



H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics online course: IBT

Bioinformatics resources and databases: Lecture 3: DNA sequence analysis

Nicola Mulder



H3ABioNet

Pan African Bioinformatics Network for H3Africa



Introduction to Bioinformatics online
course: IBT
Bioinformatics Resources & Databases: N
Mulder

Learning Objectives

- Objective: Basic DNA sequence analysis – finding sequence features
- Sub objectives:
 - Understand how to extract a DNA sequence from the database
 - Use online or local tools for simple DNA sequence analysis -finding features on the sequence and their applications

Learning Outcomes

- Understand how to find a DNA sequence and save it in the correct format
- Identify features on the sequence such as coding regions, restriction enzyme sites, etc.
- Design primers for amplification of a DNA sequence
- Interpret sequence analysis results and understand the biological impact of functional regions

Fundamental biology you need to know

- DNA is made up of 4 nucleotides/bases: ACGT
- DNA is usually double stranded through base pairing: A-T and C-G
- G-C bond is a bit stronger than A-T bond
- Double strand has top strand (sense strand) running 5' to 3'
- Bottom strand is complementary

5'
A C C T T G C G T T A A G C A A T T G G C T
3'
T G G A A C G C A A T T C G T T A A C C G A
G 5'

Reverse complement:
bottom strand read 5' to 3'

Genes can be encoded by either strand, + or – and are transcribed into RNA

How do you get the sequence?

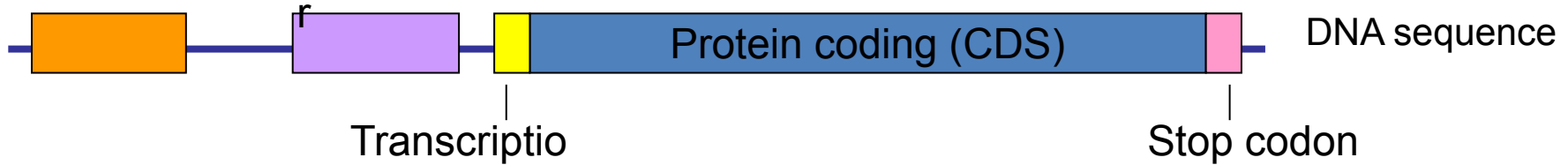
- Sequencing technologies to determine the sequence of bases
- Not in the scope of this module

Two major components to sequence analysis

- Storing and retrieving data:
 - Biological databases
 - Querying these to retrieve data
- Manipulating the data –tools e.g:
 - Finding features on sequences
 - Sequence similarity searches
 - Protein families and function prediction
 - Comparing sequences –phylogenetics
 - Etc.

Aspects of sequence analysis

Regulatory region Promote

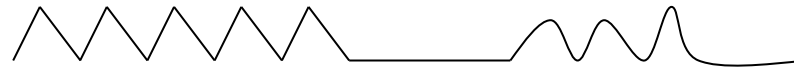


Gene and promoter prediction

Transcription start

RNA sequence

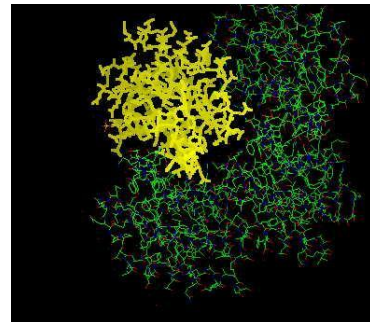
RNA secondary structure, gene expression



Protein sequence

Protein sequence analysis

Restriction mapping for cloning, primer design for PCR



Sequence formats:

Fasta

> [title]
[sequence]

>seq1

GGAAAATTAGATGCATGGGAAAAAATTA
GGAAAATTAGACAAATGGGAAAAAATTA

>seq2

AAGTCCCTGGATTACCCAATGCAGTCG
A
CATCGCATTT

Sequence formats: GenBank

```
LOCUS      525-42      1588 bp
DEFINITION 525-42      1588 bp
TITLE      525-42
FEATURES             Location/Qualifiers
     exon           39..70
                /note="exon1 is believed to have an alternative splice donor site"
ORIGIN
```

```
1      ATGTT AAGAG GGGGA AAATT AGATG CATGG GAAAA AATTA GGTTA AGGCC
51     AGGGG GAAAG AAATG CTATA NGATA AAACA CCTAG TATGG GCAAG CAGGG
101    AGCTG GAAAG ATTTG CACTT AACCC TGGCC TTTTA GAGAC ATCAG ANGGC
151    TGTA  ACAA  TAATG NAACA GATAC AACCA GCTCT TCAGA CAGGA ACAGA
```

Converting between sequence formats (save options)

DNA sequence composition

- Nucleotide composition (% GC vs AT content)
- GC bonds are stronger than AT bonds
- Applications:
 - Horizontal gene transfer analysis
 - Gene prediction
 - Primer design

Accession numbers

- **GenBank/EMBL/DDBJ**: 1 letter & digits, e.g.:
U12345 or 2 letters & 6 digits, e.g.:
AY123456
- **GenPept** Sequence Records -3 letters & 5 digits, e.g.: AAA12345
- **UniProt** -All 6 characters: [A,B,O,P,Q] [0-9] [A-Z,0-9] [A-Z,0-9] [A-Z,0-9] [0-9], e.g.:
P12345 and Q9JJS7

Cross-referencing identifiers

- So many different IDs for same thing, e.g. Ensembl, EMBL, HGNC, UniGene, UniProt, Affy ID, etc.
- Need mapping files to move between them to avoid having to parse every entry
- UniProt website mapper (www.uniprot.org)
- PICR (<http://www.ebi.ac.uk/Tools/picr/>) enables mapping between IDs

Example conversion

PICR Protein Identifier Cross-Reference

[Home](#)
[User Guide](#)
[Implementation](#)
[Webservice](#)
[RESTful](#)
[Contact Us](#)

Input Accession	ENSEMBL	REFSEQ	UNIPROT_BEST_GUESS
O60260	ENSP00000343589 ENSP00000355860 ENSP00000355862 ENSP00000355863 ENSP00000355865 ENSP00000434414 ENSG00000185345 ENSP00000343589 ENSP00000355860 ENSP00000355862 ENSP00000355863 ENSP00000434414 ENST00000338468 ENST00000366894 ENST00000366896 ENST00000366897 ENST00000366898 ENST00000479615	NP_004553.2 NP_054642.2 NP_054643.2 XP_014201025.1 NM_004562.2 NM_013987.2 NM_013988.2	O60260.2
P01130	ENSP00000397829 ENSP00000437639 ENSP00000440520 ENSP00000453346	NP_000518.1 NP_001182727.1 NP_001182728.1 NP_001182729.1	P01130.1

DNA sequence analysis

- Restriction analysis e.g. for cloning –looks for recognition sites
- Primer design
- Finding features on a sequence
- Gene prediction:
 - Translation
 - Promoter prediction

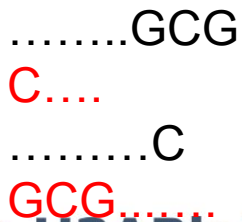
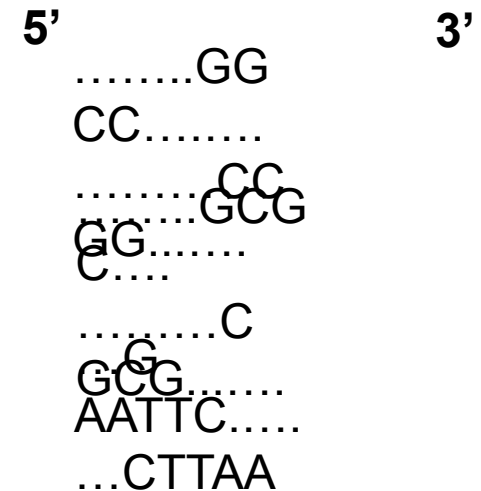
Bioinformatics and cloning

- Retrieving sequence of interest
- Identifying restriction enzyme sites
- Matching these to RE sites in cloning vector

Restriction enzyme analysis

- Restriction enzymes recognize specific or defined 4 to 8 base pair sequences on DNA and cut

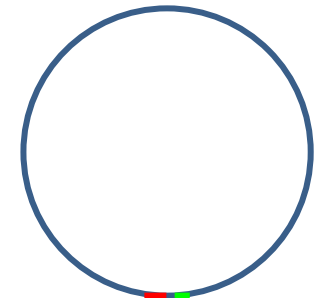
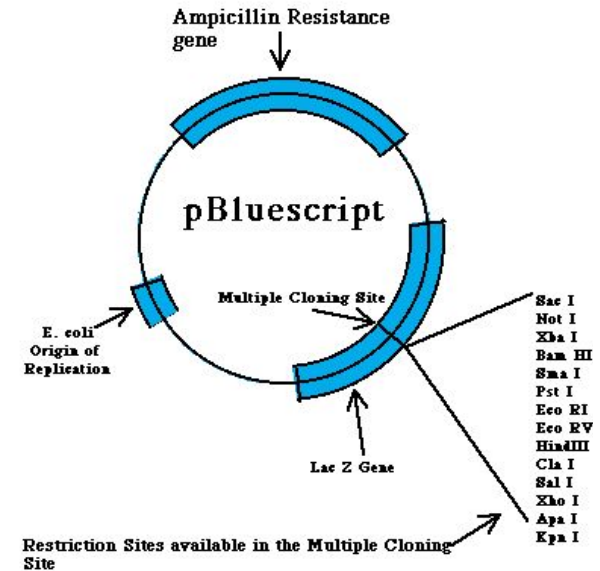
Microorganism	Enzyme	Sequences	Notes
<i>Haemophilus aegitius</i>	<i>HaeIII</i>	5'...GG CC..3' 3'...CC GG..5'	Blunt end
<i>Haemophilus haemolytica</i>	<i>HhaI</i>	5'...GC G C..3' 3'...CG C G..5'	3' single strand
<i>Escherichia coli</i>	<i>EcoRI</i>	5'...G AATT C..3' 3'...C TTAA G..5'	5' single strand



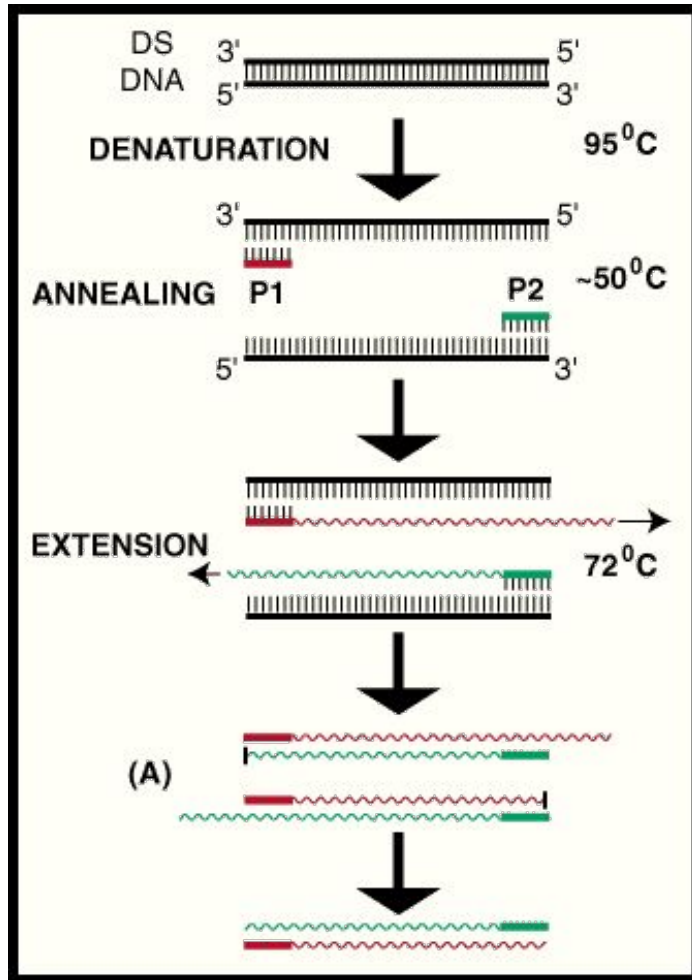
Restriction map

Restriction Enzyme Map:

1	TACATGCATGTTTCATGGTAGCATTATTTCACAAAGCCAAAAGATGCAAAACAGCCCCAATGTCCATAGATGAATAAACTGTG	80
1	ATGTACGTACAAGTACCATCGTAATAAGTGTTTTCGGTTTTCTACGTTTGTGCGGGTTACAGGTATCTACTTATTTGACAC	80
	NspI SfaNI	
	BfrBI	
	NsiI	
	NspI	
	MslI	
81	GCATACATGATACACACACACACGACACACATATACATATACACACAAAACTATTTCAGTCATAAAAAGGAATAA	160
81	CGTATGTACTATGTGTGTGTGTGCGTGTGTGTATATGTATGTGTGTGTTTGTGATAAGTCAGTATTTTCCITTATT	160
	TspDTI	
161	AGTCTGTTACATGCTACCTGAGGATGAACCTCGAAAAATGCTAAGTGAAAGACACAAAAAGTCCACACTGTGATTCCG	240
161	TCAGACAATGTACGATGGACTCCTACTTGGAGCTTTTGTACGATTCACTTTTCTGTGTTTTCAGGTGTGTGACACTAAGGC	240
	BseMII Bsu36I BstF5I TspDTI DrrI TspGWI	
	BspCNI FokI NspI Hpy8I DraIII	
	NspI MnlI TspRI	
	MnlI	
241	TTTATATGAAGTATCTAAAGTAAGTAAATATAGAGACAGAAGTAGACTGGTAATTGCCAGGGGCTGGGGGGAAGAGGGC	320
241	AAATATACTTCATAGATTTCATTTCATTTATATCTCTGTCTTCATCTGACCATTAACGGTCCCCGACCCCCCTTCTCCCG	320
	TspDTI AccI BsrI BsaJI EarI	
	BsmAI Hpy8I BslI	
	PflMI	
	AlwNI	
	BseYI	
	MnlI	



PCR and primer design



- Primers should be similar length and T_m
- Should amplify only required piece from genome

Example with Primer BLAST

Primer-BLAST

A tool for finding specific primers

BI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST).

[Reset page](#)
[Save search parameters](#)
[Retrieve recent results](#)
[Publication](#)
[Tips for finding specific primers](#)

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) [Clear](#)

Or, upload FASTA file [Browse...](#) No file selected.

Range

Forward primer [Clear](#)

Reverse primer

Primer Parameters

Use my own forward primer (5'→3' on plus strand) [Clear](#)

Use my own reverse primer (5'→3' on minus strand) [Clear](#)

PCR product size

Min Max

of primers to return

Primer melting temperatures (T_m)

Min Opt Max Max T_m difference

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section [?](#)

Exon junction span

Exon junction match

Exon at 5' side Exon at 3' side

Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction [?](#)

Intron inclusion ☐ Primer pair must be separated by at least one intron on the corresponding genomic DNA [?](#)

Intron length range

Min Max

Enter
accession
number of DNA
entry or paste
in sequence
Select where
on sequence
primers should
be

Select size
range of
product and
T_m

Primer Pair Specificity Checking Parameters

Pan African Bioinformatics Network for H3Africa



Primer BLAST output

Detailed primer reports

Primer pair 1

	Sequence (5'→3')	Length	Tm	GC%	Self complementarity
Forward primer	ATGAGGCCAAGGACCCAAGAC	21	62.08	57.14	4.00
Reverse primer	GATGAGGGGCTGACAGGAGTGG	22	64.35	63.64	5.00

Products on target templates

>[NC_000020.11](#) Homo sapiens chromosome 20, GRCh38.p7 Primary Assembly

product length = 690

Features associated with this product:

[glutathione synthetase](#)

[glutathione synthetase](#)

```
Forward primer 1      ATGAGGCCAAGGACCCAAGAC  21
Template       34929242 ..... 34929222

Reverse primer 1      GATGAGGGGCTGACAGGAGTGG  22
Template       34928553 ..... 34928574
```

product length = 1995

Features flanking this product:

[62286 bp at 5' side: zinc finger protein 217](#)

[297816 bp at 3' side: breast carcinoma-amplified sequence 1 isoform 1](#)

```
Forward primer 1      ATGAGGCCAAGGACCCAAGAC  21
Template       53645111 C.....A.TT..... 53645131

Forward primer 1      ATGAGGCCAAGGACCCAAGAC  21
Template       53647105 C.....A.T.A.....G 53647085
```

>[NC_018931.2](#) Homo sapiens chromosome 20, alternate assembly CHM1_1.1, whole genome shotgun sequence

product length = 690

Features associated with this product:

[glutathione synthetase](#)

[glutathione synthetase](#)

Other places in
the genome
primers may
bind

Gene Prediction

Wikipedia: A **gene** is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions

- Look for gene structures
- Move along sequence looking for coding regions and intergenic regions
- Check reading frame -translate
- Look for promoters and poly-adenylation signals
- In eukaryotes look for introns and exons
- Use EST or BLAST support (reduce pseudogenes)

Translation

- Can choose frame if you know it
- Otherwise 6-frame translation:
 - Choose start codon ATG
 - Otherwise lists all codons between stop codons
- Results –for bacteria, usually the longest ORF starting with Met and ending in stop, & no stop codons inside
- Can confirm this with promoter prediction
- Should use appropriate **codon usage table**

Open reading frame

- String of in-frame combinations/triplets of bases that specify an amino acid
- Starts with ATG (Meth) or Val
- Ends with stop codon
- One base insertion or deletion –out of frame/frameshift

Genetic code

- Each amino acid is specified by a triplet of 3 bases
- 4 bases:
A,C,G,T = 64 possible codons.
Actually 61 codons + 3 stop codons

	T		C		A		G		
T	TTT	phe	TCT	ser	TAT	tyr	TGT	cys	T
	TTC		TCC		TAC		TGC		C
	TTA	leu	TCA		TAA	stop	TGA	stop	A
	TTG		TCG		TAG		TGG	try	G
C	CTT	leu	CCT	pro	CAT	his	CGT	arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	ile	ACT	thr	AAT	asp	AGT	ser	T
	ATC		ACC		AAC		AGC		C
	ATA	ile	ACA		AAA	lys	AGA	arg	A
	ATG	met	ACG		AAG		AGG		G
G	GTT	val	GCT	ala	GAT	asp	GGT	gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	glu	GGA		A
	GTG		GCG		GAG		GGG		G

Translating sequences

- 6 possible reading frames, 3 in each direction

5'-AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT-3' +strand
 3'-TCAGCCGACTGACGCAAATGCTTACGCTAATGAGGGAA-5' -strand

Reverse complement

5'-AAGGGAGTAATCGCATTTCGTAAACGCAGTCAGCCGACT-3'

Translating sequences

- 6 possible reading frames, 3 in each direction

AGTCGGCTGACTGCGTTTACGAATGCGATTACT

+1

Reverse complement

AAGGGAGTAATCGCATTTCGTAAACGCAGTCAG

-1

	T		C		A		G		
T	TTT	phe	TCT	ser	TAT	tyr	TGT	cys	T
	TTC		TCC		TAC		TGC		C
	TTA	leu	TCA		TAA	stop	TGA	stop	A
	TTG		TCG		TAG		TGG	try	G
C	CTT	leu	CCT	pro	CAT	his	CGT	arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	ile	ACT	thr	AAT	asp	AGT	ser	T
	ATC		ACC		AAC		AGC		C
	ATA	met	ACA		AAA	lys	AGA	arg	A
	ATG		ACG		AAG		AGG		G
G	GTT	val	GCT	ala	GAT	asp	GGT	gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	glu	GGA		A
	GTG		GCG		GAG		GGG		G

Translating sequences

- 6 possible reading frames, 3 in each direction

Ser Arg Leu

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+1

Reverse complement

AAGGGAGTAATCGCATTCTGTAAACGCAGTCAGCCGACT

-1

Translating sequences

- 6 possible reading frames, 3 in each direction

Val Gly
 Stop
 AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT
 +2

AAGGGAGTAATCGCATTCTGTAAACGCAGTCAGCCGACT
 -2

Translating sequences

- 6 possible reading frames, 3 in each direction

Ser Ala Asp

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+3

AAGGGAGTAATCGCATTCTGTAAACGCAGTCAGCCGACT

-3

Translating sequences

- 6 possible reading frames, 3 in each direction

Arg Leu Thr

AGT CGG CTG ACT GCGTTTACGAATGCGATTACTCCCTT

+1

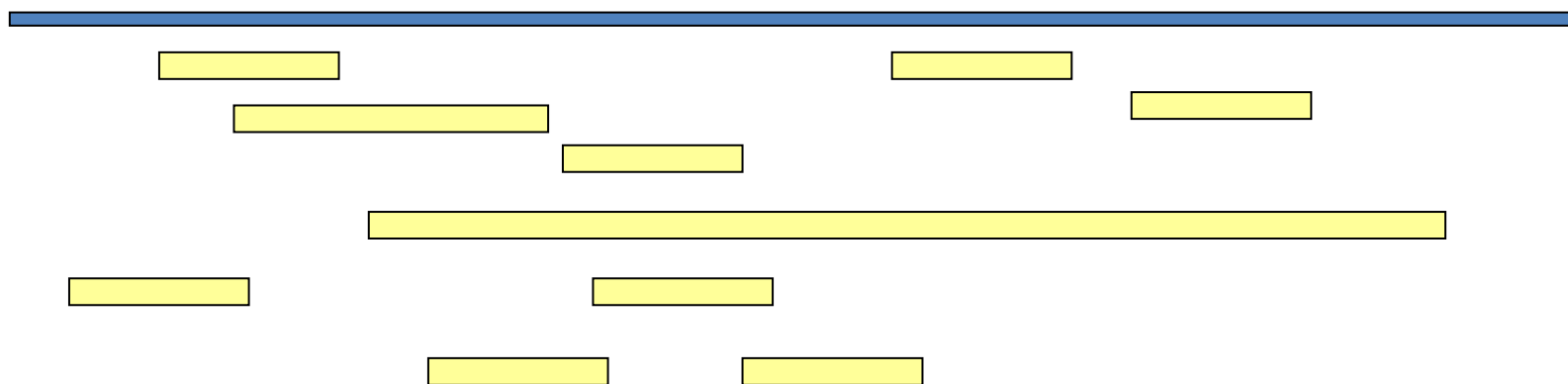
Reverse complement

AAG GGAG TAATCGCATTGTAACGCAGTCAGCCGACT

-1

Getting the final protein

- Six-frame translation
- Find longest ORF with initiation site, start codon and ending with stop codon



Gene Prediction -bacteria

Promoter

Start
codon

CDS

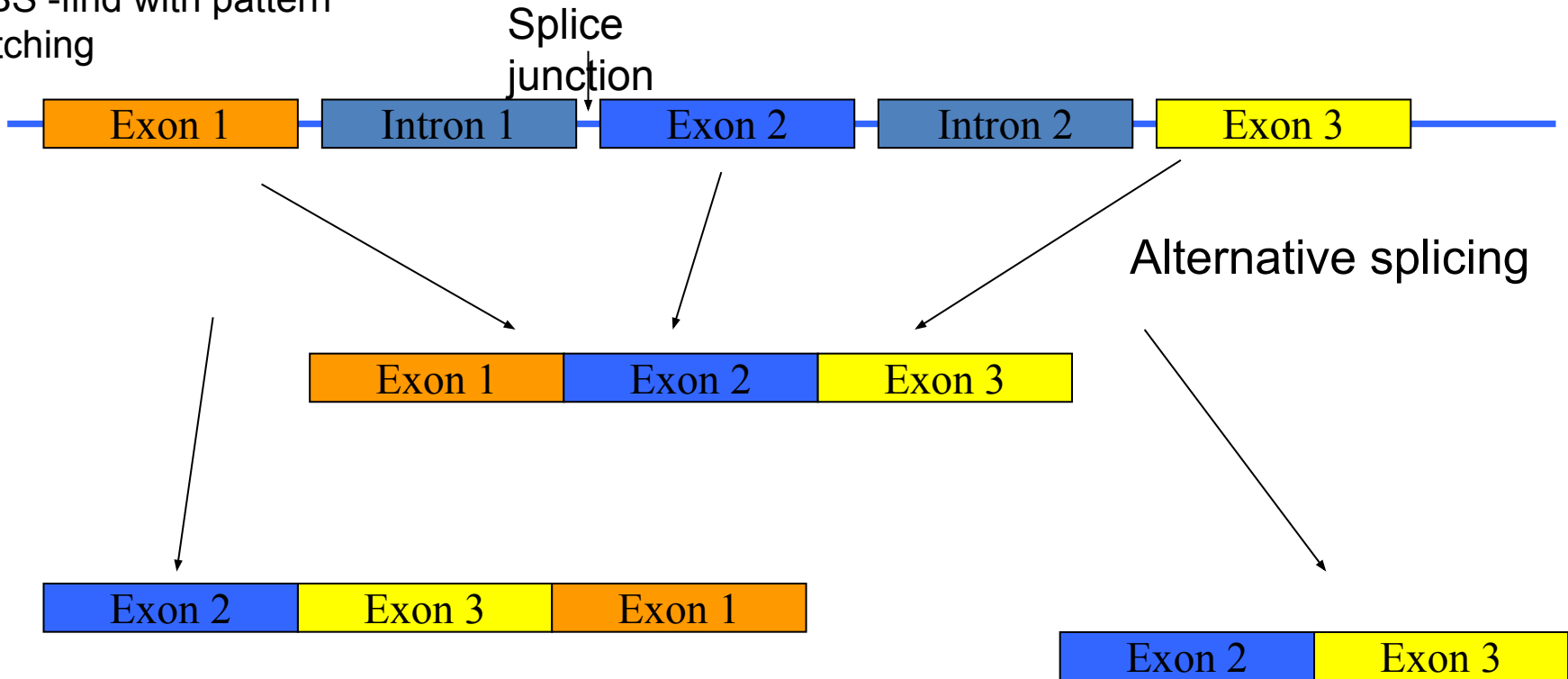
Stop
codon

```

1 GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT
                                     TT
41 CCAA AATCGCCTTTTGCTG TATATACTCA CAGCATAACTG
   CCAA -35          -10 TATACT      >
81 TATA TACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA
   +10 GGGGG Ribosomal binding site
121 GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG
161 CCAGACAGGTATG CCGCGACGCGTGCGGAAATCGCGCAG
201 CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC
241 TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC
281 CGGCGCATCACGCGGGATTTCGTCTGTTGCAGGAAGAGGAA
321 GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC
361 CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT
401 CGATCCTTCCTTATTC AAGCCGAATGCTGATTTCCTGCTG
441 CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG
481 ATGGTGACTTGCTGGCAGTGCATAAACTCAGGATGTACG
521 TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT
561 ACCGTTAAGCGCCTGAAAAAACAGGGCAATAAAGTCGAAC
601 TGTGCCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA
641 CCTTCGTCAGCAGAGCTTC ACCATTGAAGGGCTGGCGGTT
681 GGGGTTATTTCGCAACGGCGACTGGCTGTAACATATCTCTG
721 AGACCGCGATGCCGCTGCGCGTCCGGTTTGTITTTTCATC
761 TCTCTTCATCAGGCTTGTC TGCATGGCATTCTC ACTTCA
801 TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT
841 TCTCAATATCACCGTTCGTTGCTGGGACTGGTCGATAC
881 GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTGCGG
921 GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT
  
```

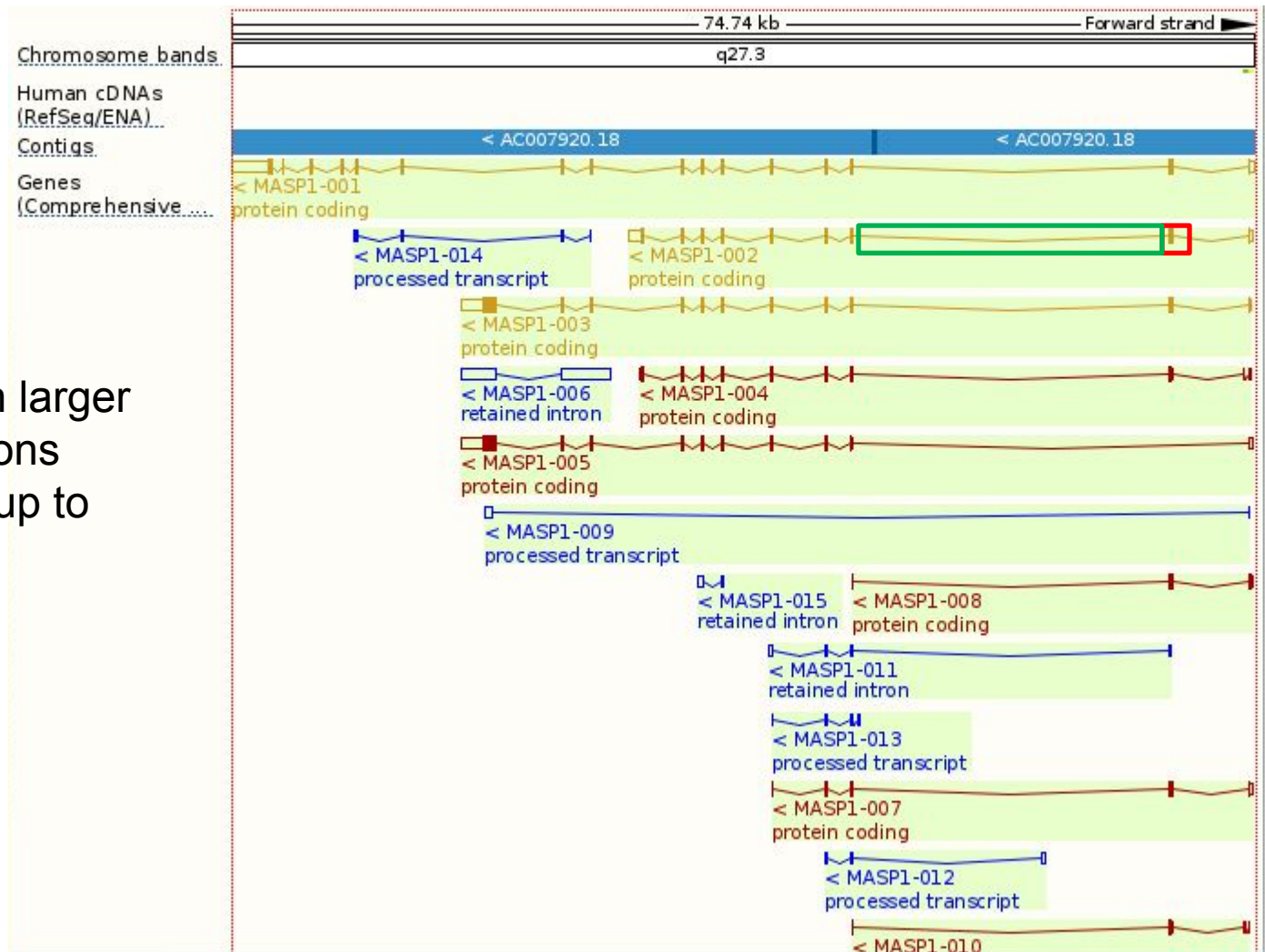

Complex Eukaryotic systems

Promoter region –many
TFBS -find with pattern
matching



Human introns and exons

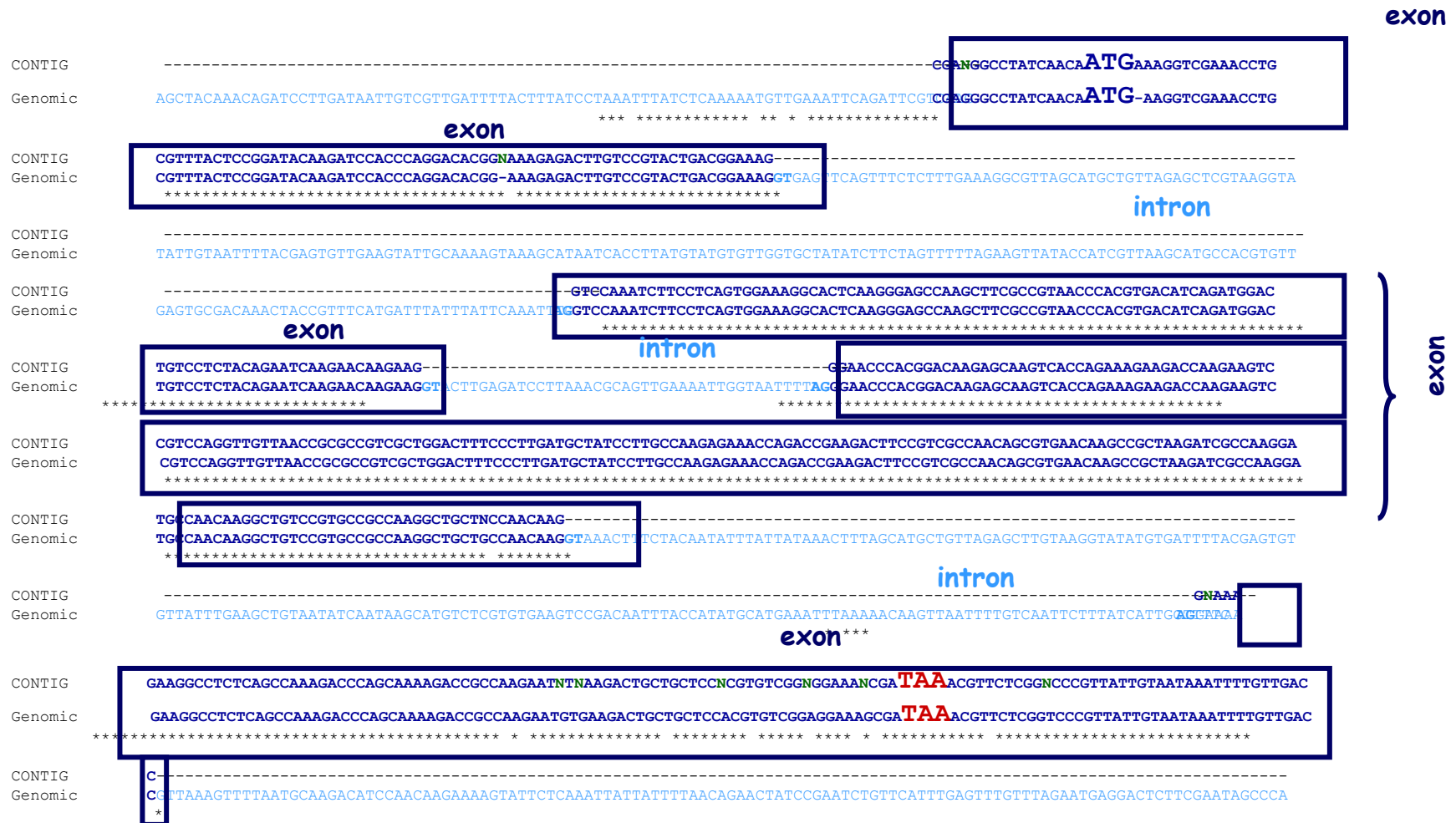
Introns are much larger than exons, introns could represent up to 95% of gene



Gene prediction in eukaryotes

- Identifying features (sometimes by PSSMs):
 - splice sites
 - start and stop sites
- Predict exons based on these signals
- Score exons based on signals and exon characteristics (coding sequences may have compositional biases)
- Use composition and homology information
- Assemble components into predicted gene structure
- Some methods use HMMs -features are states
- Use EST (expressed sequence tag –sequenced RNA) info

Using EST data: mRNA against genomic sequence



Gene Prediction software

- GeneMark –gene prediction for prokaryotes, eukaryotes and viruses:
<http://opal.biology.gatech.edu/GeneMark/>
- GENSCAN –for vertebrate, maize and Arabidopsis sequences: <http://genes.mit.edu/GENSCAN.html>
- Microbial Gene Prediction System
<http://compbio.ornl.gov/generation/>
- Glimmer –bacteria, archae and viruses
<http://www.tigr.org/software/glimmer/>
- GRAIL –for eukaryotes, includes splice info, homology, etc. <http://compbio.ornl.gov/grailexp/>

Other translators and promoter prediction

- NCBI ORF Finder:
(<http://www.ncbi.nlm.nih.gov/gorf/gorf.htm>)
- Promoter 2.0 Prediction Server
(<http://www.cbs.dtu.dk/services/Promoter/>)
- MCPromoter MM:II
(<http://genes.mit.edu/McPromoter.html>)
- BPRROM -prediction of bacterial promoters, etc.

RNA sequence analysis

- Many different types of RNA e.g. tRNA, rRNA, mRNA etc.
- Some have activities e.g. ribozymes
- Many new programs for identification of non-coding RNA, miRNAs etc and their targets
- Secondary structure of RNA is NB for stability and often function
- RNA levels are NB for final protein levels, they measure gene expression –ESTs, microarrays

Summary and conclusions

- Basic sequence analysis is finding features on a sequence
- This could be small features
 - Restriction sites -> cloning
 - Primer sites -> PCR
- Or combinations of features:
 - Gene signals -> gene prediction
- Features found by nature of their “conservation” or pattern matching
- ***Practical assignment*** –retrieve a DNA sequence, run some basic analysis programs