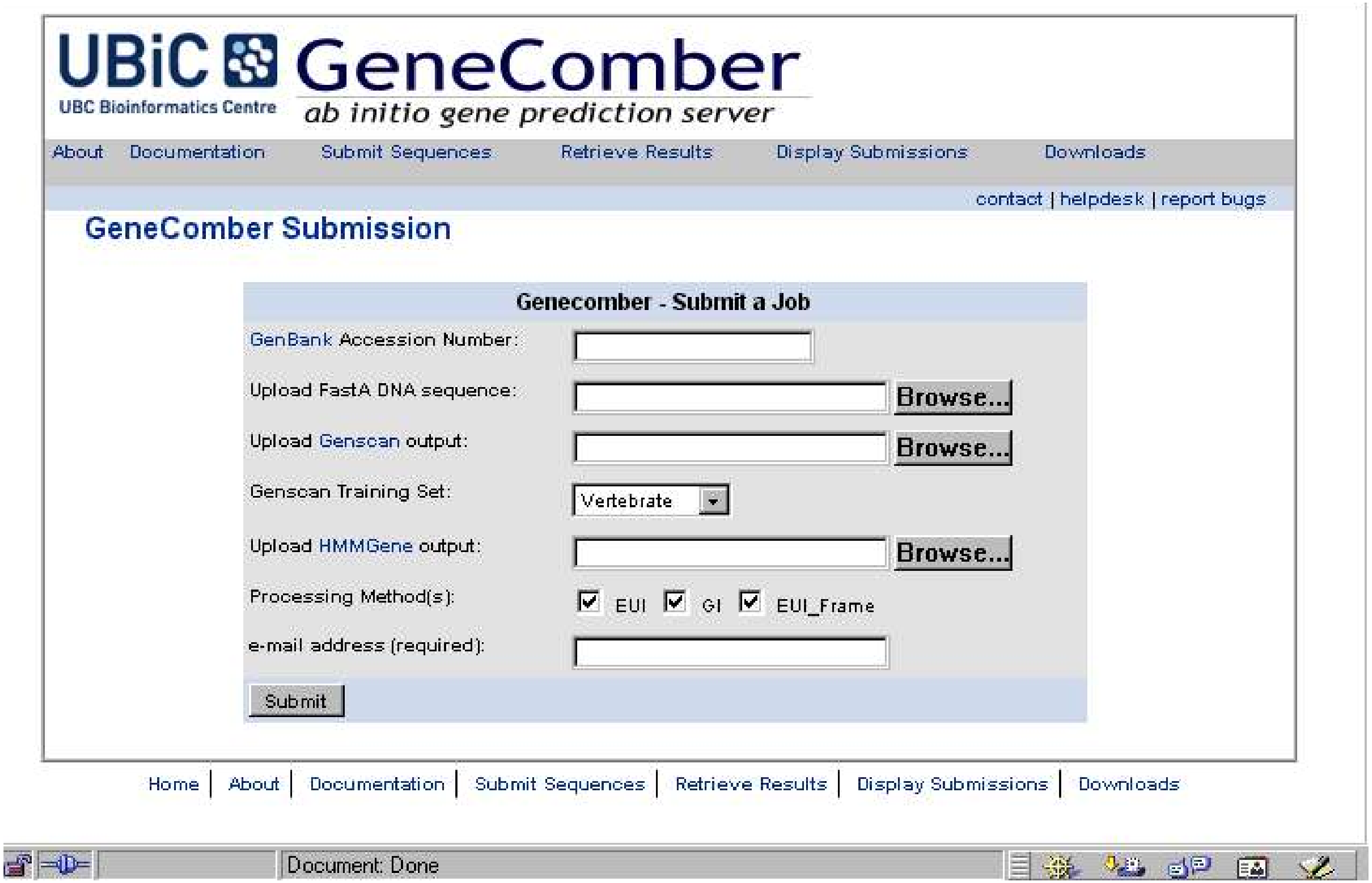
Enter your email address (for obtaining results):

The first sequence (in FASTA format):

The second sequence (in FASTA format):

GeneComber -

http://www.bioinformatics.ubc.ca/genecomber/submit.php



UBiC
UBC Bioinformatics Centre
GeneComber
ab initio gene prediction server
About
Documentation
Submit Sequences
Retrieve Results
Display Submissions
Downloads
contact | helpdesk | report bugs

GeneComber Submission

Genecomber - Submit a Job

GenBank Accession Number:

Upload FastA DNA sequence: Browse...

Upload Genscan output: Browse...

Genscan Training Set: Vertebrate

Upload HMMGene output: Browse...

Processing Method(s): ☑ EUI ☑ GI ☑ EUI_Frame

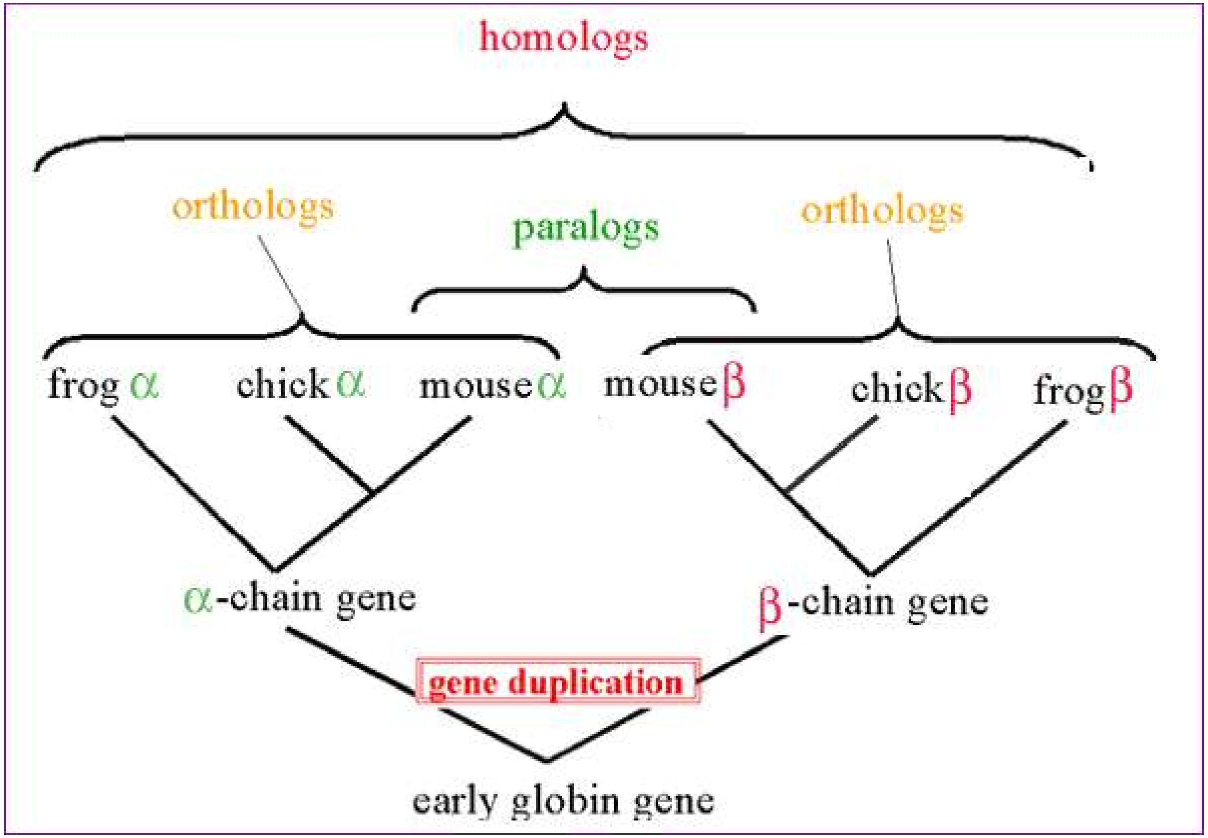
e-mail address (required):

Submit

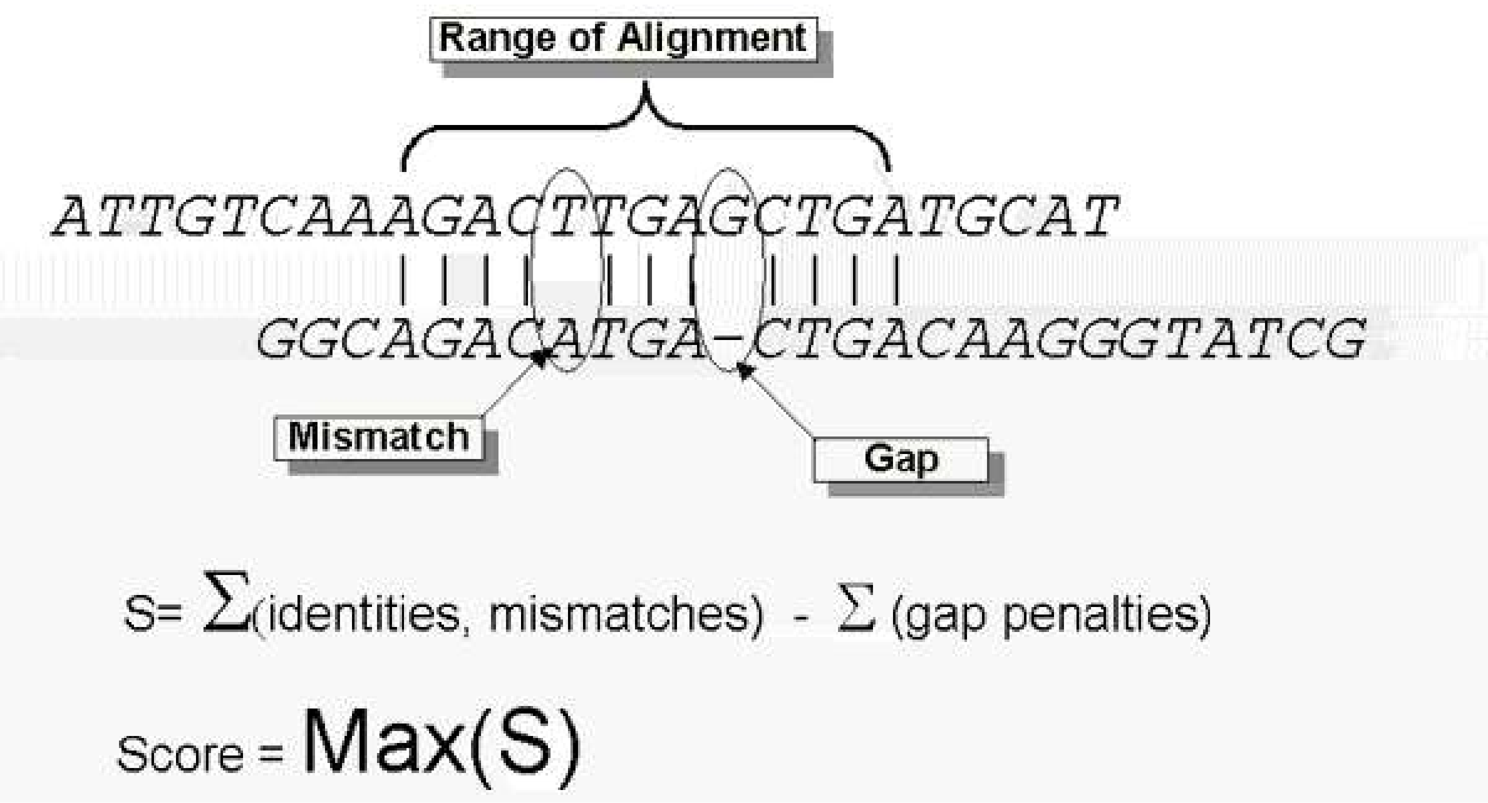
Home | About | Documentation | Submit Sequences | Retrieve Results | Display Submissions | Downloads
Document: Done

Srovnávání sekvencí

Různé kategorie podobnosti



Hodnocení podobností



$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

Zarovnání sekvencí

```
ACGTGA    ->   ACGTGA    ->
CGTG      ->    CGTG     ->   4

ACGTGA
TCGTA

ACGTGATGCAG
GGAGAGCACG

ACAGTTGACGAGATGGCAGGATGCGCGATGCAGCA
GACGAGCGTGAGTGCGATCGATGACAGTGTATAT
```

Zarovnání sekvencí

```
ACGTGA
 ::::          4
 CGTG

ACGTGA
 ::: :         4
TCGT-A

ACGTGATGCA-G
  : :: ::: :   7
 GGAGA-GCACG
```

Aligning Two Sequences

ATTGCAGTGATCG

ATTGCGTCGATCG

Solution 1:

```
ATTGCAGTGATCG
/////   /////
ATTGCGTCGATCG
```

Solution 2:

```
ATTGCAGT-GATCG
///// // /////
ATTGC-GTCGATCG
```

Which alignment is better?

```
ATTGCAGTGATCG
ATTGCGTCGATCG
```

Solution 1:

```
ATTGCAGTGATCG
|||||   |||||
ATTGCGTCGATCG
```

10 matches+ 3 mismatches

Solution 2:

```
ATTGCAGT-GATCG
||||| || |||||
ATTGC-GTCGATCG
```

12 matches+2 gaps

Scoring Scheme

Match	+1
Mismatch	-1
Indel	-2

Which alignment is better?

```
ATTGCAGTGATCG
ATTGCGTCGATCG
```

Solution 1:

```
ATTGCAGTGATCG
|||||   |||||
ATTGCGTCGATCG
```

Score=7

Solution 2:

```
ATTGCAGT-GATCG
||||| || |||||
ATTGC-GTCGATCG
```

Score=8

Finding the best alignment for long sequences is tedious

For two sequences of length 300 bases there are 10^{179} different alignments

↓

Dynamic programming

Dynamické programování

Needleman-Wunsch (1970)
Smith-Waterman (1981)

- První krok je triviální a pokrývá částečné řešení
- Každé další řešení je hodnoceno na základě předcházejících zjištění
- Zarovnání je tak postupně prodlužováno o další triviální úseky
- Opakování předchozích kroků vyústí v konečné řešení

Dynamic Programming Algorithm

```
Seq 1) * A G C
Seq 2) * A A A C
```

	A	G	C
0	1	2	3
0			
A 1			
A 2			
A 3			
C 4			

Needelman-Wunsch algorithm (1970)

Dynamic Programming Algorithm

```
* - - - - A G C
* A A A C

match=1
mismatch=-1
indel=-2
```

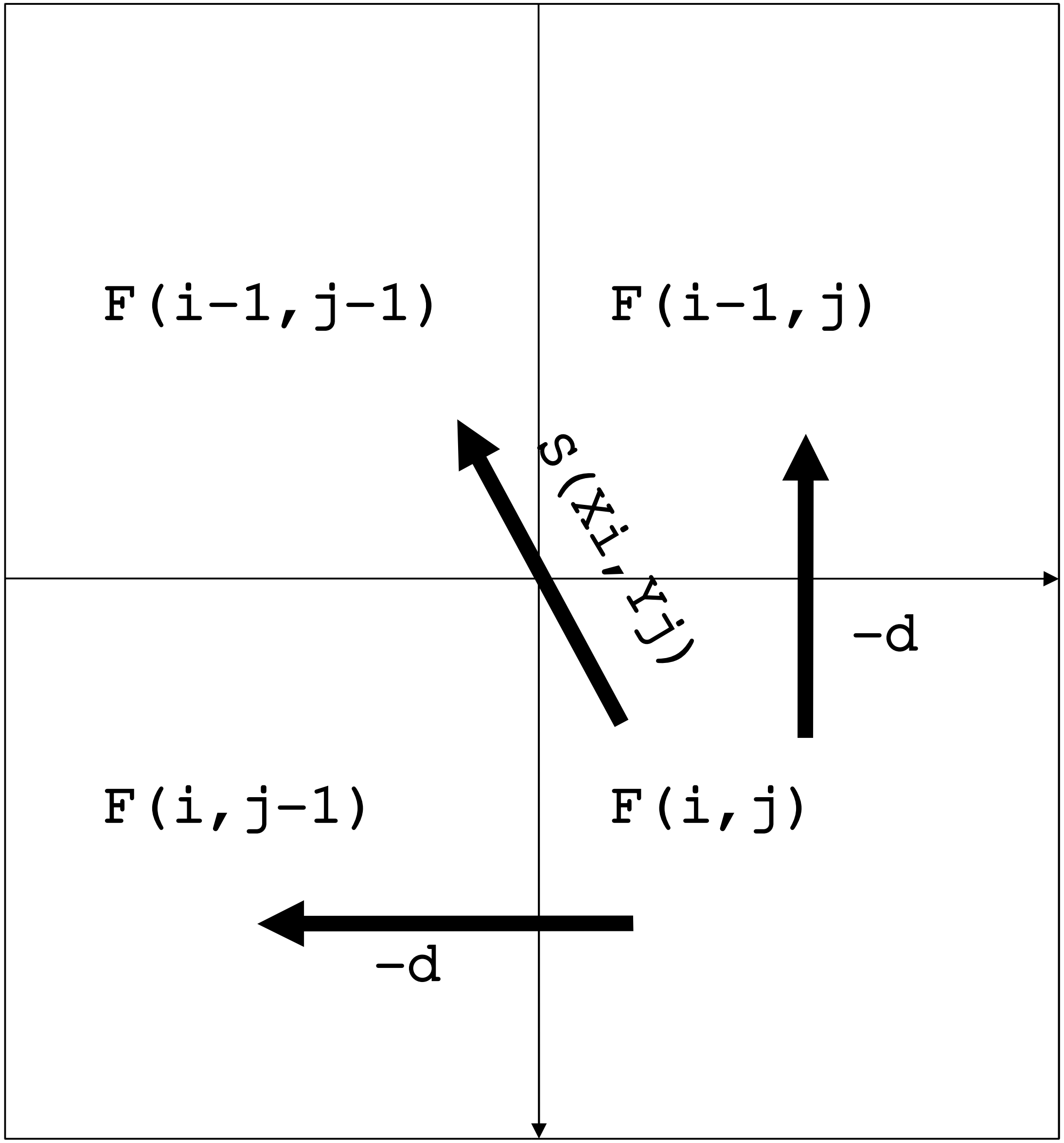
			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2			
A	2	-4			
A	3	-6			
C	4	-8			

Dynamic Programming Algorithm

```
* A G C
* A - - A A C

match=1
mismatch=-1
indel=-2
```

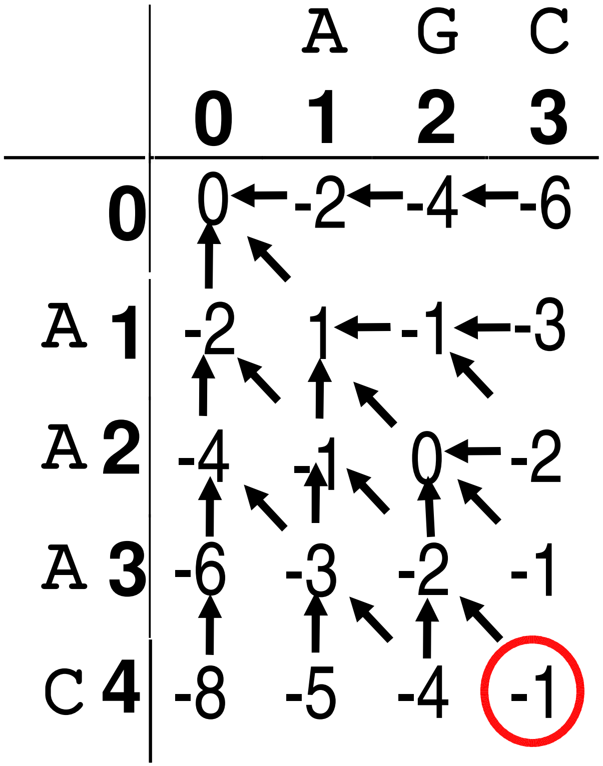
			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	↖ 1	← -1	← -3
A	2	-4			
A	3	-6			
C	4	-8			


F(i-1,j-1)
F(i-1,j)
S(xi,yj)
-d
F(i,j-1)
F(i,j)
-d

Global pairwise alignment

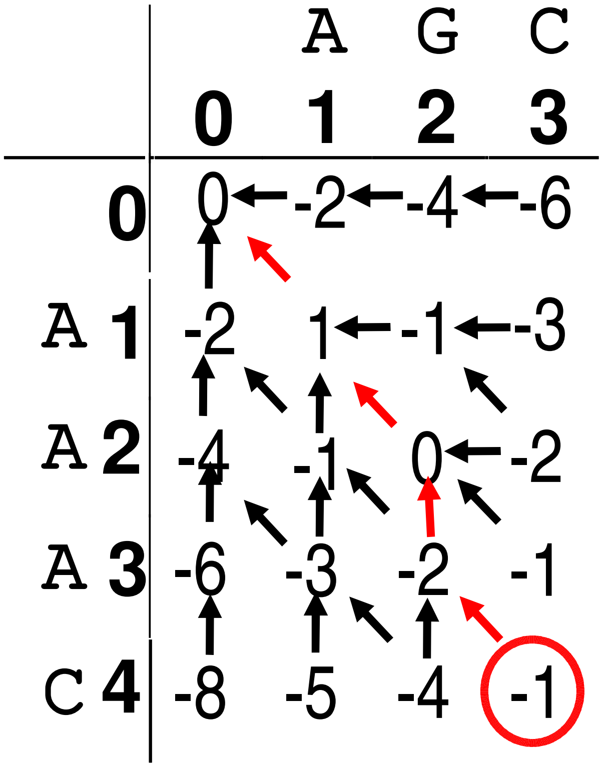
$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Finding the Best Score



			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	1	-1	-3
A	2	-4	-1	0	-2
A	3	-6	-3	-2	-1
C	4	-8	-5	-4	-1

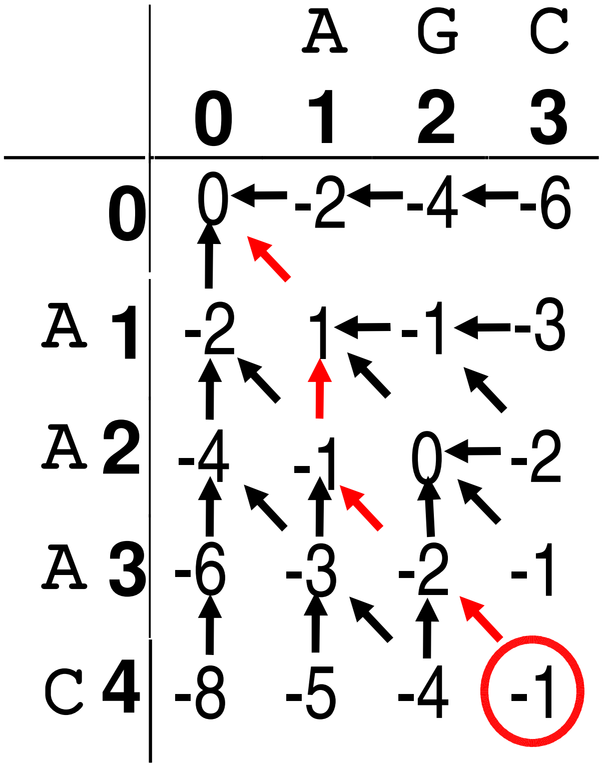
Tracing the Best Alignment



			A	G	C
		0	**1**	**2**	**3**
	0	0	-2	-4	-6
A	**1**	-2	1	-1	-3
A	**2**	-4	-1	0	-2
A	**3**	-6	-3	-2	-1
C	**4**	-8	-5	-4	-1

A G – C
A A A C

Tracing the Best Alignment

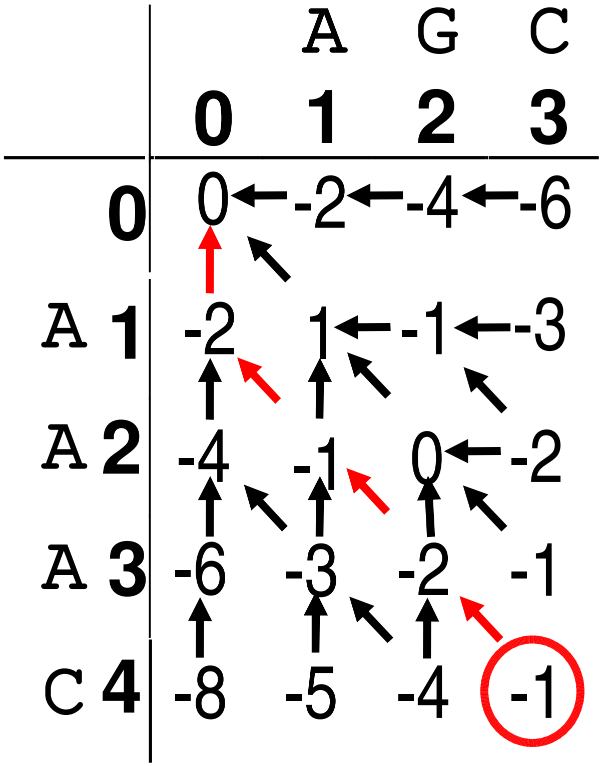


			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	1	-1	-3
A	2	-4	-1	0	-2
A	3	-6	-3	-2	-1
C	4	-8	-5	-4	-1

A – G C

A A A C

Tracing the Best Alignment



			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	1	-1	-3
A	2	-4	-1	0	-2
A	3	-6	-3	-2	-1
C	4	-8	-5	-4	-1

–	A	G	C
A	A	A	C

Local Alignment Example

ATCTAA
TAATA

			A	T	C	T	A	A
		0	1	2	3	4	5	6
	0							
T	1							
A	2							
A	3							
T	4							
A	5							

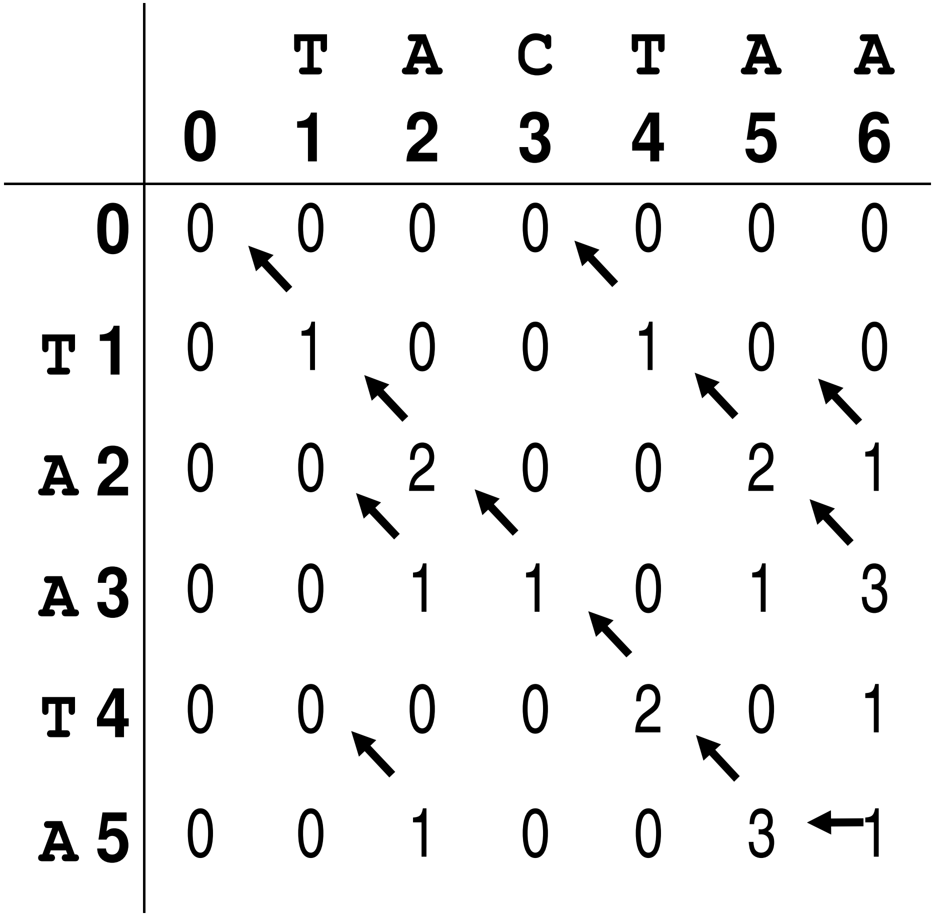
Smith-Waterman algorithm, 1981

Local Alignment

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \\ 0 \end{cases}$$

Local Alignment Example

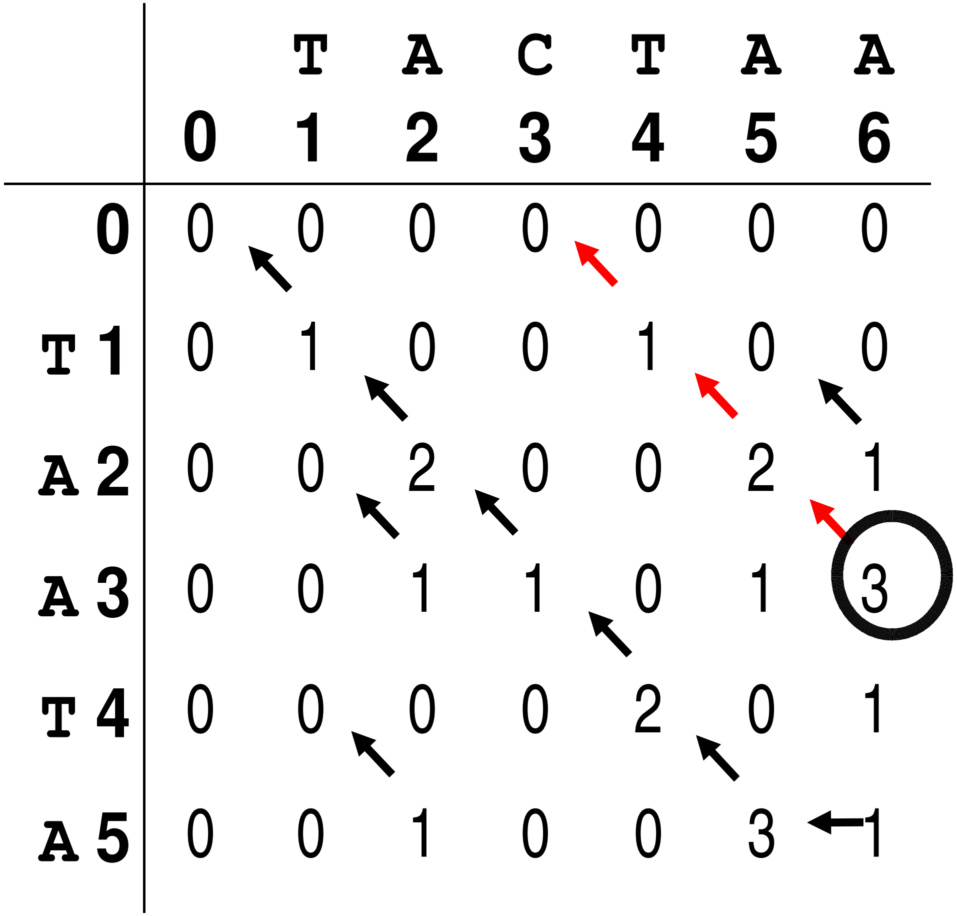
```
TCATAA
TAATA
```



			T	A	C	T	A	A
		0	1	2	3	4	5	6
	0	0	0	0	0	0	0	0
T	1	0	1	0	0	1	0	0
A	2	0	0	2	0	0	2	1
A	3	0	0	1	1	0	1	3
T	4	0	0	0	0	2	0	1
A	5	0	0	1	0	0	3	1

Local Alignment Example

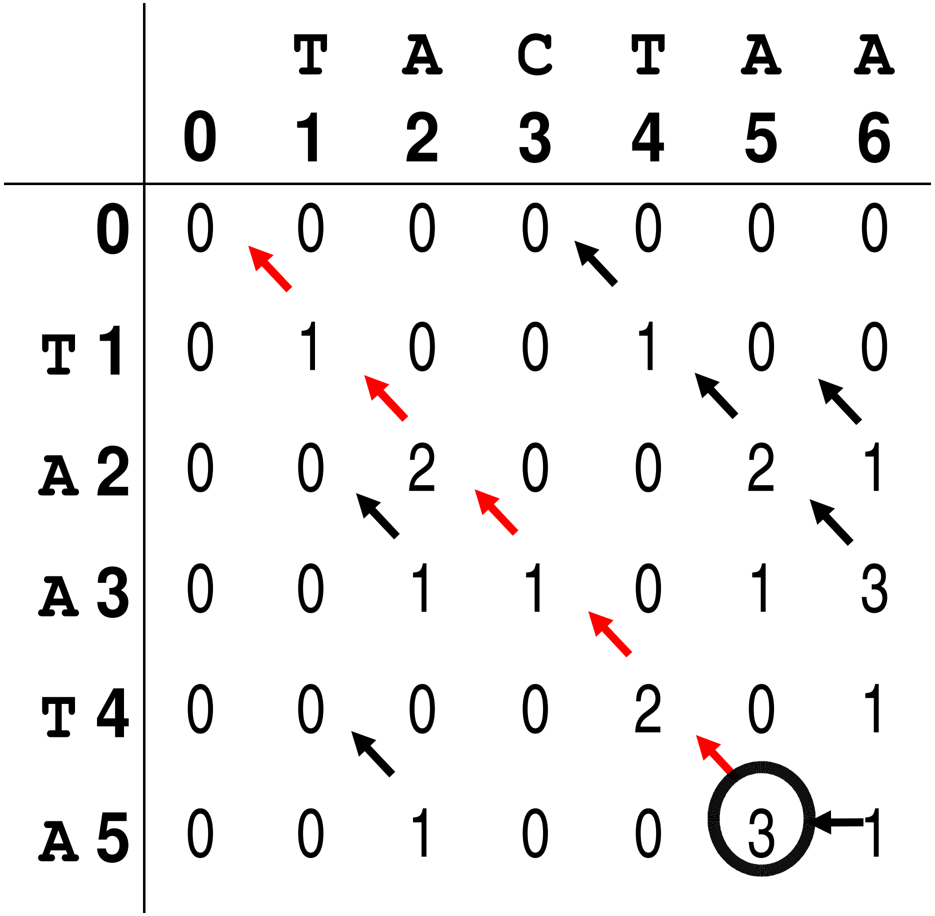
```
TACTAA
   TAATA
```



	0	T 1	A 2	C 3	T 4	A 5	A 6
0	0	0	0	0	0	0	0
T 1	0	1	0	0	1	0	0
A 2	0	0	2	0	0	2	1
A 3	0	0	1	1	0	1	3
T 4	0	0	0	0	2	0	1
A 5	0	0	1	0	0	3	1

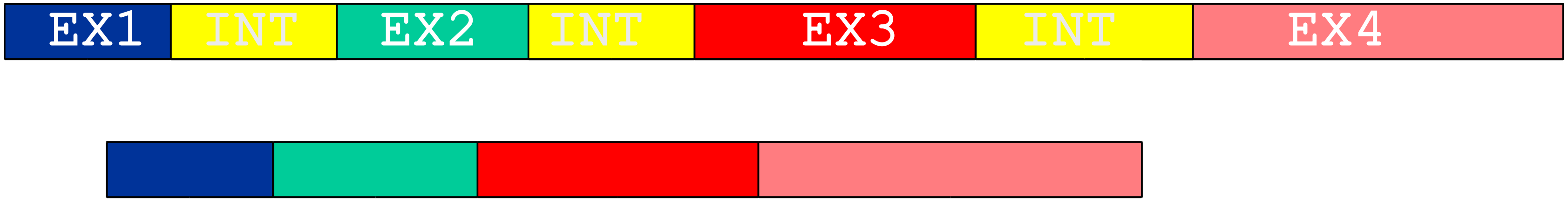
Local Alignment Example

TACTAA
TAATA

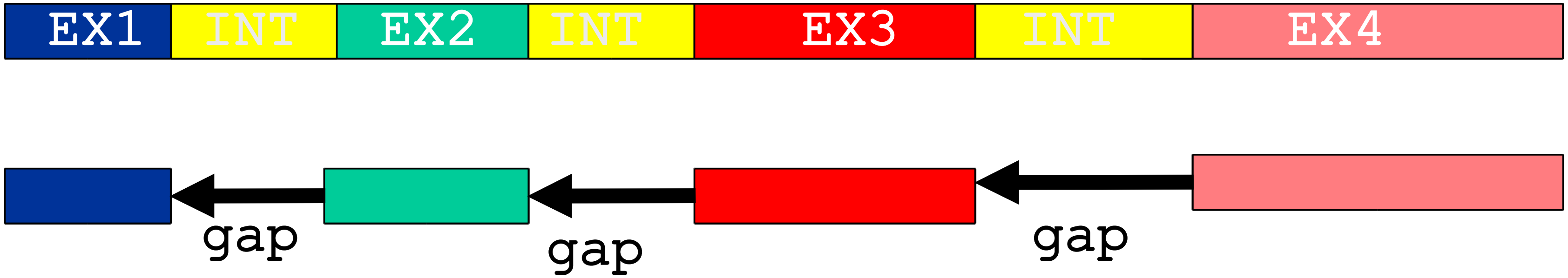


		T	A	C	T	A	A
	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
T 1	0	1	0	0	1	0	0
A 2	0	0	2	0	0	2	1
A 3	0	0	1	1	0	1	3
T 4	0	0	0	0	2	0	1
A 5	0	0	1	0	0	3	1

Examples :
Genomic DNA versus mRNA



Alignment



Gap Penalties

```
AAC-AATTAAG-ACTAC-GTTCATGAC

A-CGA-TTA-GCAC-ACTG-T-A-GA-
```

```
AACAATTAAGACTACGTTCATGAC---

AACAATT--------GTTCATGACGCA
```

Scoring Gaps

```
   AAC-AATTAAG-ACTAC-GTTCATGAC
I                                -6
   A-CGA-TTA-GCAC-ACTG-T-A-GA-

   AACAATTAAGACTACGTTCATGAC---
II                               12
   AACAATT--------GTTCATGACGCA
```

Scoring parameters
match:+1;Gap_open:-2

Scoring Insertions/Deletions

```
   AAC-AATTAAG-ACTAC-GTTCATGAC
                               -6
I
   A-CGA-TTA-GCAC-ACTG-T-A-GA-


   AACAATTAAGACTACGTTCATGAC---
                               -6
II
   AACAATT--------GTTCATGACGCA
```

Scoring parameters

match:+1;indel:-2

Considering Gap Opening and Gap Extension

```
   AAC-AATTAAG-ACTAC-GTTCATGAC
                               -17
I
   A-CGA-TTA-GCAC-ACTG-T-A-GA-
```

```
    AACAATTAAGACTACGTTCATGAC---
II                             1
    AACAATT--------GTTCATGACGCA
```

```
Scoring parameters
match:+1;Gap_open:-2; Gap_exten:-1
```