QBIO: Applied Bioinformatics

The Flow of Genetic Information

Dr. Jędrzej Jakub Szymański
Dr. Simon Maria Zumkeller
Prof. Björn Usadel

QBIO: Applied Bioinformatics

The Flow of Genetic Information

Where? → To Where?

Dr. Jędrzej Jakub Szymański
Dr. Simon Maria Zumkeller
Prof. Björn Usadel

QBIO: Applied Bioinformatics

The Flow of Genetic Information

Genetic sequence

Phenotype

Dr. Jędrzej Jakub Szymański
Dr. Simon Maria Zumkeller
Prof. Björn Usadel

QBIO: Applied Bioinformatics

The Flow of Genetic Information

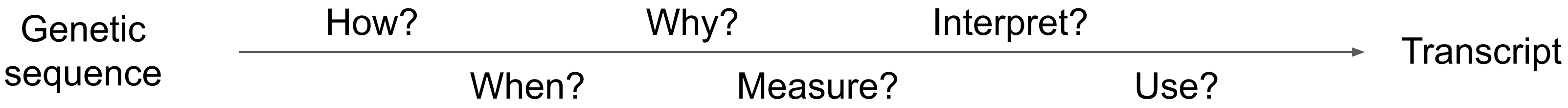
Genetic sequence

Transcript

Dr. Jędrzej Jakub Szymański
Dr. Simon Maria Zumkeller
Prof. Björn Usadel

QBIO: Applied Bioinformatics

The Flow of Genetic Information



Genetic sequence

How?

When?

Why?

Measure?

Interpret?

Use?

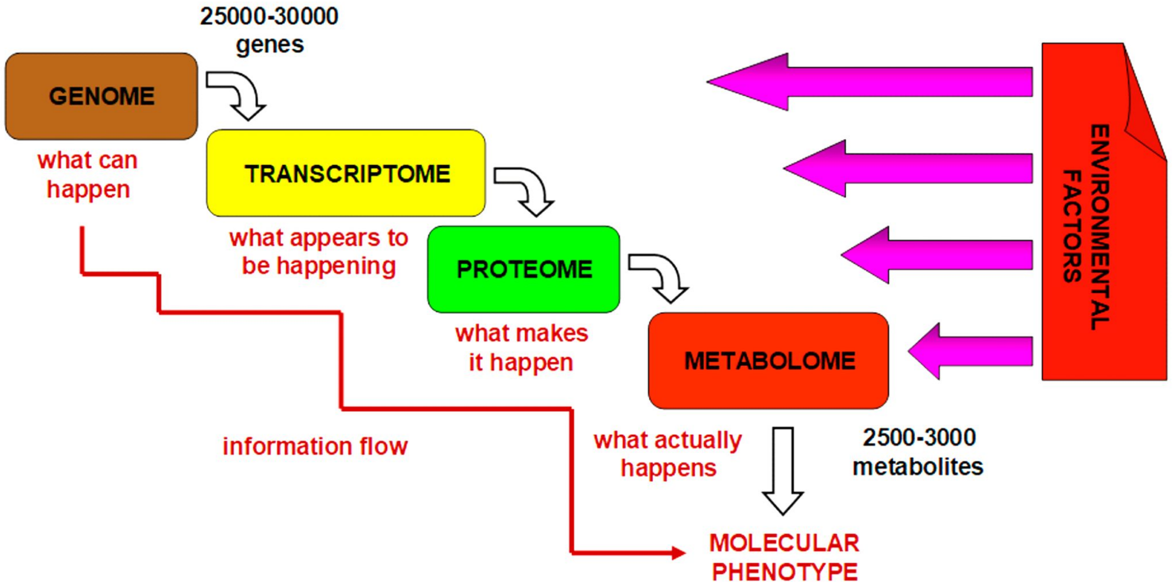
Transcript

Dr. Jędrzej Jakub Szymański
Dr. Simon Maria Zumkeller
Prof. Björn Usadel

Monday 2/10	Wednesday 4/10	Thursday 5/10	Friday 6/10	Friday 13/10	Friday 20/10
Setup	Sequencing technologies				
Gene transcription					
cis-egulatory elements					
Transcriptomics					
Experiment design					

Monday 2/10	Wednesday 4/10	Thursday 5/10	Friday 6/10	Friday 13/10	Friday 20/10
Setup		Normalization	Functional Enrichment		
Gene transcription	Sequencing technologies				
cis-egulatory elements		Data exploration	Networks		
Transcriptomics	Units				
Experiment design	Read mapping	DGE analysis	Public data		

Monday 2/10	Wednesday 4/10	Thursday 5/10	Friday 6/10	Friday 13/10	Friday 20/10
Setup		Normalization	Functional Enrichment	Lab 1	Lab 2
Gene transcription	Sequencing technologies				
cis-egulatory elements		Data exploration	Networks		Assignment
Transcriptomics	Units				
Experiment design	Read mapping	DGE analysis	Public data		



25000-30000 genes

GENOME

what can happen

TRANSCRIPTOME

what appears to be happening

PROTEOME

what makes it happen

METABOLOME

what actually happens

2500-3000 metabolites

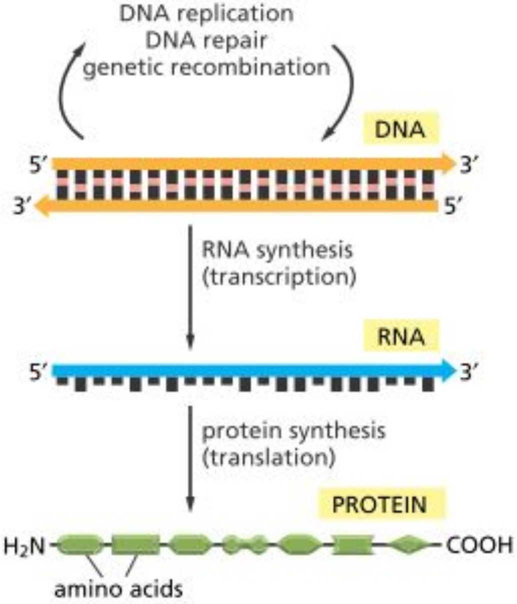
ENVIRONMENTAL FACTORS

information flow

MOLECULAR PHENOTYPE

J. Pers. Med. **2018**, *8*(3), 28; **https://doi.org/10.3390/jpm8030028**

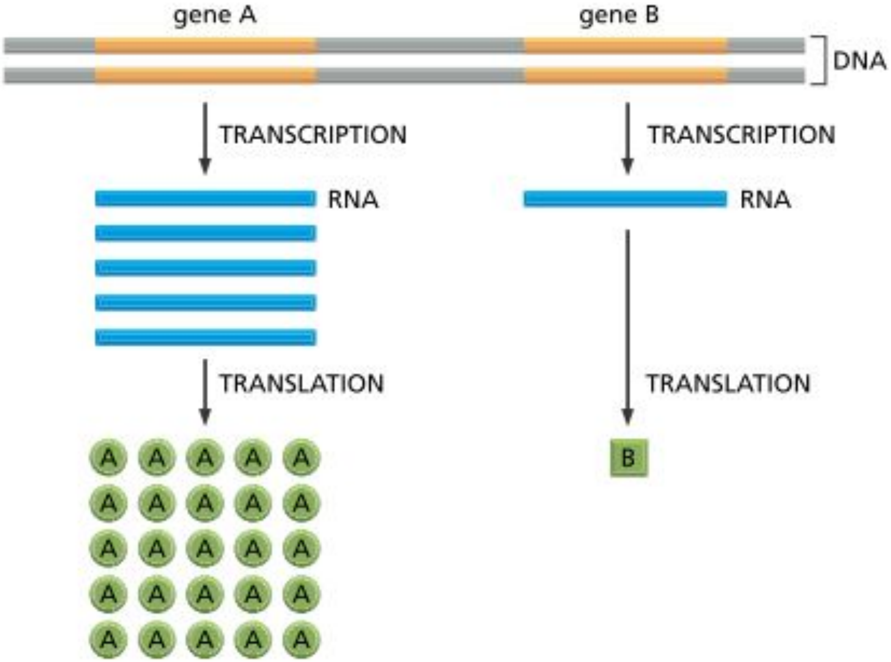
Just as a reminder



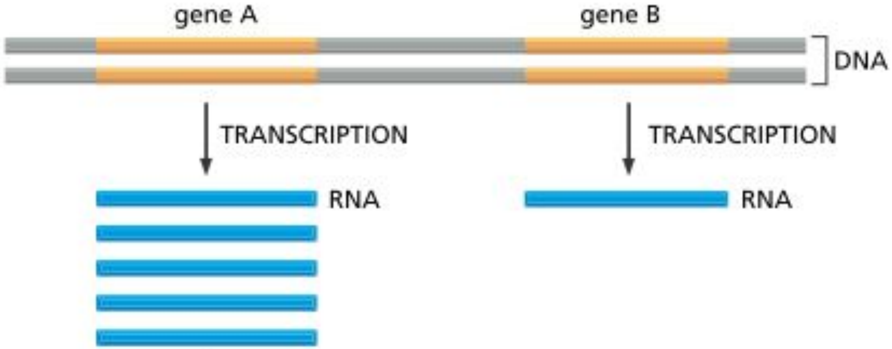
Alberts: Molecular Biology of the Cell

The amount of transcript being made is **not** the same for each gene.

Thus there a multiple levels of regulation affecting protein amount.



Here we will focus on the DNA → RNA story



The cell contains many different types of RNA.

mRNA is studied for its role in gene regulation,

rRNA often makes up the **bulk amount of RNA in a cell ~80%**

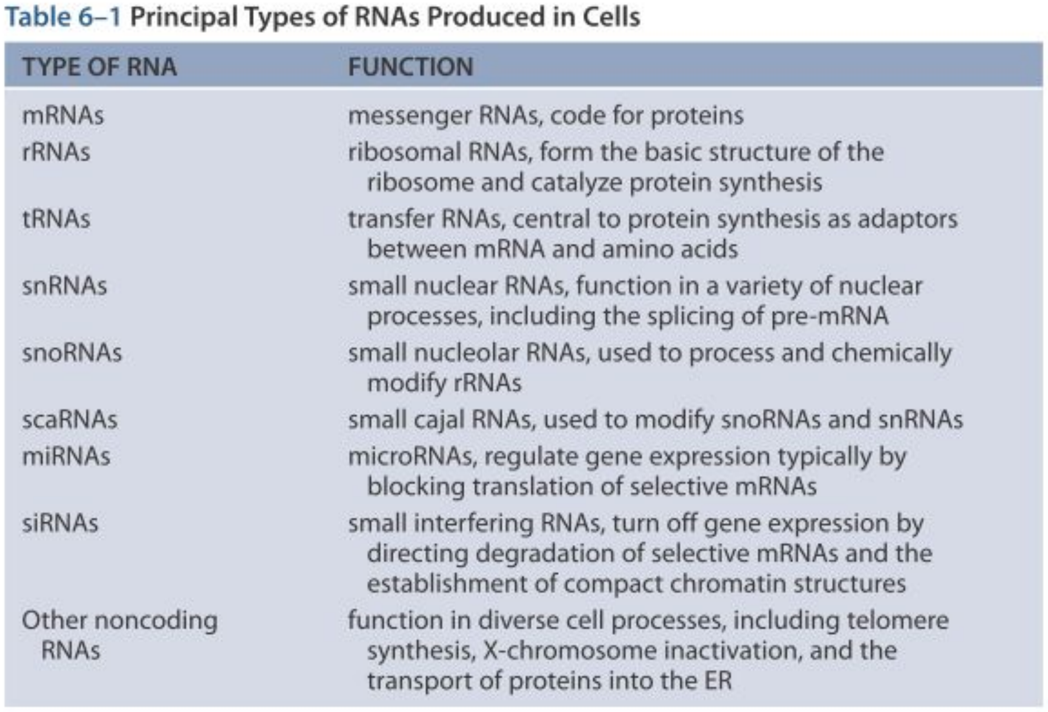
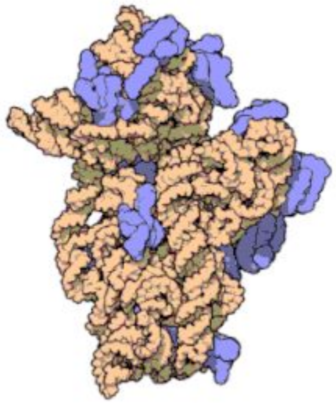


Table 6–1 Principal Types of RNAs Produced in Cells

TYPE OF RNA	FUNCTION
mRNAs	messenger RNAs, code for proteins
rRNAs	ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	small nucleolar RNAs, used to process and chemically modify rRNAs
scaRNAs	small cajal RNAs, used to modify snoRNAs and snRNAs
miRNAs	microRNAs, regulate gene expression typically by blocking translation of selective mRNAs
siRNAs	small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
Other noncoding RNAs	function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER



Alberts: Molecular Biology of the Cell

RNA is made by RNA polymerases which are large multi-subunit enzymes in eukaryotes

Eukaryotes have **at least three RNA polymerases**:

- **RNA polymerase I (Pol I)** transcribes large ribosomal RNA (rRNA) genes
- **RNA polymerase II (Pol II)** transcribes messenger RNA (mRNA) genes
- **RNA polymerase III (Pol III)** transcribes a variety of RNAs including transfer RNA (tRNA) and 5S ribosomal RNA
- Plants have a fourth RNA polymerase that transcribes regulatory RNAs
- some plants have a fifth RNA polymerase

Bacteria and archaea have a **single RNA polymerase**

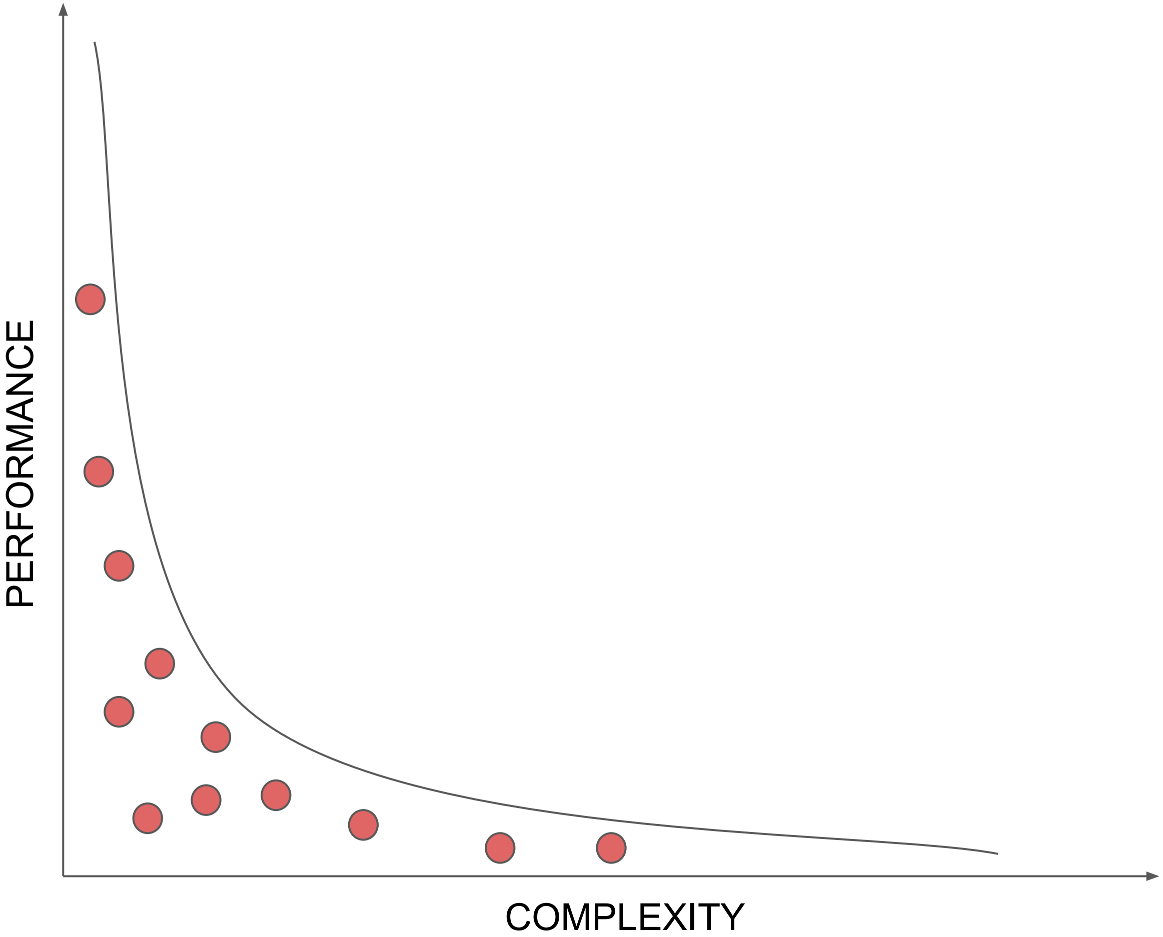
RNA is made by RNA polymerases which are large multi-subunit enzymes in eukaryotes

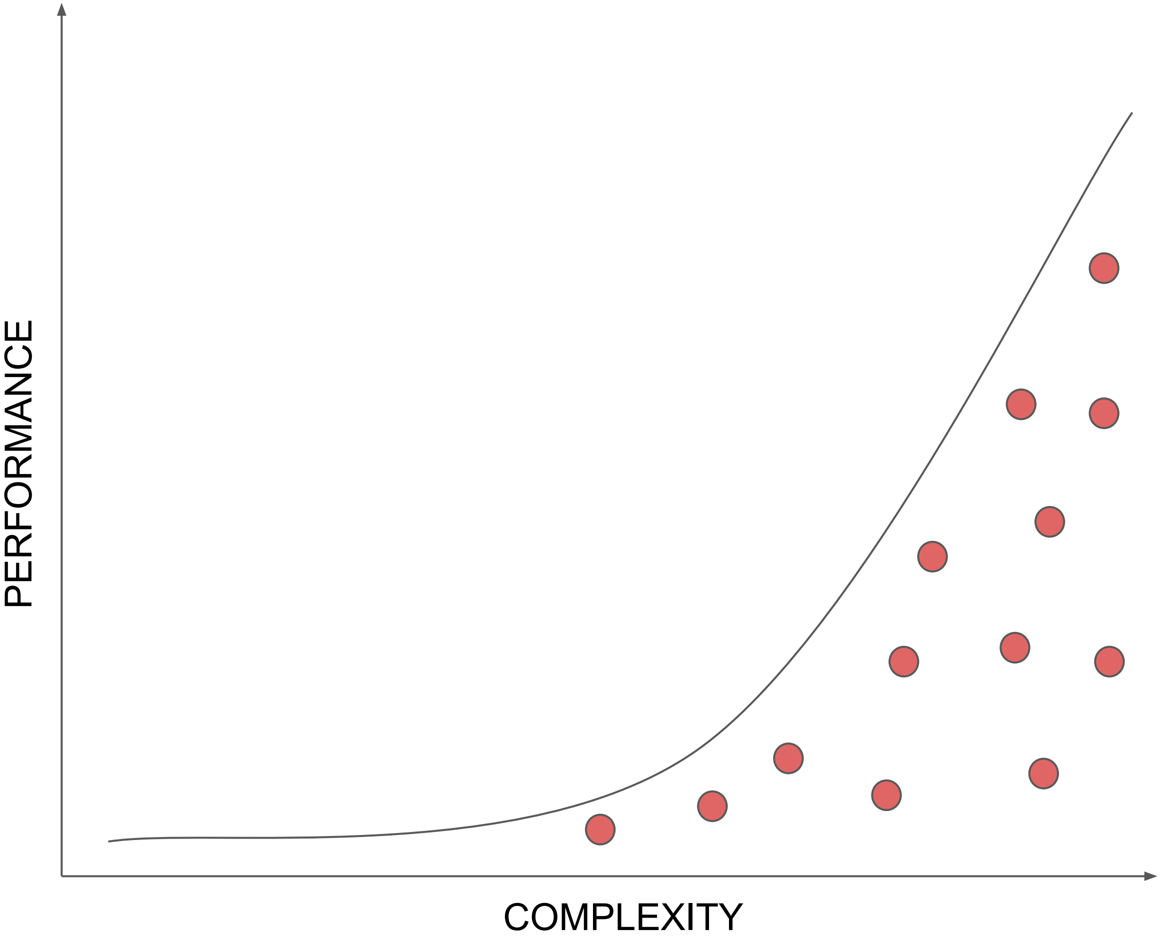
Eukaryotes have **at least three RNA polymerases**:

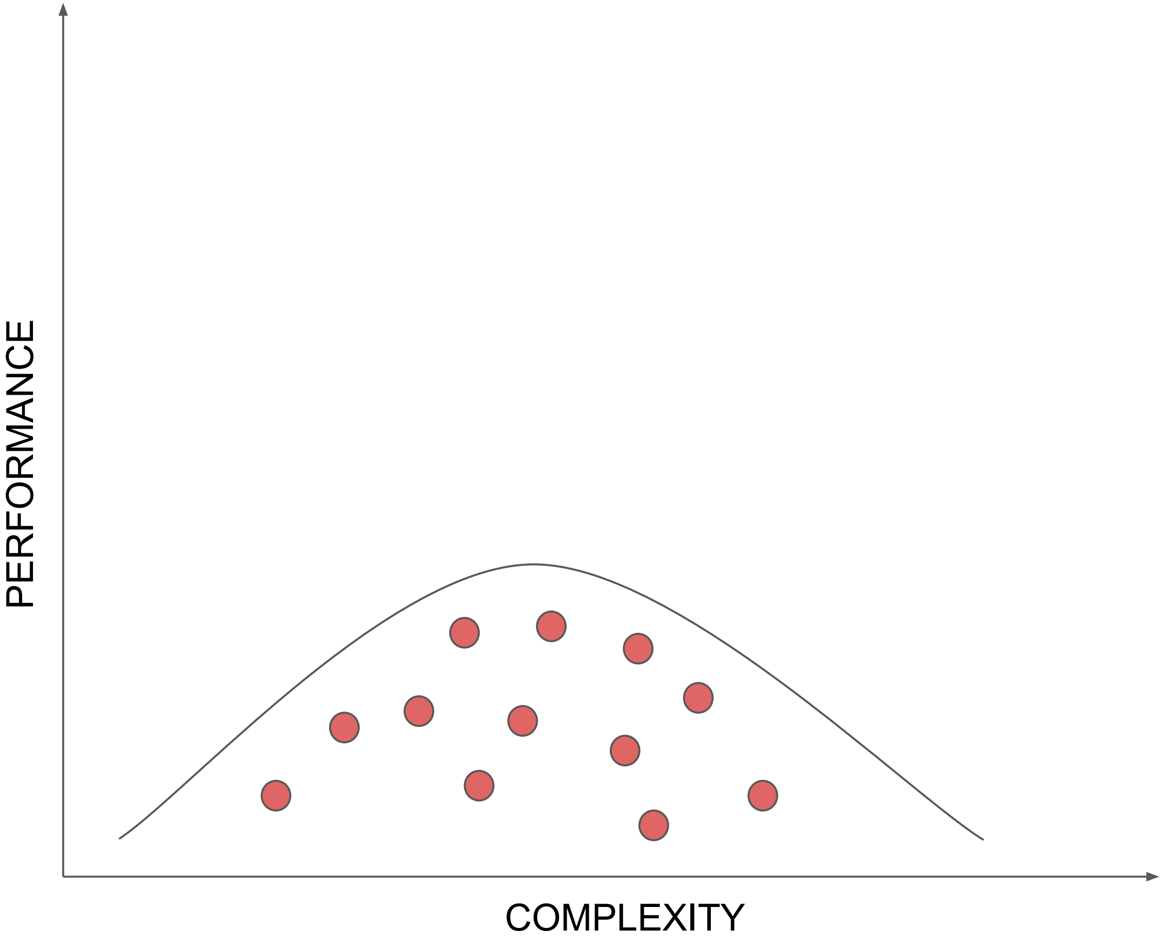
- **RNA polymerase I (Pol I)** transcribes large ribosomal RNA (rRNA) genes
- **RNA polymerase II (Pol II)** transcribes messenger RNA (mRNA) genes
- **RNA polymerase III (Pol III)** transcribes a variety of RNAs including transfer RNA (tRNA) and 5S ribosomal RNA
- Plants have a fourth RNA polymerase that transcribes regulatory RNAs
- some plants have a fifth RNA polymerase

WHY COMPLEXITY?

Bacteria and archaea have a **single RNA polymerase**


PERFORMANCE
COMPLEXITY


PERFORMANCE
COMPLEXITY


PERFORMANCE
COMPLEXITY

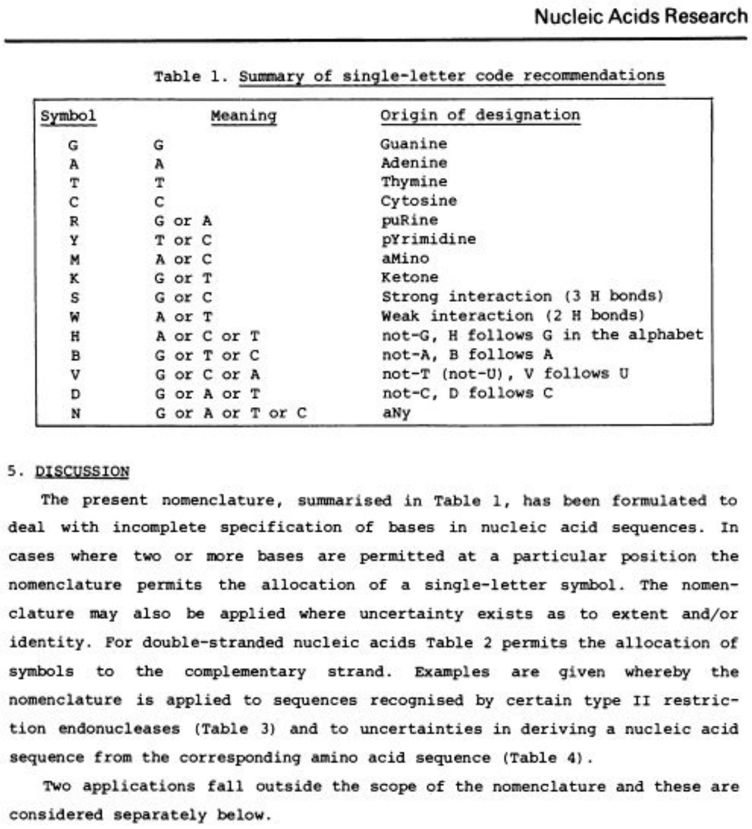
RNA is made by RNA polymerases which are large multi-subunit enzymes in eukaryotes

Eukaryotes have **at least three RNA polymerases**:

- **RNA polymerase I (Pol I)** transcribes large ribosomal RNA (rRNA) genes
- **RNA polymerase II (Pol II)** transcribes messenger RNA (mRNA) genes
- **RNA polymerase III (Pol III)** transcribes a variety of RNAs including transfer RNA (tRNA) and 5S ribosomal RNA
- Plants have a fourth RNA polymerase that transcribes regulatory RNAs
- some plants have a fifth RNA polymerase

Bacteria and archaea have a **single RNA polymerase**

CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA



Nucleic Acids Research

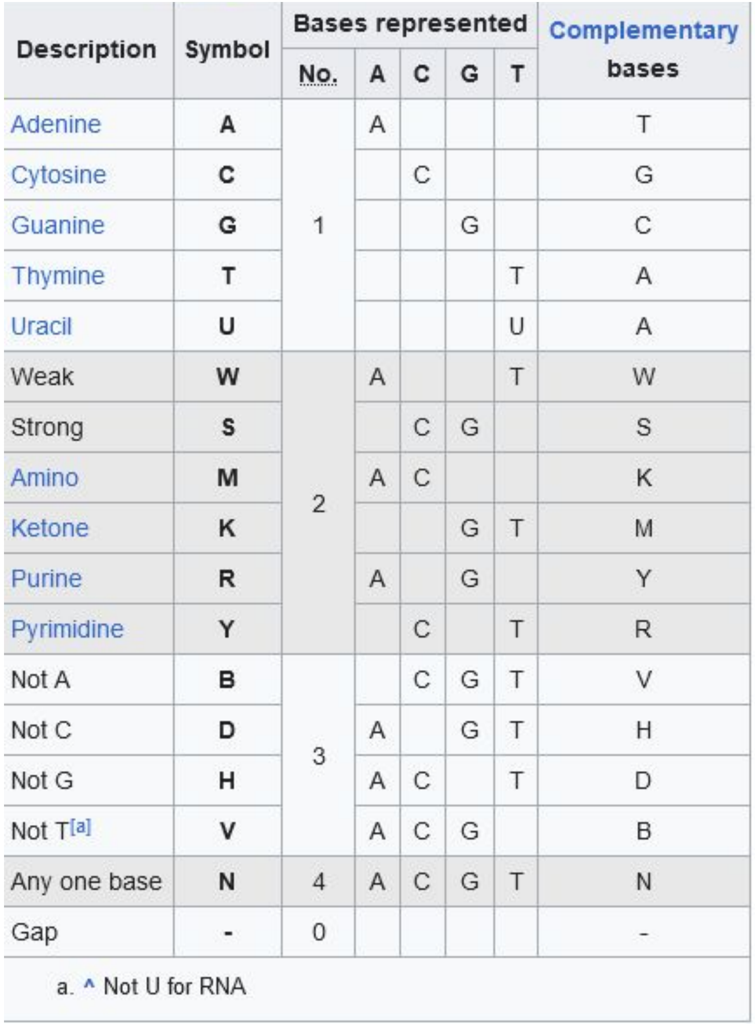
Table 1. Summary of single-letter code recommendations

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Ketone
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

5. DISCUSSION

The present nomenclature, summarised in Table 1, has been formulated to deal with incomplete specification of bases in nucleic acid sequences. In cases where two or more bases are permitted at a particular position the nomenclature permits the allocation of a single-letter symbol. The nomenclature may also be applied where uncertainty exists as to extent and/or identity. For double-stranded nucleic acids Table 2 permits the allocation of symbols to the complementary strand. Examples are given whereby the nomenclature is applied to sequences recognised by certain type II restriction endonucleases (Table 3) and to uncertainties in deriving a nucleic acid sequence from the corresponding amino acid sequence (Table 4).

Two applications fall outside the scope of the nomenclature and these are considered separately below.



Description	Symbol	Bases represented					Complementary bases
		No.	A	C	G	T	
Adenine	A	1	A				T
Cytosine	C			C			G
Guanine	G				G		C
Thymine	T					T	A
Uracil	U					U	A
Weak	W	2	A			T	W
Strong	S			C	G		S
Amino	M		A	C			K
Ketone	K				G	T	M
Purine	R		A		G		Y
Pyrimidine	Y			C		T	R
Not A	B	3		C	G	T	V
Not C	D		A		G	T	H
Not G	H		A	C		T	D
Not T[a]	V		A	C	G		B
Any one base	N	4	A	C	G	T	N
Gap	-	0					-

a. ^ Not U for RNA

Cornish-Bowden (1985) 13(9): 3021–3030. wikipedia

CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTCCCTCCTCTGCACCATGACTACCTGCAGCCGCCAGTTCACCTCCTC
CAGCTCCATGAAGGGCTCCTGCGGCATCGGGGGCGGCATCGGGGCGGGCTCCAGCCGCATCTCCTCCGTC
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
CCTCTGGGGGAGCCTATGGGTTGGGGGGCGGCTATGGCGGTGGCTTCAGCAGCAGCAGCAGCAGCTTTGG
TAGTGGCTTTGGGGGAGGATATGGTGGTGGCCTTGGTGCTGGCTTGGGTGGTGGCTTTGGTGGTGGCTTT
GCTGGTGGTGATGGGCTTCTGGTGGGCAGTGAGAAGGTGACCATGCAGAACCTCAATGACCGCCTGGCCT
CCTACCTGGACAAGGTGCGTGCTCTGGAGGAGGCCAACGCCGACCTGGAAGTGAAGATCCGTGACTGGTA
CCAGAGGCAGCGGCCTGCTGAGATCAAAGACTACAGTCCCTACTTCAAGACCATTGAGGACCTGAGGAAC
AAGGTGGGTGAATGGGCAGCAGAAGGCACCATTCCAGCTAGCTCCTTCTGGGAACAATTCATGCCCCAGG
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC
```

We will in almost all cases only write **one strand** of DNA in the FASTA format

FASTA format
One line with “>” then identifier
Multiple lines with sequence typically 80,120 etc characters per line

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTCCCTCCTCTGCACCATGACTACCTGCAGCCGCCAGTTCACCTCCTC
CAGCTCCATGAAGGGCTCCTGCGGCATCGGGGGCGGCATCGGGGCGGGCTCCAGCCGCATCTCCTCCGTC
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
CCTCTGGGGGAGCCTATGGGTTGGGGGCGGCTATGGCGGTGGCTTCAGCAGCAGCAGCAGCAGCTTTGG
TAGTGGCTTTGGGGGAGGATATGGTGGTGGCCTTGGTGCTGGCTTGGGTGGTGGCTTTGGTGGTGGCTTT
GCTGGTGGTGATGGGCTTCTGGTGGGCAGTGAGAAGGTGACCATGCAGAACCTCAATGACCGCCTGGCCT
CCTACCTGGACAAGGTGCGTGCTCTGGAGGAGGCCAACGCCGACCTGGAAGTGAAGATCCGTGACTGGTA
CCAGAGGCAGCGGCCTGCTGAGATCAAAGACTACAGTCCCTACTTCAAGACCATTGAGGACCTGAGGAAC
AAGGTGGGTGAATGGGCAGCAGAAGGCACCATTCCAGCTAGCTCCTTCTGGGAACAATTCATGCCCCAGG
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC
```

We will in almost all cases only write **one strand** of DNA in the FASTA format

FASTA format

One line with “>” then identifier

Multiple lines with sequence typically 80,120 etc characters per line

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
```

Some older programs can only parse few characters in the identifier or expect certain line lengths

```
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
CCTCTGGGGGAGCCTATGGGTTGGGGGGCGGCTATGGCGGTGGCTTCAGCAGCAGCAGCAGCAGCTTTGG
```

Many programs have issues with line endings

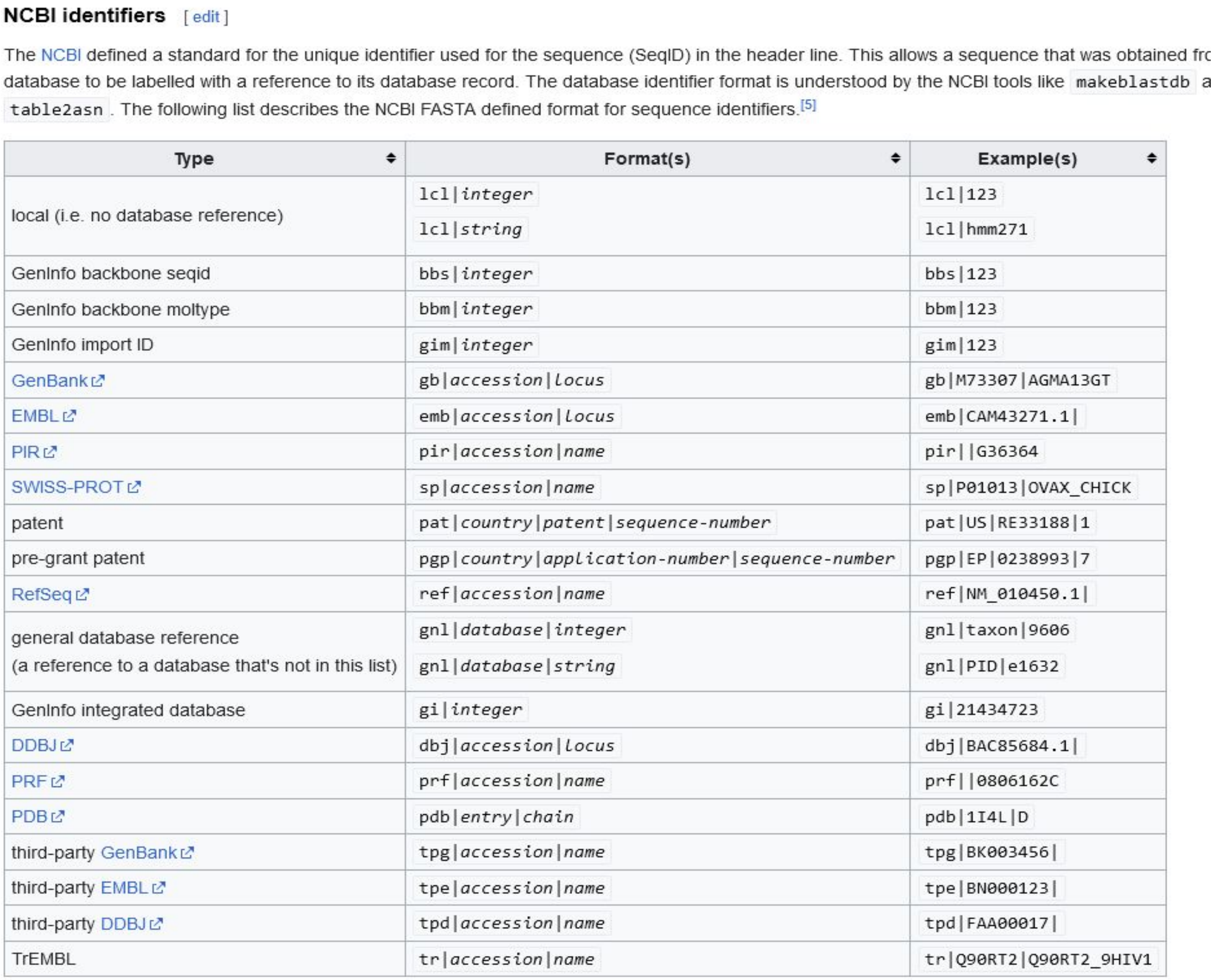
This is /r/n in windows CR LF

/n in Linux LF

And

/r in some old Macs CR

```
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC
```

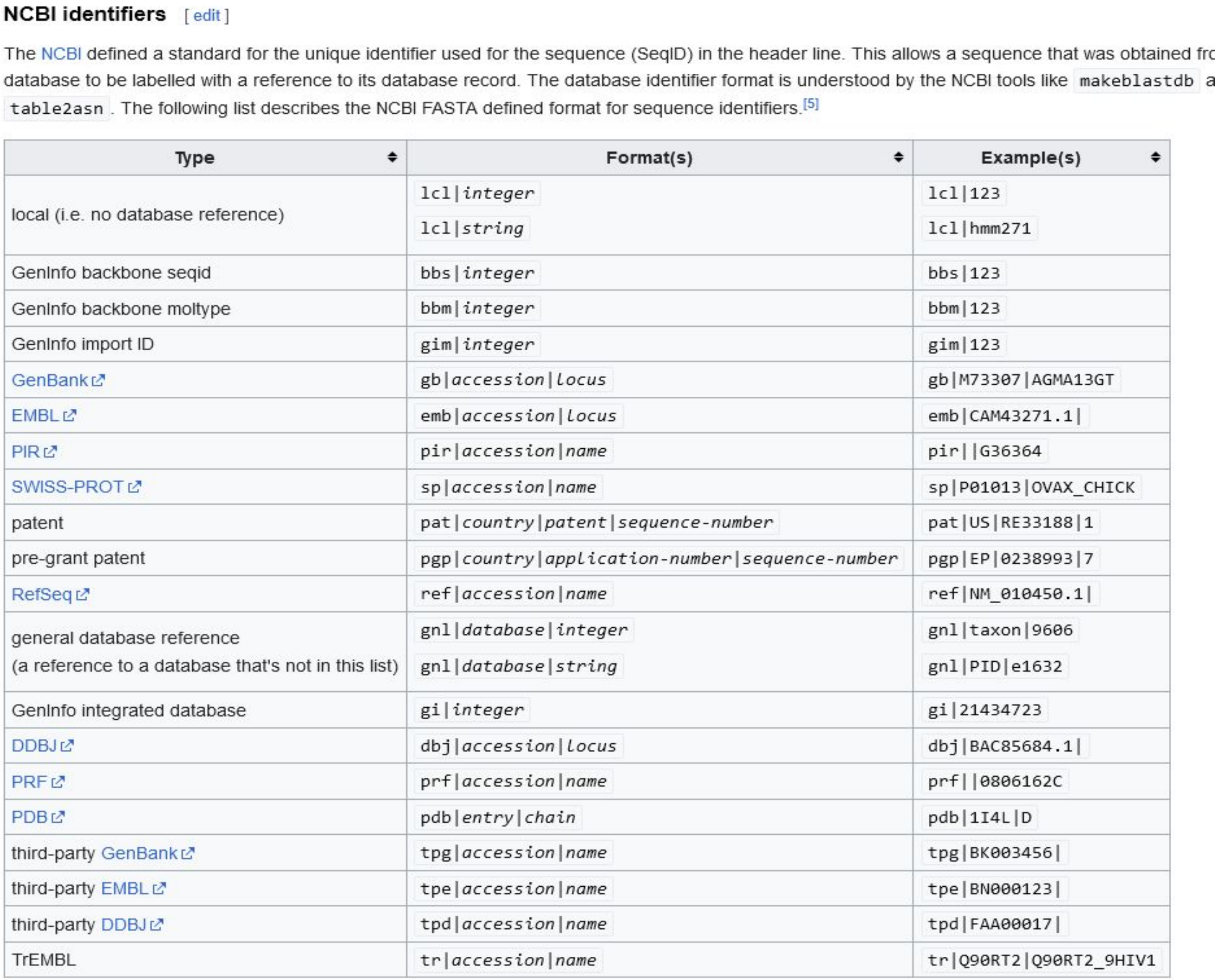


NCBI identifiers [edit]

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. This allows a sequence that was obtained frc database to be labelled with a reference to its database record. The database identifier format is understood by the NCBI tools like `makeblastdb` a `table2asn`. The following list describes the NCBI FASTA defined format for sequence identifiers.[5]

Type	Format(s)	Example(s)
local (i.e. no database reference)	lcl|*integer* lcl|*string*	lcl|123 lcl|hmm271
GenInfo backbone seqid	bbs|*integer*	bbs|123
GenInfo backbone moltype	bbm|*integer*	bbm|123
GenInfo import ID	gim|*integer*	gim|123
GenBank	gb|*accession*|*locus*	gb|M73307|AGMA13GT
EMBL	emb|*accession*|*locus*	emb|CAM43271.1|
PIR	pir|*accession*|*name*	pir||G36364
SWISS-PROT	sp|*accession*|*name*	sp|P01013|OVAX_CHICK
patent	pat|*country*|*patent*|*sequence-number*	pat|US|RE33188|1
pre-grant patent	pgp|*country*|*application-number*|*sequence-number*	pgp|EP|0238993|7
RefSeq	ref|*accession*|*name*	ref|NM_010450.1|
general database reference (a reference to a database that's not in this list)	gnl|*database*|*integer* gnl|*database*|*string*	gnl|taxon|9606 gnl|PID|e1632
GenInfo integrated database	gi|*integer*	gi|21434723
DDBJ	dbj|*accession*|*locus*	dbj|BAC85684.1|
PRF	prf|*accession*|*name*	prf||0806162C
PDB	pdb|*entry*|*chain*	pdb|1I4L|D
third-party GenBank	tpg|*accession*|*name*	tpg|BK003456|
third-party EMBL	tpe|*accession*|*name*	tpe|BN000123|
third-party DDBJ	tpd|*accession*|*name*	tpd|FAA00017|
TrEMBL	tr|*accession*|*name*	tr|Q90RT2|Q90RT2_9HIV1

wikipedia



NCBI identifiers [edit]

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. This allows a sequence that was obtained frc database to be labelled with a reference to its database record. The database identifier format is understood by the NCBI tools like `makeblastdb` a `table2asn`. The following list describes the NCBI FASTA defined format for sequence identifiers.[5]

Type	Format(s)	Example(s)
local (i.e. no database reference)	lcl|*integer* lcl|*string*	lcl|123 lcl|hmm271
GenInfo backbone seqid	bbs|*integer*	bbs|123
GenInfo backbone moltype	bbm|*integer*	bbm|123
GenInfo import ID	gim|*integer*	gim|123
GenBank	gb|*accession*|*locus*	gb|M73307|AGMA13GT
EMBL	emb|*accession*|*locus*	emb|CAM43271.1|
PIR	pir|*accession*|*name*	pir||G36364
SWISS-PROT	sp|*accession*|*name*	sp|P01013|OVAX_CHICK
patent	pat|*country*|*patent*|*sequence-number*	pat|US|RE33188|1
pre-grant patent	pgp|*country*|*application-number*|*sequence-number*	pgp|EP|0238993|7
RefSeq	ref|*accession*|*name*	ref|NM_010450.1|
general database reference (a reference to a database that's not in this list)	gnl|*database*|*integer* gnl|*database*|*string*	gnl|taxon|9606 gnl|PID|e1632
GenInfo integrated database	gi|*integer*	gi|21434723
DDBJ	dbj|*accession*|*locus*	dbj|BAC85684.1|
PRF	prf|*accession*|*name*	prf||0806162C
PDB	pdb|*entry*|*chain*	pdb|1I4L|D
third-party GenBank	tpg|*accession*|*name*	tpg|BK003456|
third-party EMBL	tpe|*accession*|*name*	tpe|BN000123|
third-party DDBJ	tpd|*accession*|*name*	tpd|FAA00017|
TrEMBL	tr|*accession*|*name*	tr|Q90RT2|Q90RT2_9HIV1

WHY COMPLEXITY?

wikipedia

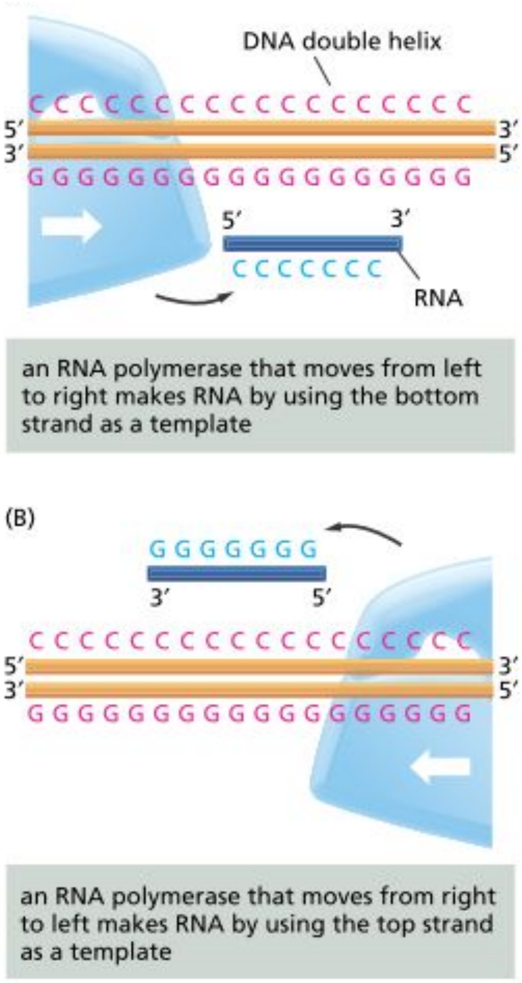
We will in almost all cases only write **one strand** of DNA in the FASTA format

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTC...
```

We will in almost all cases only write **one strand** of DNA in the FASTA format

When studying genomes it is important to keep in mind that genes can be **encoded on both strands of the DNA**.

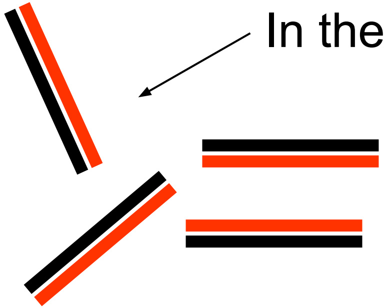
```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTC...
```


DNA double helix
C C C C C C C C C C C C C C C C C C
5′
3′
3′
5′
G G G G G G G G G G G G G G G G G G
5′
3′
C C C C C C C
RNA
(B)
G G G G G G G
3′
5′
C C C C C C C C C C C C C C C C C C
5′
3′
3′
5′
G G G G G G G G G G G G G G G G G G

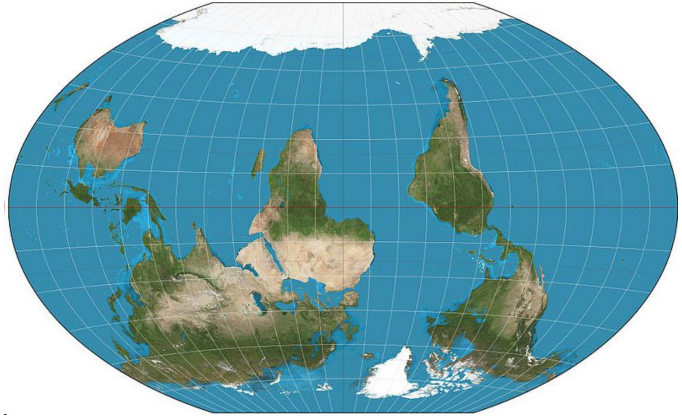
an RNA polymerase that moves from left to right makes RNA by using the bottom strand as a template

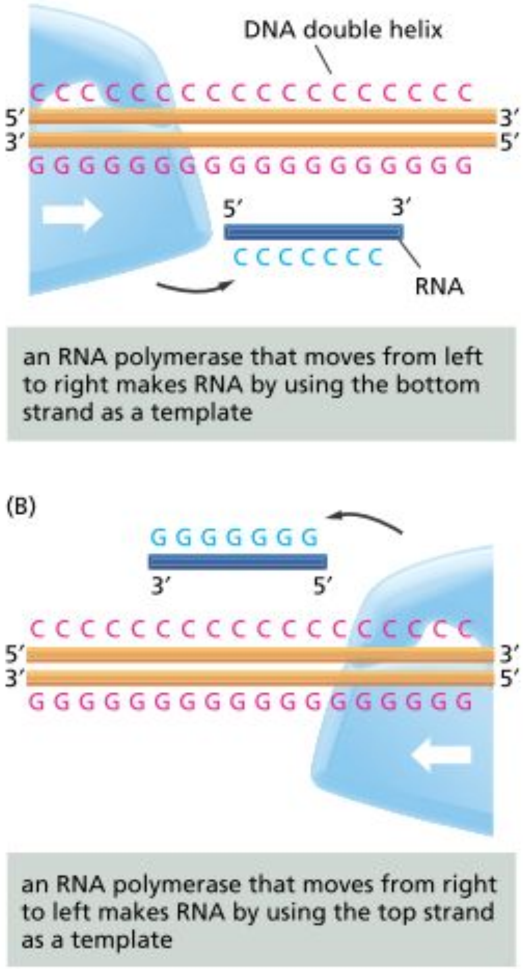
an RNA polymerase that moves from right to left makes RNA by using the top strand as a template

Alberts: Molecular Biology of the Cell



In the cell




DNA double helix
C C C C C C C C C C C C C C C C C C
5′
3′
3′
5′
G G G G G G G G G G G G G G G G G G
5′
3′
C C C C C C C
RNA
(B)
G G G G G G G
3′
5′
C C C C C C C C C C C C C C C C C C
5′
3′
3′
5′
G G G G G G G G G G G G G G G G G G

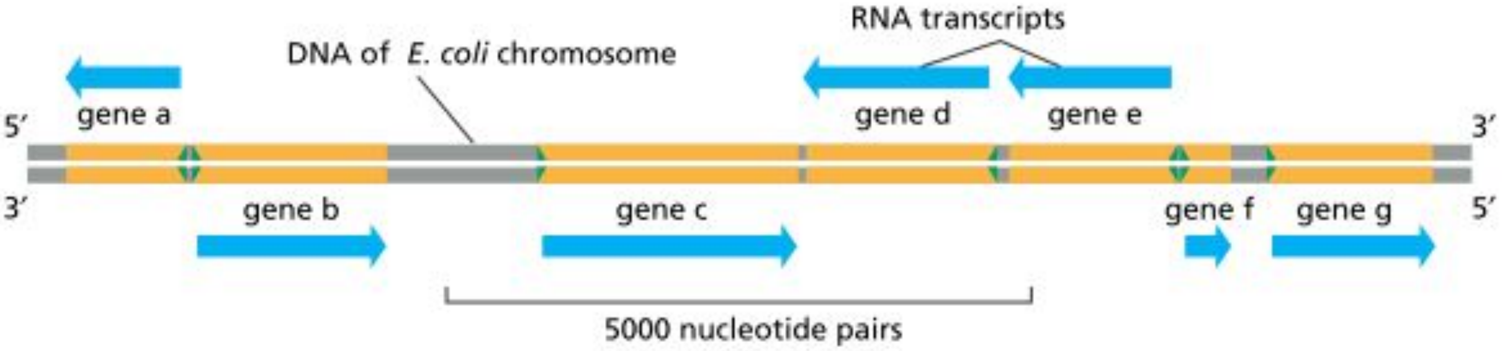
an RNA polymerase that moves from left to right makes RNA by using the bottom strand as a template

an RNA polymerase that moves from right to left makes RNA by using the top strand as a template

Alberts: Molecular Biology of the Cell

But how does an RNA polymerase know which strand to read from and where to start transcription?

Promoter elements direct the RNA Polymerase. These regions on the DNA often consist of short **DNA stretches with conserved sequence** to which the some auxiliary factors bind.



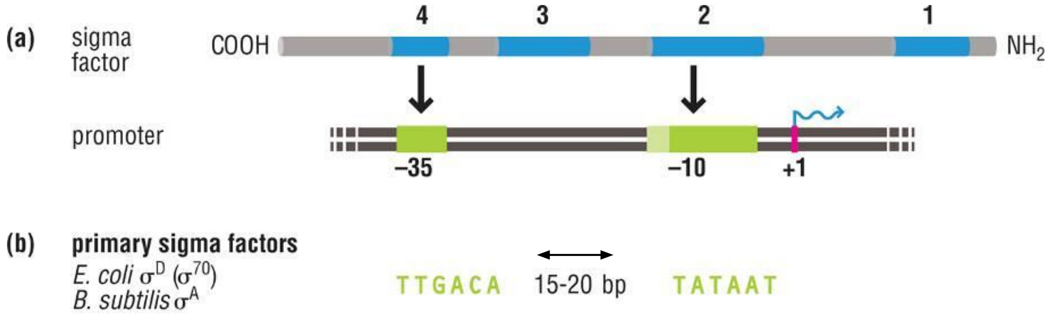
▲ Promoter elements

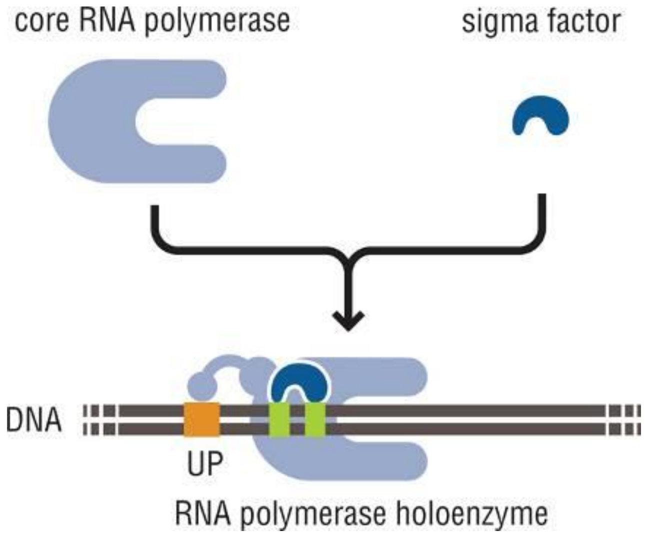
Alberts: Molecular Biology of the Cell

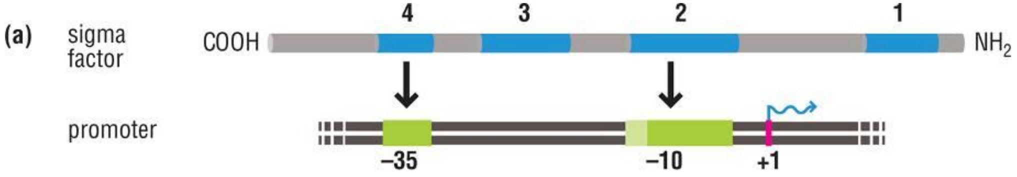
Basal bacterial promoters generally have two elements: a -35 element and a -10 element. These are roughly 35 and 10 bases upstream of the transcription start site

Sigma factors bind sequences that define the bacterial promoters and each sigma factor has sequences it prefers to bind to, and has a preferred spacing between -35 and -10

Some promoters might have additional elements, e.g. very active ones have an AT rich sequence the ***UP element*** which is contacted by the C-terminal domain of RNA Polymerase α subunit





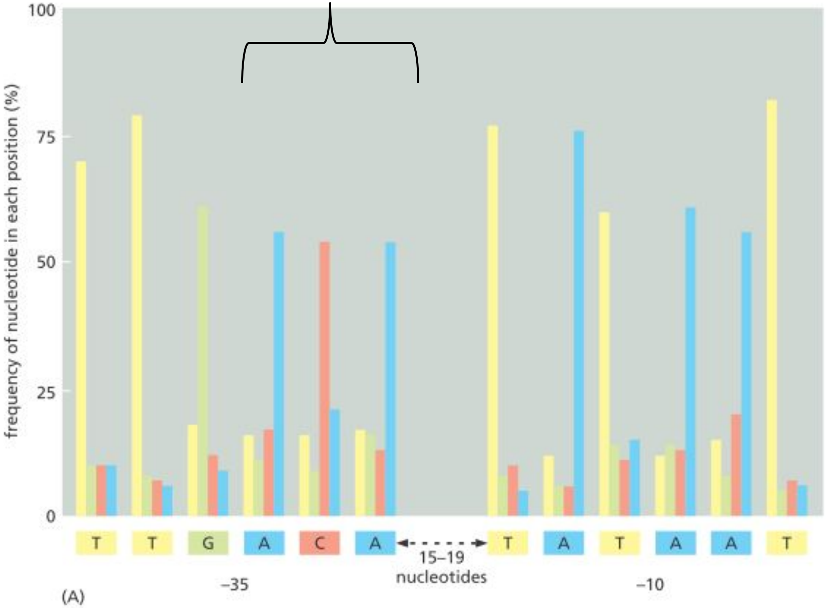


(b) **primary sigma factors**

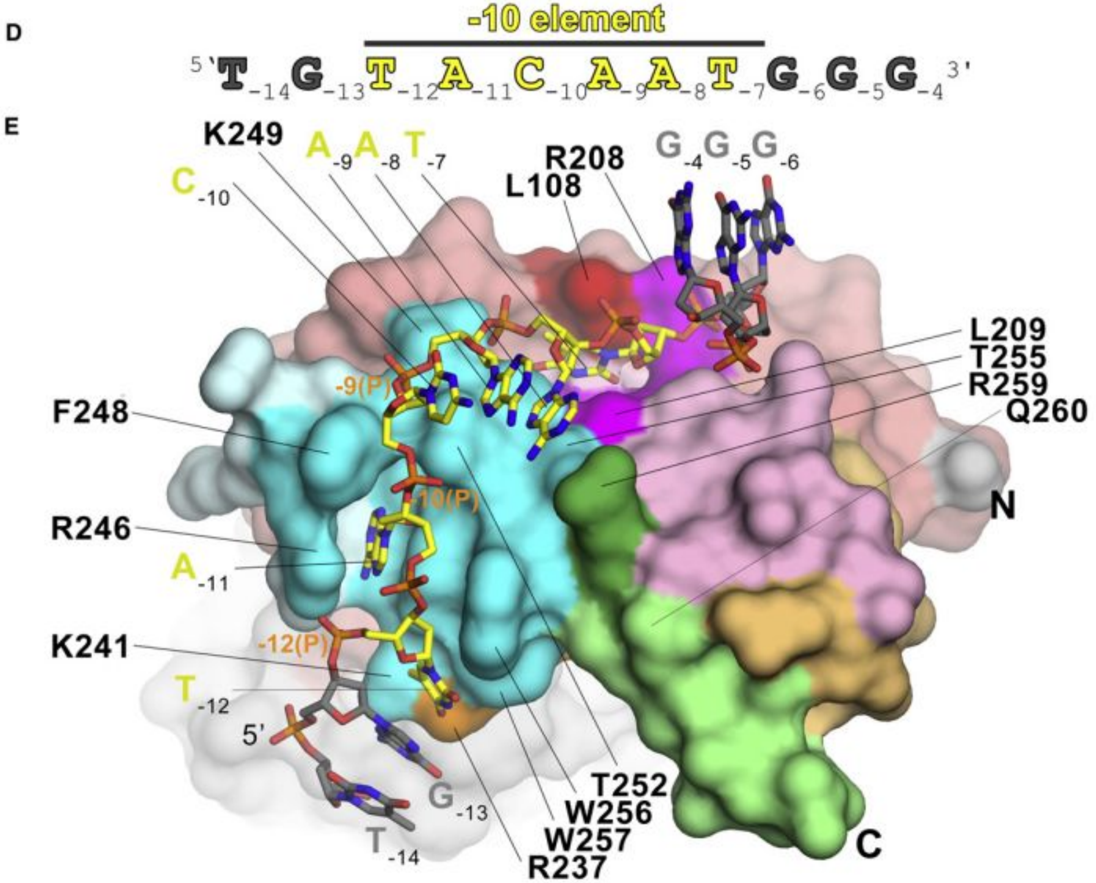
E. coli σ^D (σ^{70})
B. subtilis σ^A

TTGACA 15-20 bp TATAAT

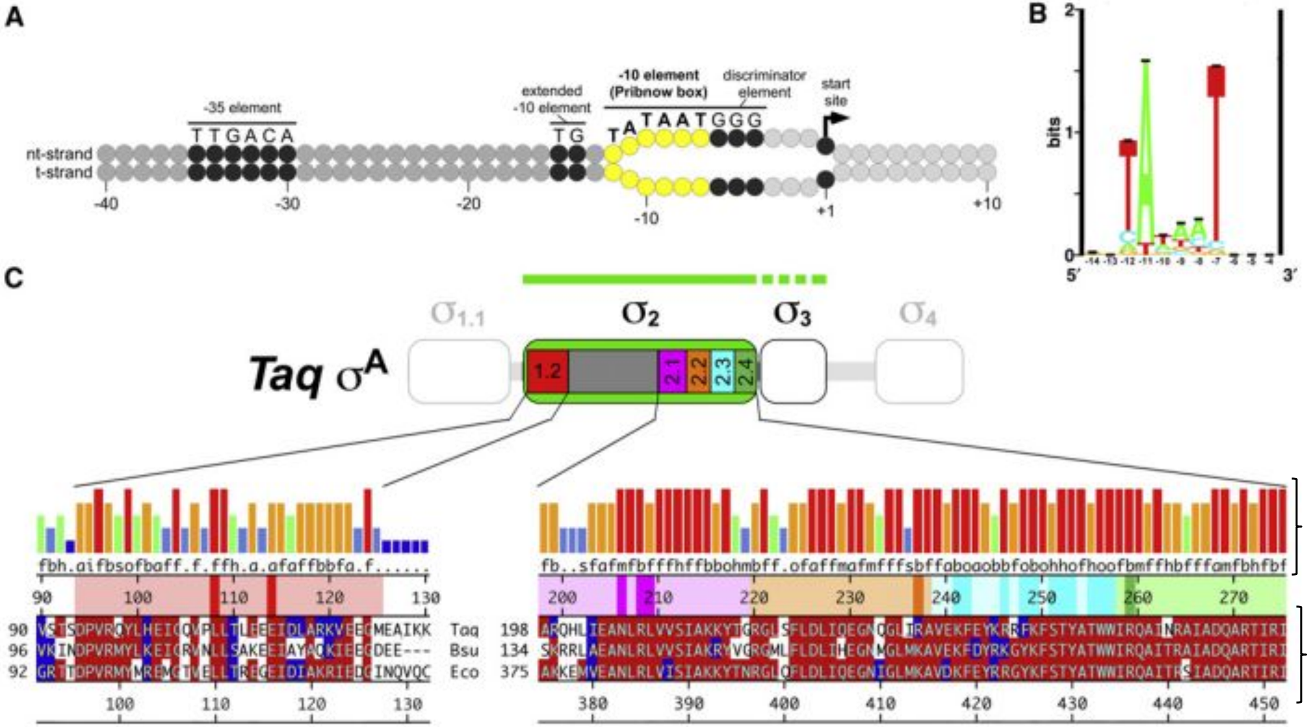
Actually less conserved
ACA only in slightly more than 50% each



Alberts: Molecular Biology of the Cell



Cell 147 , 1257–1269



Sequence conservation

Sequence alignment

Promoter motifs recognized by primary bacterial RNAP factors. Circles are bases and black and yellow circles are recognized.

The resulting sequence logo (another better? Way to sequence frequency histograms)

Cell 147 , 1257–1269

```
gatgaca
gatcacc
gatgaag
gatgact
gatgaca
```

```
g   a   t   g   a   c   a
g   a   t   c   a   c   c
g   a   t   g   a   a   g
g   a   t   g   a   c   t
g   a   t   g   a   c   a
```

Count Matrix

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

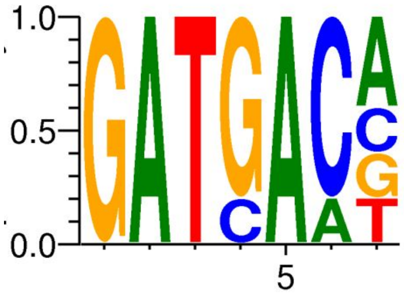
```
g   a   t   g   a   c   a
g   a   t   c   a   c   c
g   a   t   g   a   a   g
g   a   t   g   a   c   t
g   a   t   g   a   c   a
```

Count Matrix

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

Frequency Matrix

	1	2	3	4	5	6	7
A	0	1	0	0	1	0.2	0.4
T	0	0	1	0	0	0	0.2
G	1	0	0	0.8	0	0	0.2
C	0	0	0	0.2	0	0.8	0.2



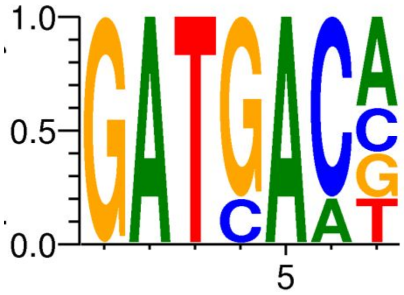
```
g   a   t   g   a   c   a
g   a   t   c   a   c   c
g   a   t   g   a   a   g
g   a   t   g   a   c   t
g   a   t   g   a   c   a
```

Count Matrix

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

Frequency Matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**
A	0	1	0	0	1	0.2	0.4
T	0	0	1	0	0	0	0.2
G	1	0	0	0.8	0	0	0.2
C	0	0	0	0.2	0	0.8	0.2

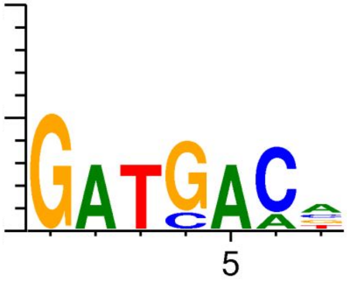


log-odds ratio= $\log_2\left(\frac{\text{Nucleotide Frequency}}{\text{Background Frequency}}\right)$

IC=∑(Nucleotide Frequency×log-odds ratio)

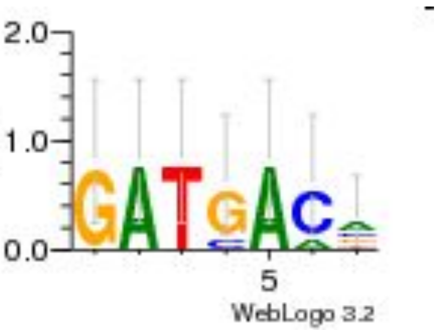
Information Content based Weight matrix $i_{ij}=f_{ij}\log_2 f_{ij}/p_i$

	1	**2**	**3**	**4**	**5**	**6**	**7**
A	NA	1.51	NA	NA	1.51	-0.16	0.08
T	NA	NA	1.51	NA	NA	NA	-0.16
G	2.73	NA	NA	1.93	NA	NA	0.08
C	NA	NA	NA	0.08	NA	1.93	0.08

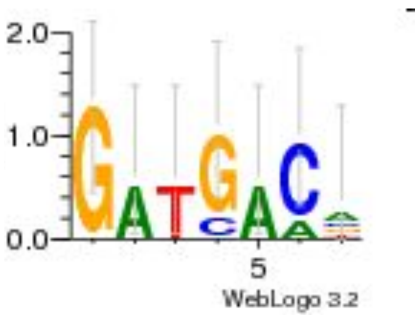


As the GC content of the organism influences the probability of individual bases and thus the information content, a different GC percentage is reflected in grossly different figures.

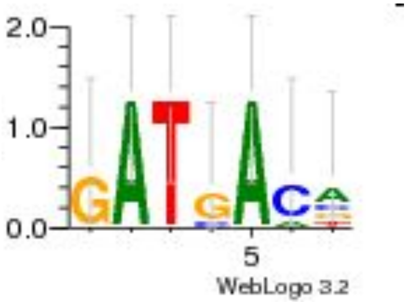
Auto



20% GC

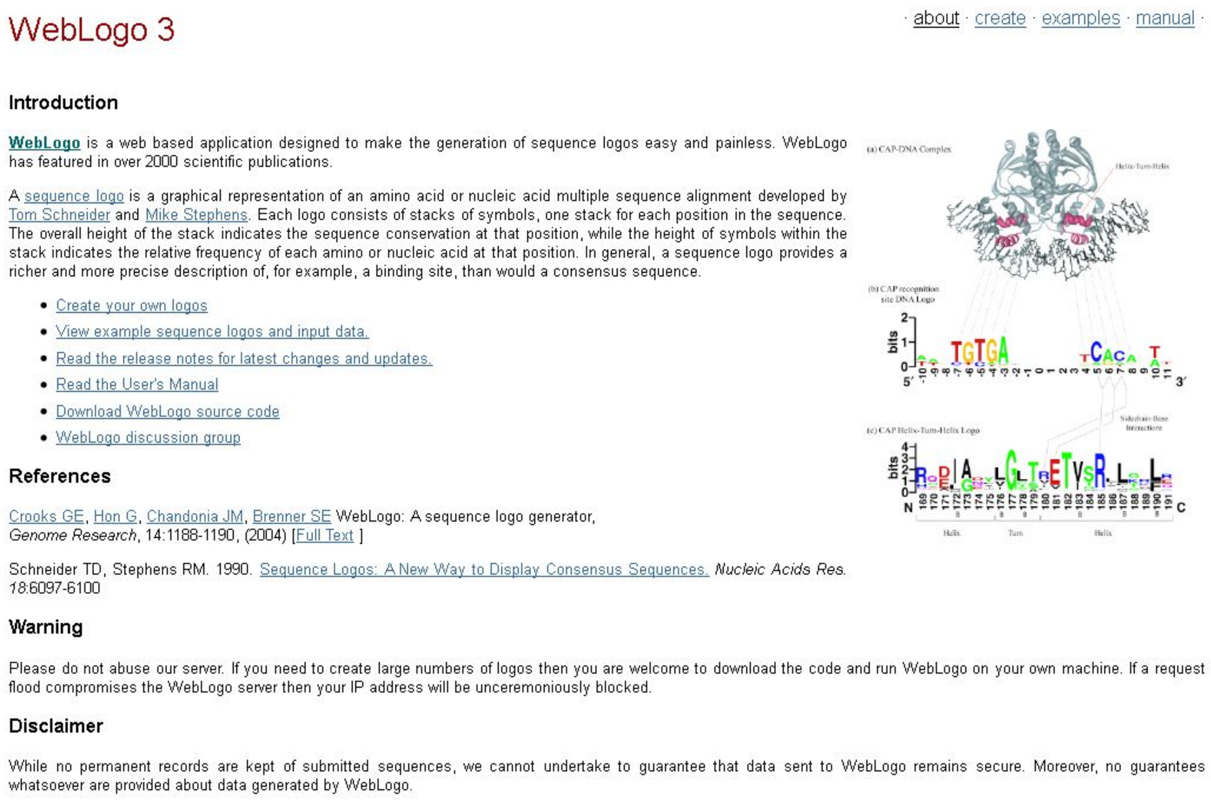


80% GC



You can generate your own sequence logos online

http://weblogo.threeplusone.com/



WebLogo 3

Introduction

WebLogo is a web based application designed to make the generation of sequence logos easy and painless. WebLogo has featured in over 2000 scientific publications.

A sequence logo is a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

- Create your own logos
- View example sequence logos and input data.
- Read the release notes for latest changes and updates.
- Read the User's Manual
- Download WebLogo source code
- WebLogo discussion group

References

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator,
Genome Research, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* *18*:6097-6100

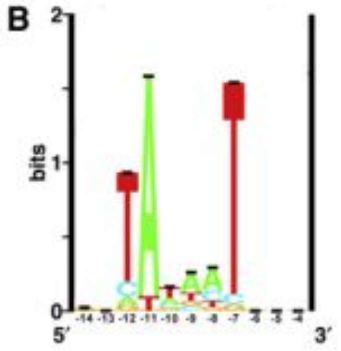
Warning

Please do not abuse our server. If you need to create large numbers of logos then you are welcome to download the code and run WebLogo on your own machine. If a request flood compromises the WebLogo server then your IP address will be unceremoniously blocked.

Disclaimer

While no permanent records are kept of submitted sequences, we cannot undertake to guarantee that data sent to WebLogo remains secure. Moreover, no guarantees whatsoever are provided about data generated by WebLogo.

Summary so far



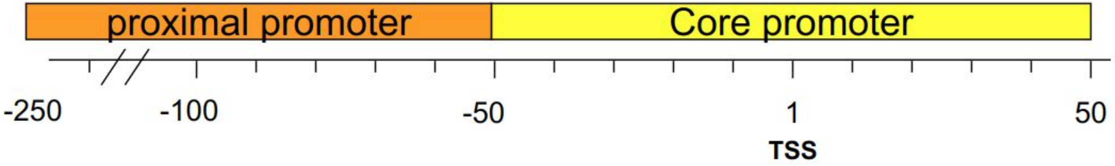
- Sigma Factors bind to certain sequences
- These ones are not absolute
- But one can display them using sequence logos
- These can be scaled using bits or using frequency (the latte being much simpler but less meaningful)

Eukaryotes and transcriptional start points

Eukaryotic RNA polymerases need the TATA binding protein (TBP) to initiate transcription (this is part of TFIID) TBP binds to the TATA box if this is present (~25-30 bp upstream of TS start site)

It was presumed that not all genes have a TATA box but only about 1/3 of genes.

- The “core-promoter” is a sequence region of nearly 100 bp surrounding the transcription start site (TSS).
- This core promoter might extend from about 50 bp upstream of the TSS to 50 bp downstream of the TSS in eukaryotic organisms.
- The core promoter alone is enough to drive basal transcription by nucleating the assembly of the pre-initiation complex, consisting of RNA Polymerase II and associated general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH).



- Eukaryotic polymerases need the **TATA binding protein (TBP)** to initiate transcription (this is part of TFIID)
- Promoters for Pol II often have a **TATA box** (TATAA consensus sequence) ~25-30bp upstream of the transcription start
- Some other elements that may be present include the **TFIIB** recognition element (**BRE**), initiator element (**INR**) and downstream promoter element (**DPE**; found downstream of the transcription start)
- There are however many promoters that do not have any of these elements

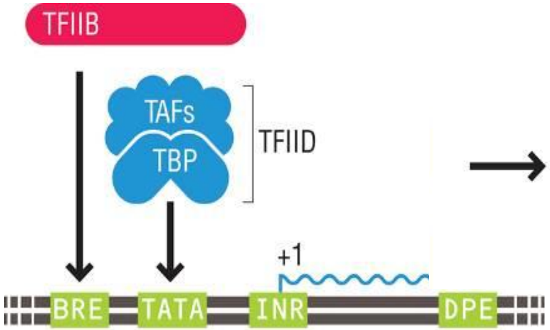
- The first step in assembling one transcription initiation complex is often binding of TFIID to the TATA box

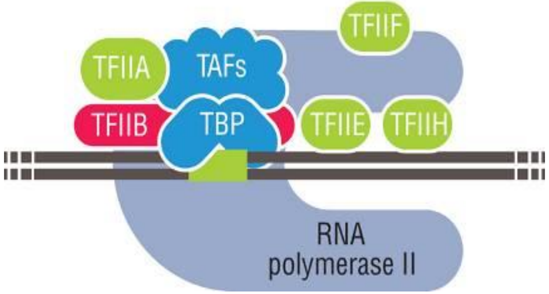
- TFIID binds to the TATA box via TBP, which binds to the minor groove of DNA, inducing strong distortions in the DNA and thus local DNA unwinding

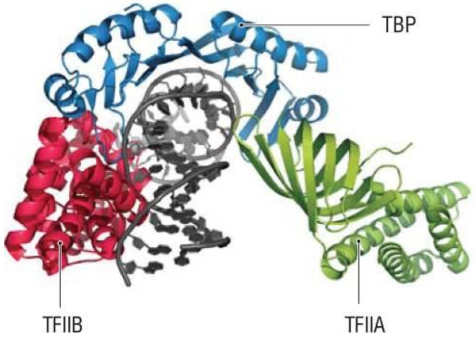
- Other components of TFIID, called TBP-associated factors (TAFs), mediate recognition of other promoter elements like INR and DPE

- After TFIID has associated with DNA, TFIIB is recruited. This recognizes the BRE promoter element and binds asymmetrically, helping to determine the transcription direction. TFIIB has some similarities to bacterial sigma factor

- After TFIID and TFIIB have bound, TFIIA binds, and stabilizes the TBP-DNA interactions, then TFIIE and TFIIH (TFIIH catalyzes ATP-powered DNA unwinding)

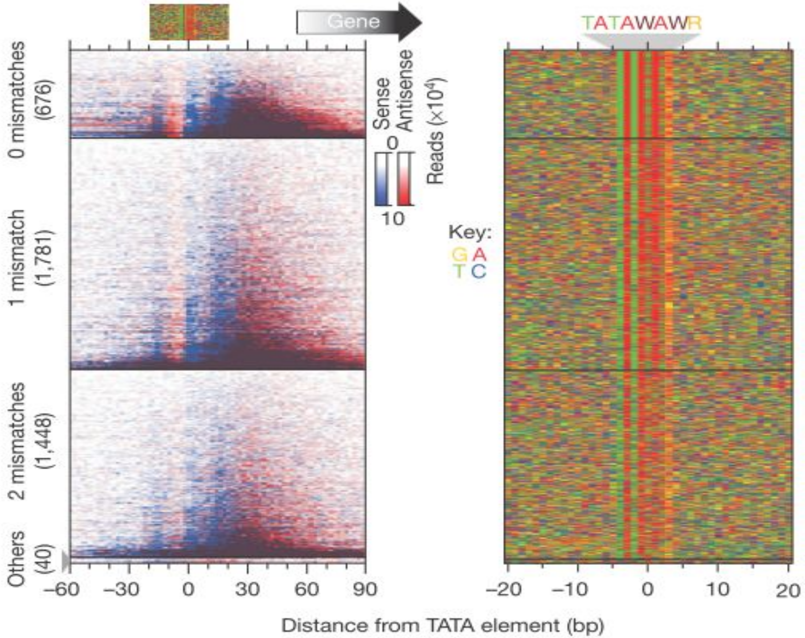






eukaryotic RNA polymerases need the TATA binding protein (TBP) to initiate transcription (this is part of TFIID) TBP binds to the TATA box if this is present (~25-30 bp upstream of TS start site)

It was presumed that not all genes have a TATA box but only about 1/3 of genes. This might be challenged by some novel experiments showing a TATA like element.

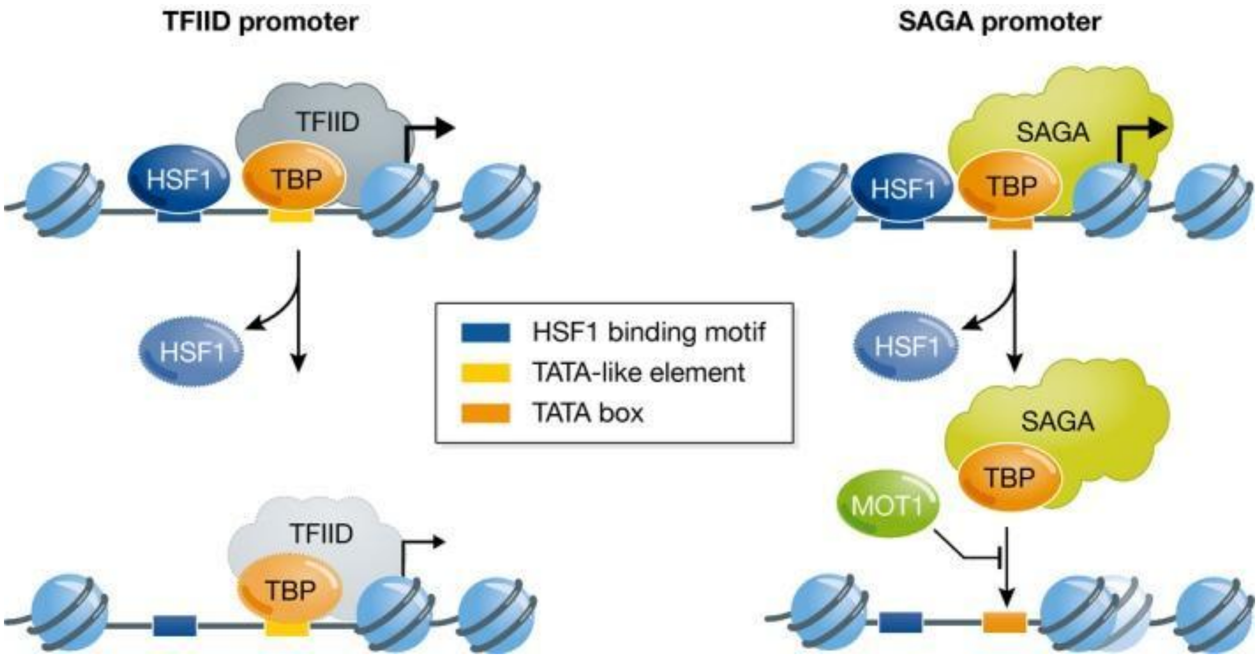


Finding Transcriptional start sites, mapping reads
Right DNA centered around putative TATA Elements

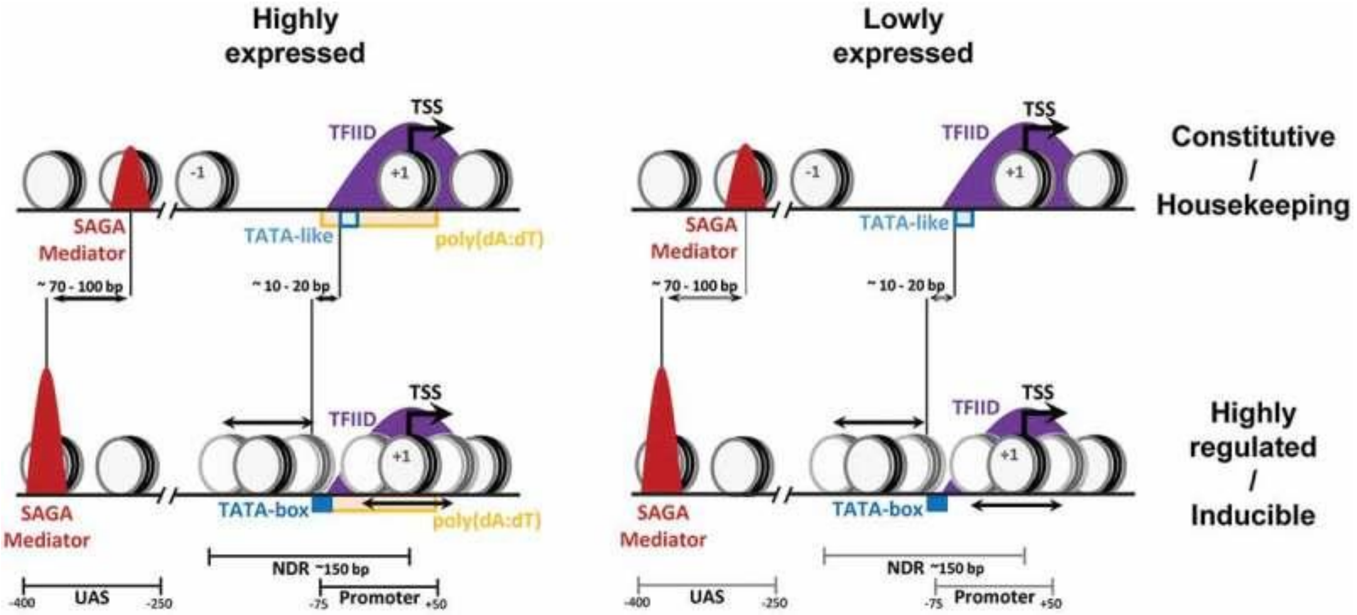
Nature 483,295–301

SAGA-dominated promoters often contain strong TATA- box motifs and are associated with genes responsive to stress

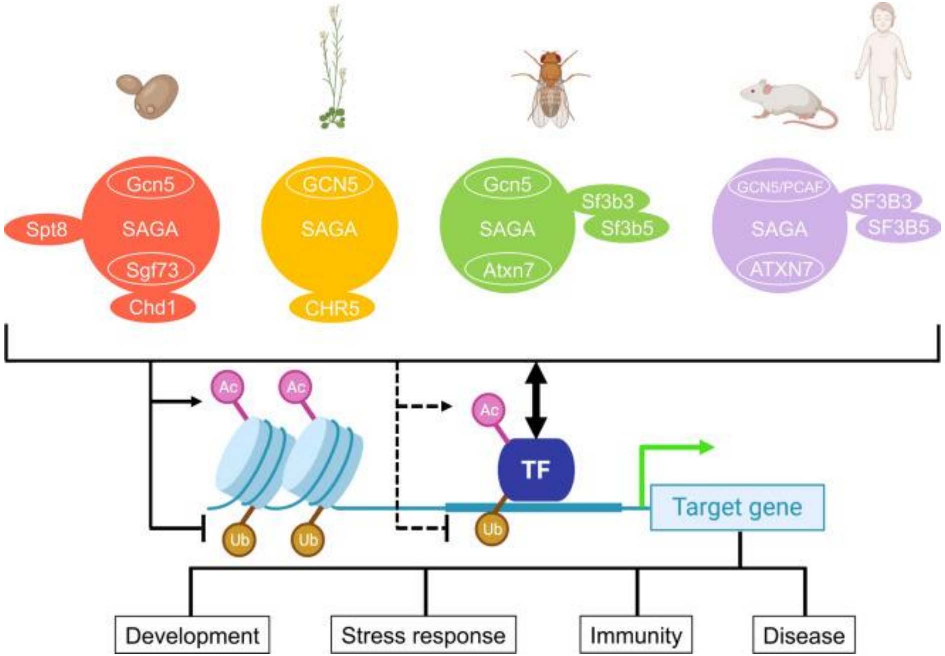
TFIID- dominated promoters are depleted of such strong TATA-box motifs



Kubik et all EMBO J. 2017 Feb 1; 36(3): 248–249.



Fischer V (2019) Global role for coactivator complexes in RNA polymerase II transcription. Transcription. 2019 Feb;10(1):29-36



Chen & Dent (2021) Conservation and diversity of the eukaryotic SAGA coactivator complex across kingdoms. Epigenetics Chromatin. 14(1):26

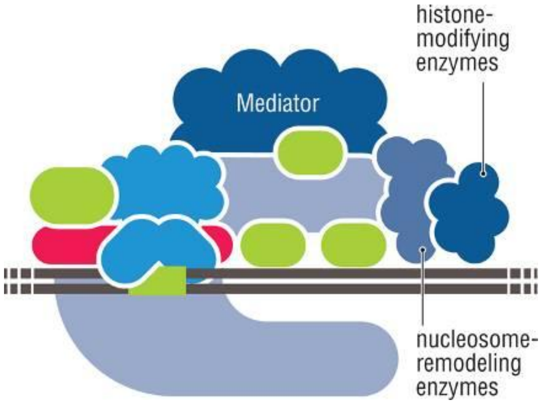
- The pre-initiation complex components are:

  - RNA polymerase II core enzyme (light blue)
  - General transcription factors (blue, green, red)

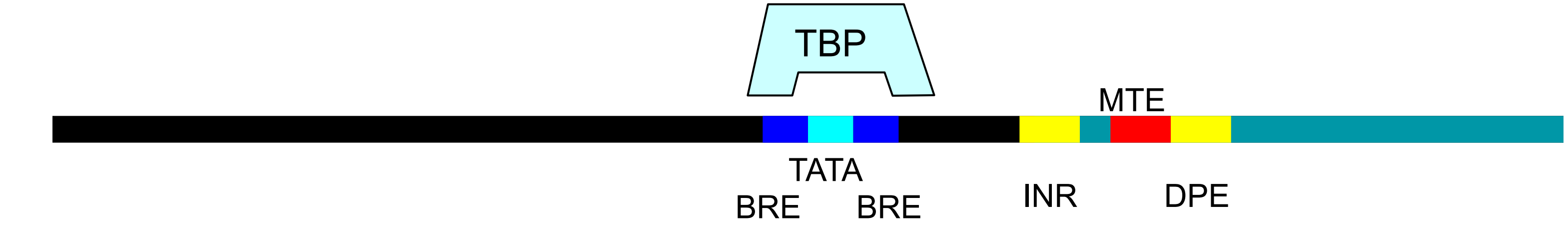
- The pre-initiation complex is competent to initiate transcription *in vitro*. *In vivo* transcription requires additional protein complexes

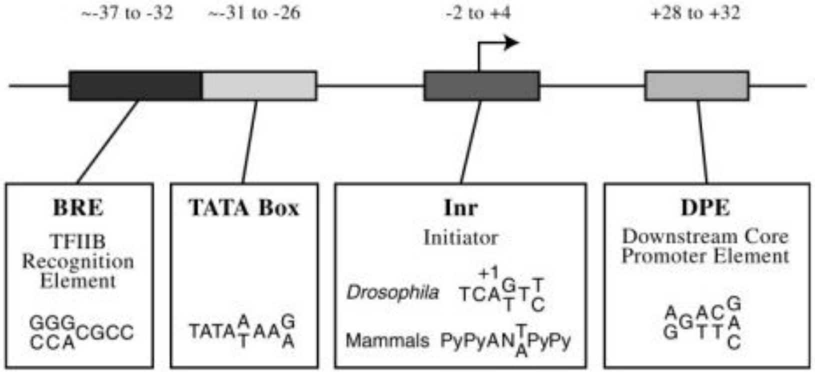
- These include enzymes that alter chromatin structure (to remove histone barriers to transcription)

- Another large complex, Mediator (which has more than 20 subunits), is needed to activate many Pol II transcribed genes



Certain conserved elements can be found in the core promoter. The most notable ones are TATA, BRE, INR, DPE




~-37 to -32
~-31 to -26
-2 to +4
+28 to +32
BRE
TFIIB Recognition Element
$\substack{\text{GGG}\\\text{CCA}}$CGCC
TATA Box
TATA$\substack{A\\T}$AA$\substack{G\\A}$
Inr
Initiator
+1
Drosophila TCA$\substack{G\\T}$T$\substack{T\\C}$
Mammals PyPyAN$\substack{T\\A}$PyPy
DPE
Downstream Core Promoter Element
$\substack{A\\G}$G$\substack{A\\T}$$\substack{C\\T}$$\substack{G\\A\\C}$

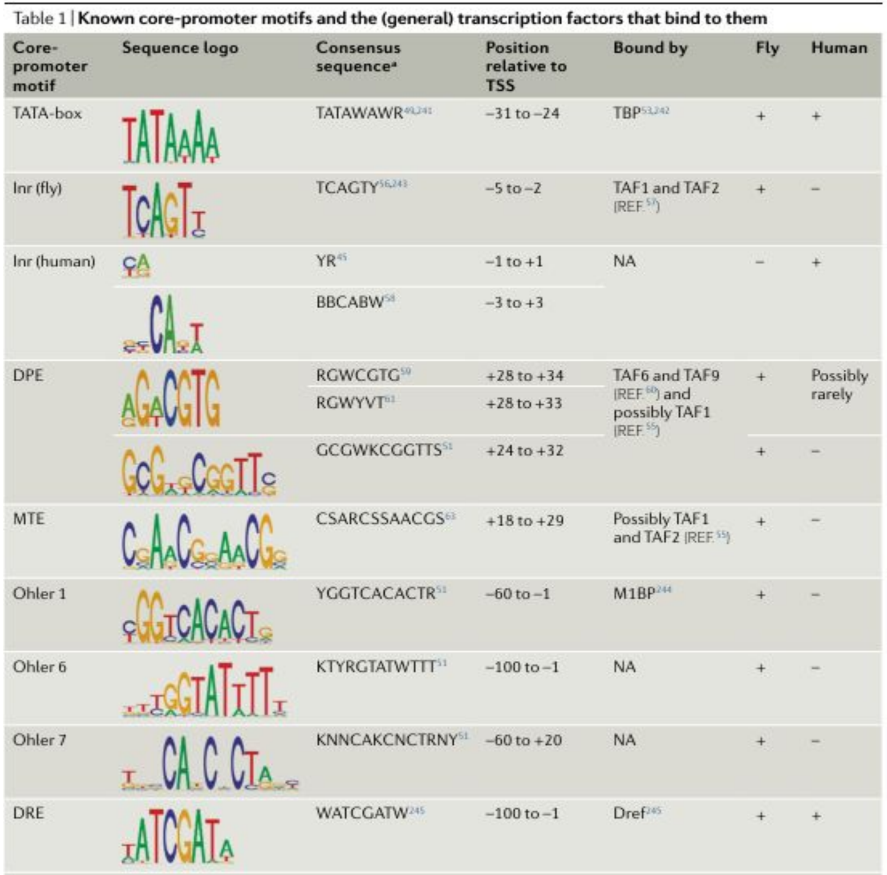
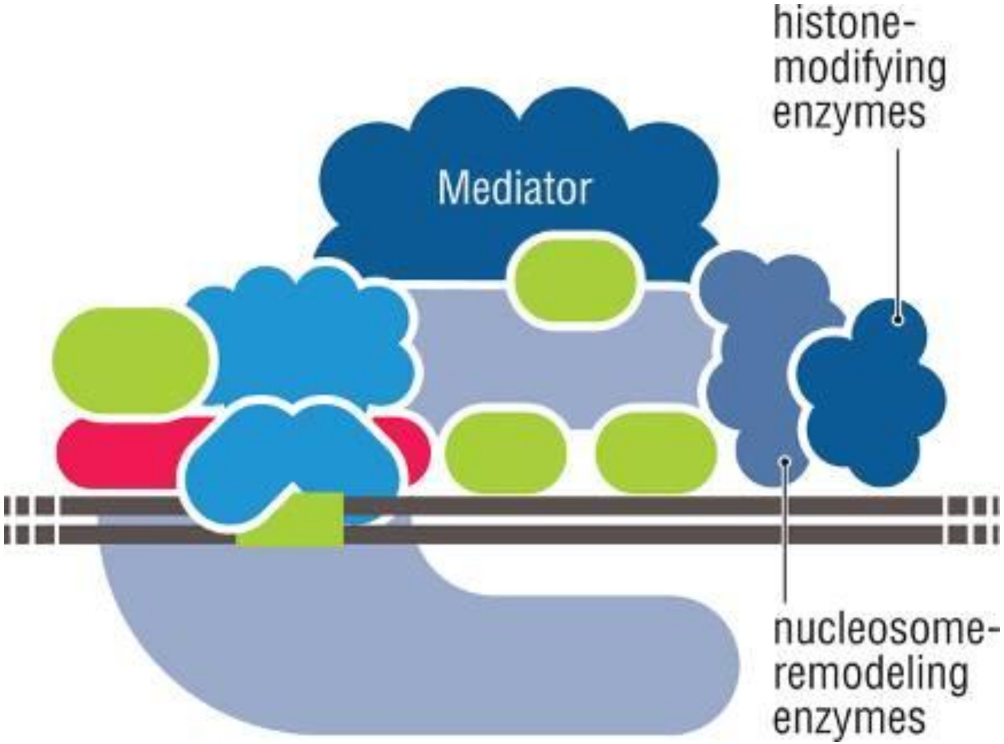


Table 1 | **Known core-promoter motifs and the (general) transcription factors that bind to them**

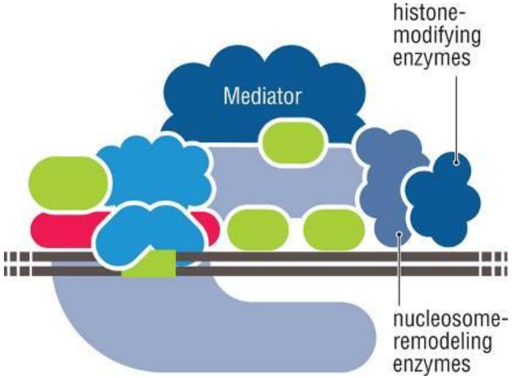
Core-promoter motif	Sequence logo	Consensus sequence[a]	Position relative to TSS	Bound by	Fly	Human
TATA-box		TATAWAWR[49,241]	−31 to −24	TBP[53,242]	+	+
Inr (fly)		TCAGTY[56,243]	−5 to −2	TAF1 and TAF2 (REF.[57])	+	–
Inr (human)		YR[45]	−1 to +1	NA	–	+
		BBCABW[58]	−3 to +3			
DPE		RGWCGTG[59]	+28 to +34	TAF6 and TAF9 (REF.[60]) and possibly TAF1 (REF.[55])	+	Possibly rarely
		RGWYVT[61]	+28 to +33			
		GCGWKCGGTTS[51]	+24 to +32		+	–
MTE		CSARCSSAACGS[63]	+18 to +29	Possibly TAF1 and TAF2 (REF.[55])	+	–
Ohler 1		YGGTCACACTR[51]	−60 to −1	M1BP[244]	+	–
Ohler 6		KTYRGTATWTTT[51]	−100 to −1	NA	+	–
Ohler 7		KNNCAKCNCTRNY[51]	−60 to +20	NA	+	–
DRE		WATCGATW[245]	−100 to −1	Dref[245]	+	+

Haberle and Stark (2018) Nature Reviews Molecular Cell Biology19: 621–637


histone-modifying enzymes
Mediator
nucleosome-remodeling enzymes

Summary Eukaryotes

Promoter is complex and different transcription starts can be found



Of cis and trans

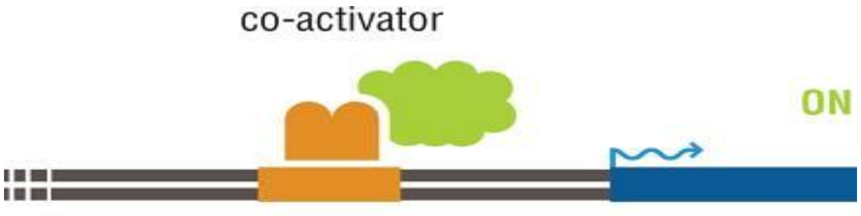
A locus is ***cis*-acting** on a second locus if it must be on the same DNA molecule in order to have an effect. The operator is a *cis*-acting element because it works only when physically attached to the gene whose expression it regulates.

A locus is ***trans*-acting** if it can affect a second locus even when on a different DNA molecule. The gene for the lactose repressor (*lacI*) is *trans*-acting because it can regulate expression of the lactose operon even when removed from the *Escherichia coli* chromosome and placed on a plasmid.

To a molecular biologist and bioinformatician, a *cis*-acting regulatory element is usually a target site for a DNA-binding protein, upstream of the gene whose expression is being regulated. A *trans*-acting element is the regulatory protein itself, which can diffuse through the cell from its site of synthesis to its DNA-binding site.

Regulatory proteins must specifically recognize the right regulatory sequence

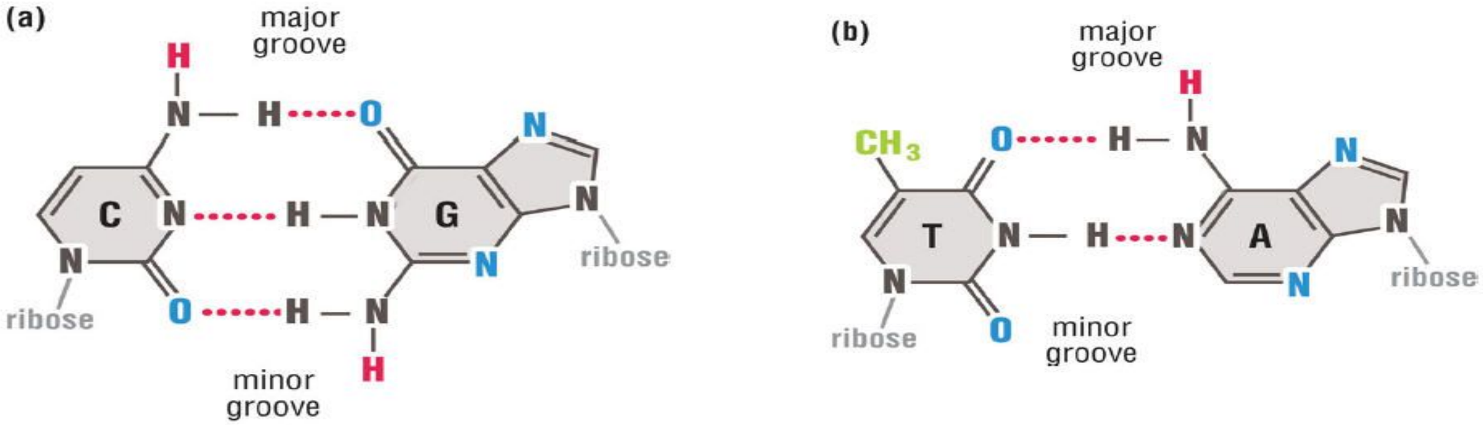
Each regulator usually has a DNA binding domain that recognizes a specific sequence and additional domains



Regulators must be able to recognize certain DNA sequences with high specificity and bind to them non-covalently this is done through contacts

Each DNA base has different available chemical groups – hydrogen bond donors and acceptors (and a methyl group on thymine)

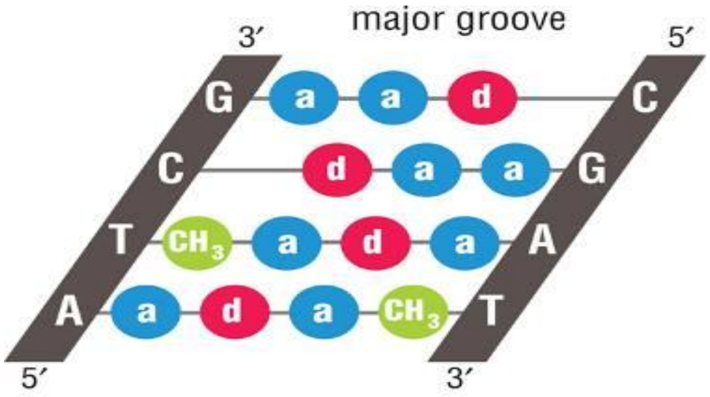
The surface of the protein is adapted to the DNA surface. Often positively charged amino acids are used.

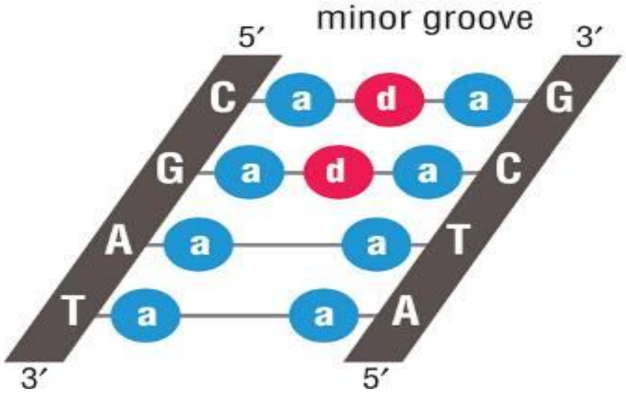


Craig Molecular Biology

There is **less variability** in the minor groove, because T-A and A-T are the same, and G-C and C-G are the same, so these can't be distinguished by binding proteins

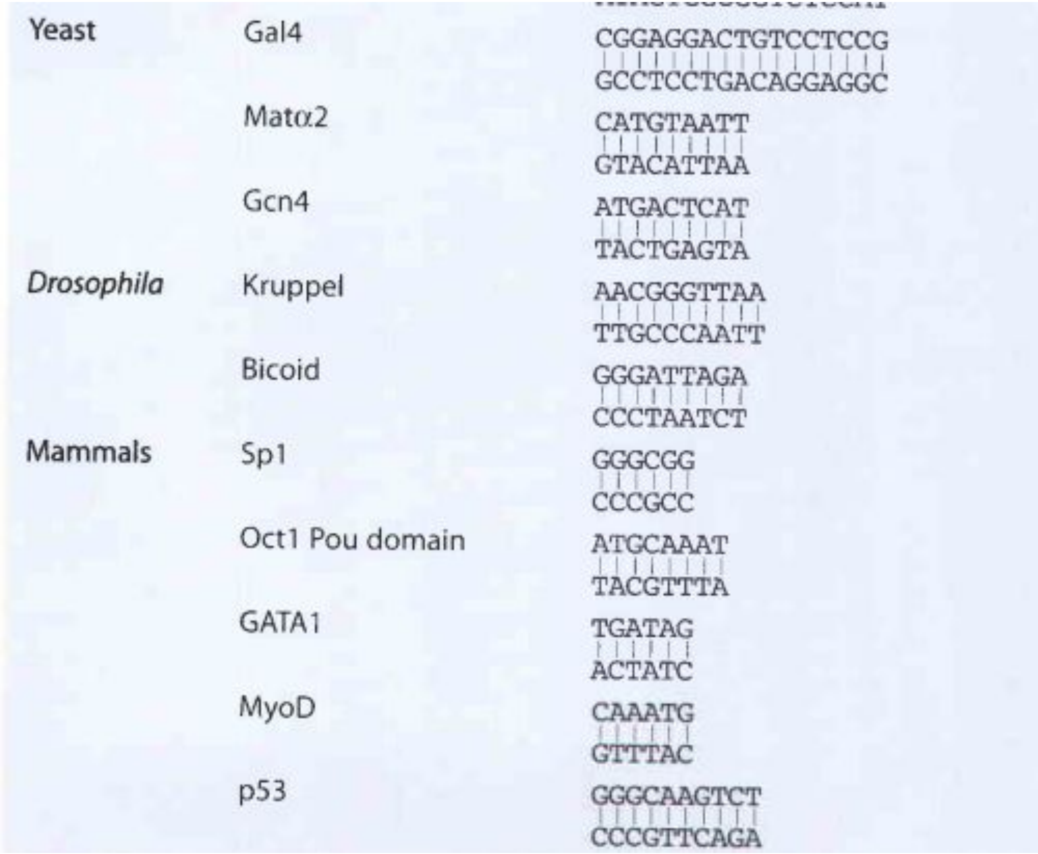
Many regulatory proteins interact primarily through the major groove, via non-covalent interactions with the available groups but not all!



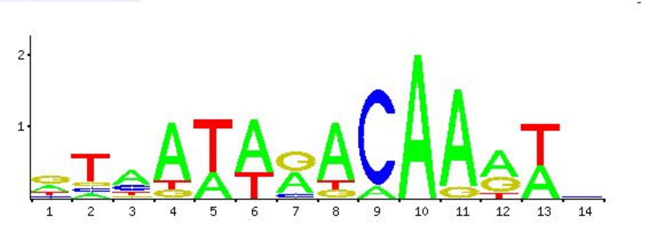


Craig Molecular Biology

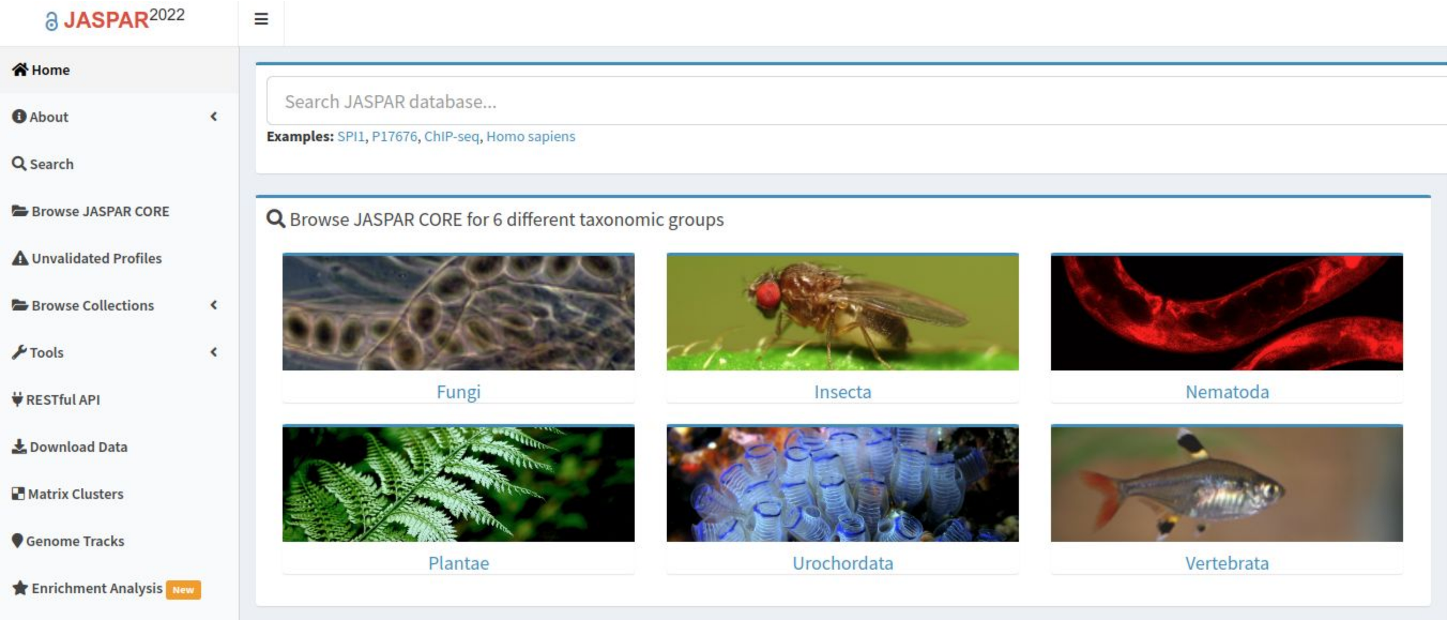
Typically binding sites are short. Also usually these are not very well conserved



Yeast	Gal4	CGGAGGACTGTCCTCCG GCCTCCTGACAGGAGGC
	Matα2	CATGTAATT GTACATTAA
	Gcn4	ATGACTCAT TACTGAGTA
Drosophila	Kruppel	AACGGGTTAA TTGCCCAATT
	Bicoid	GGGATTAGA CCCTAATCT
Mammals	Sp1	GGGCGG CCCGCC
	Oct1 Pou domain	ATGCAAAT TACGTTTA
	GATA1	TGATAG ACTATC
	MyoD	CAAATG GTTTAC
	p53	GGGCAAGTCT CCCGTTCAGA



Alberts: Molecular Biology of the Cell



Examples: SPI1, P17676, ChIP-seq, Homo sapiens

Browse JASPAR CORE for 6 different taxonomic groups