H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics online course: IBT

Bioinformatics resources and databases: Lecture 3: DNA sequence analysis

Nicola Mulder

cc BY NC SA

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

cc BY NC SA

Learning Outcomes

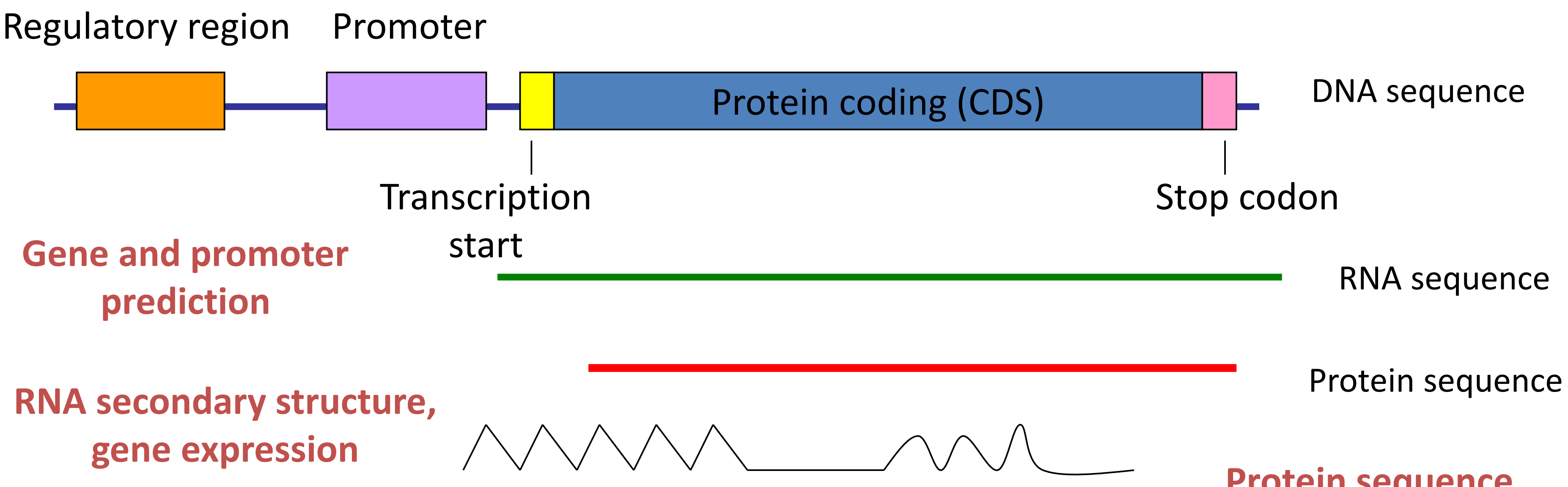
- Understand how to find a DNA sequence and save it in the correct format
- Identify features on the sequence such as coding regions, restriction enzyme sites, etc.
- Design primers for amplification of a DNA sequence
- Interpret sequence analysis results and understand the biological impact of functional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

cc BY NC SA

Two major components to Bioinformatics

- 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

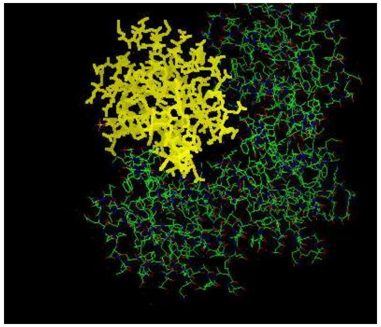


Gene and promoter prediction

RNA secondary structure, gene expression

Protein sequence analysis

Restriction mapping for cloning, primer design for PCR



cc BY NC SA

Overview

- Assume sequence is retrieved from the database
- General text/format manipulation and accession numbers
- DNA sequences
  - Restriction analysis
  - Primer design
  - Finding features –coding and non-coding
  - Gene prediction
- RNA sequence analysis
  - Summary of kinds of analyses possible

cc BY NC SA

Sequence formats: Fasta

```
> [title]
[sequence]

>seq1
GGAAAATTAGATGCATGGGAAAAAATTA
GGAAAATTAGACAAATGGGAAAAAATTA
>seq2
AAGTCCCTGGATTTACCCAATGCAGTCGA
CATCGCATTT
```

BY NC SA

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   TGTAA ACAAA TAATG NAACA GATAC AACCA GCTCT TCAGA CAGGA ACAGA
```

Converting between sequence formats (save options)

cc BY NC SA

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

cc BY NC SA

Accession numbers

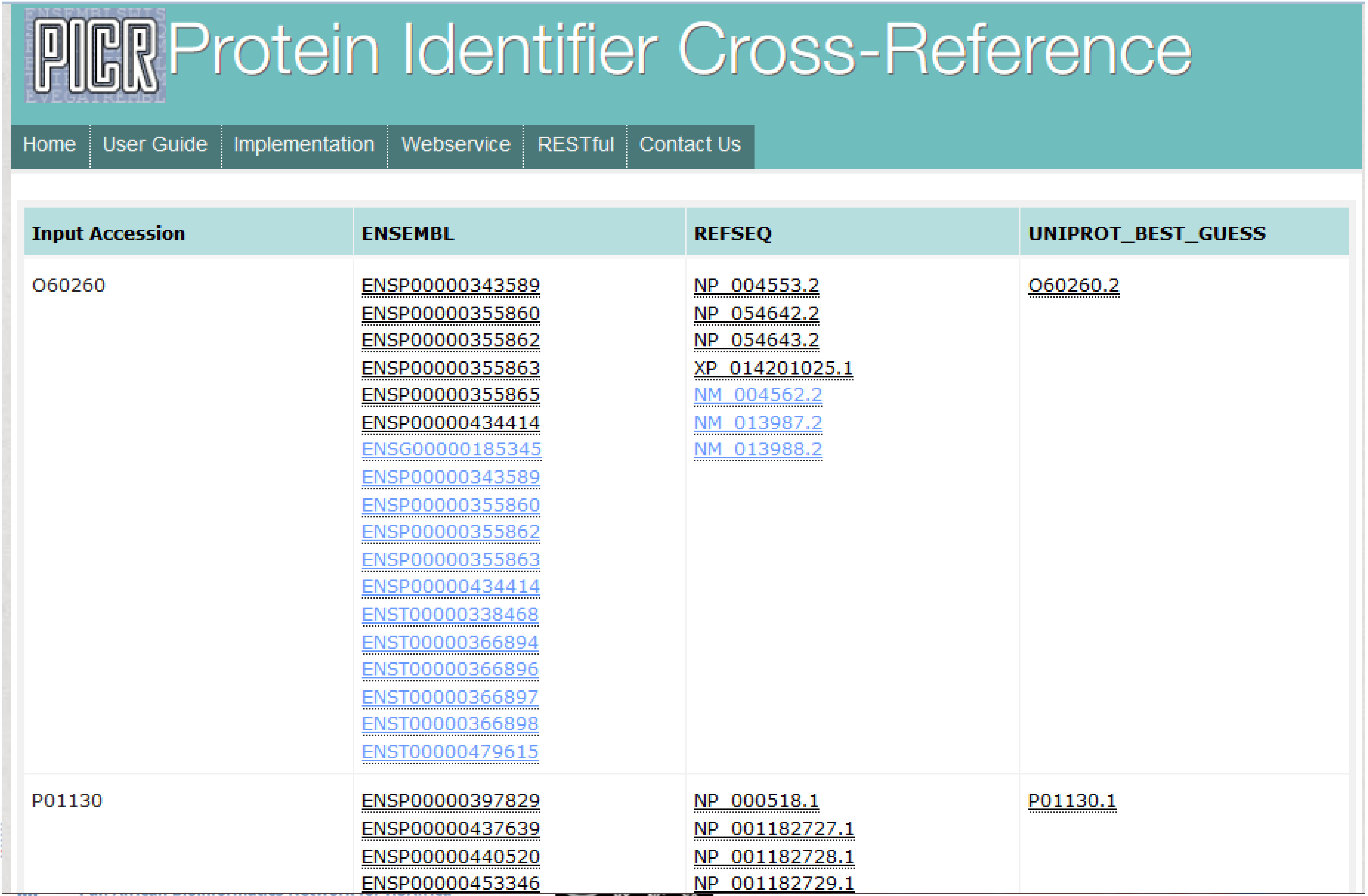
- **GenBank/EMBL/DDBJ**: 1 letter & digits, e.g.: U12345 or 2 letters & 6 digits, e.g.: AY123456
- **GenPept** Sequence Records -3 letters & 5 digits, e.g.: AAA12345
- **UniProt** -All 6 characters: [A,B,O,P,Q] [0-9] [A-Z,0-9] [A-Z,0-9] [A-Z,0-9] [0-9], e.g.: P12345 and Q9JJS7

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

cc BY NC SA

Example conversion



PICR Protein Identifier Cross-Reference

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

BY NC SA

Bioinformatics and cloning

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

cc BY NC SA

Restriction enzyme analysis

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

Microorganism	Enzyme	Sequences	Notes
Haemophilus aegitius	*Hae*III	5’…GG CC..3’ 3’…CC GG..5’	Blunt end
Haemophilus haemolytica	*Hha*I	5’…GC G C..3’ 3’…CG C G..5’	3’ single strand
Escherichia coli	*Eco*RI	5’…G AATT C..3’ 3’…C TTAA G..5’	5’ single strand

5’ **3’**

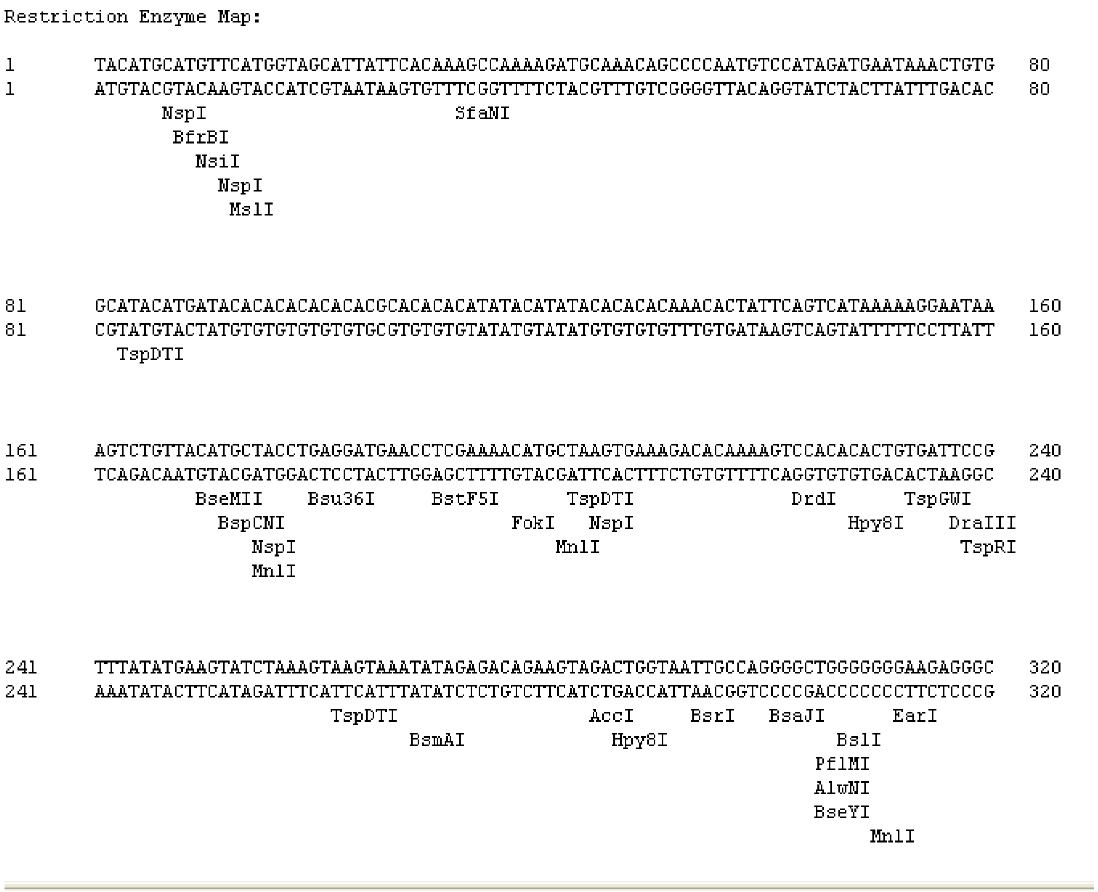
……..GG CC……..
………CC GG…….

……..GCG C….
………C GCG…….

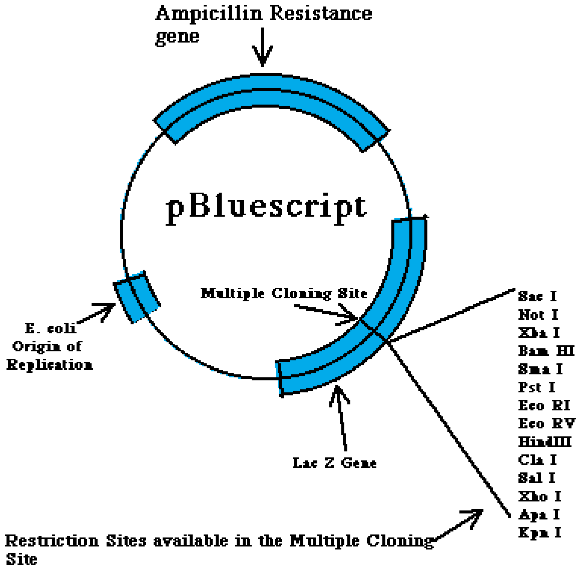
…G AATTC…..
…CTTAA G…..

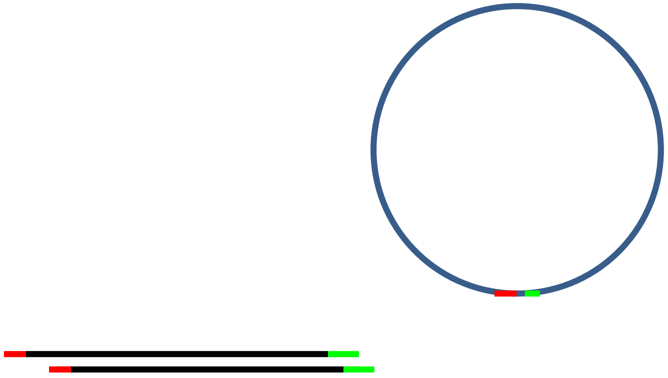
BY NC SA

Restriction map



```
Restriction Enzyme Map:

1       TACATGCATGTTCATGGTAGCATTATTCACAAAGCCAAAAGATGCAAACAGCCCCAATGTCCATAGATGAATAAACTGTG   80
1       ATGTACGTACAAGTACCATCGTAATAAGTGTTTCGGTTTTCTACGTTTGTCGGGGTTACAGGTATCTACTTATTTGACAC   80
              NspI                      SfaNI
               BfrBI
                 NsiI
                   NspI
                    MslI


81      GCATACATGATACACACACACACACGCACACACATATACATATACACACACAAACACTATTCAGTCATAAAAAGGAATAA   160
81      CGTATGTACTATGTGTGTGTGTGTGCGTGTGTGTATATGTATATGTGTGTGTTTGTGATAAGTCAGTATTTTTCCTTATT   160
          TspDTI


161     AGTCTGTTACATGCTACCTGAGGATGAACCTCGAAAACATGCTAAGTGAAAGACACAAAAGTCCACACACTGTGATTCCG   240
161     TCAGACAATGTACGATGGACTCCTACTTGGAGCTTTTGTACGATTCACTTTCTGTGTTTTCAGGTGTGTGACACTAAGGC   240
                 BseMII    Bsu36I     BstF5I      TspDTI              DrdI      TspGWI
                   BspCNI                    FokI   NspI                   Hpy8I    DraIII
                      NspI                       MnlI                                TspRI
                      MnlI


241     TTTATATGAAGTATCTAAAGTAAGTAAATATAGAGACAGAAGTAGACTGGTAATTGCCAGGGGCTGGGGGGGAAGAGGGC   320
241     AAATATACTTCATAGATTTCATTCATTTATATCTCTGTCTTCATCTGACCATTAACGGTCCCCGACCCCCCCTTCTCCCG   320
                             TspDTI                 AccI     BsrI   BsaJI      EarI
                                    BsmAI             Hpy8I               BslI
                                                                        PflMI
                                                                        AlwNI
                                                                        BseYI
                                                                             MnlI
```





cc BY NC SA

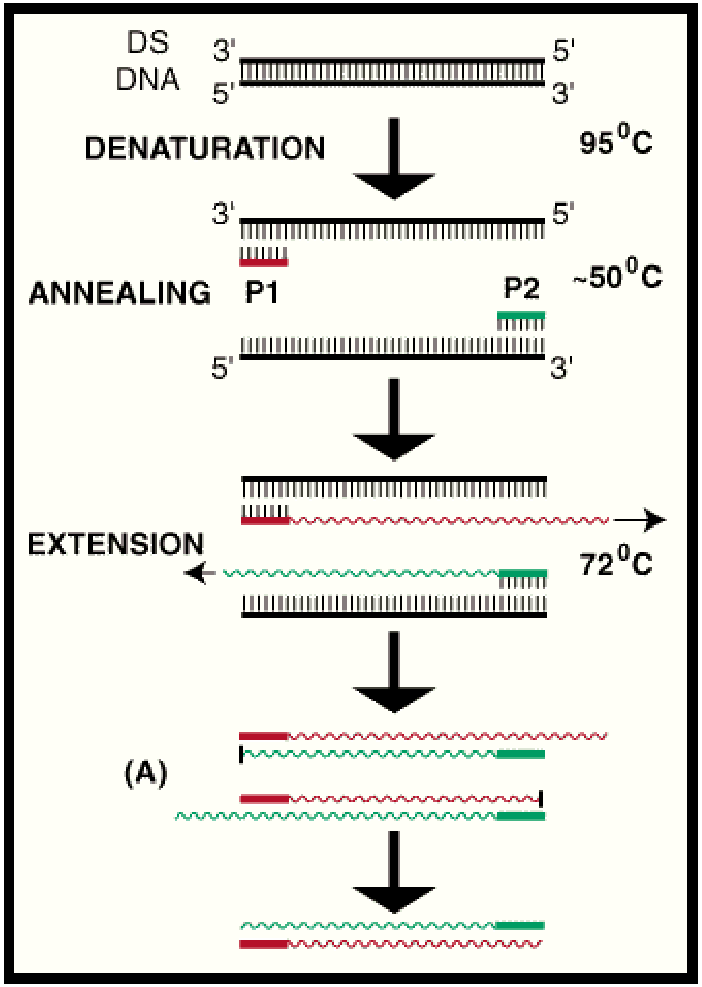
Removing vector sequence

- Vector contamination can be identified by searching your sequence against a database of vector sequences (UniVec) e.g. http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html –uses BLASTN
- Need to hope vector is only at extremities and not in insert (contamination!)

cc BY NC SA

PCR and primer design

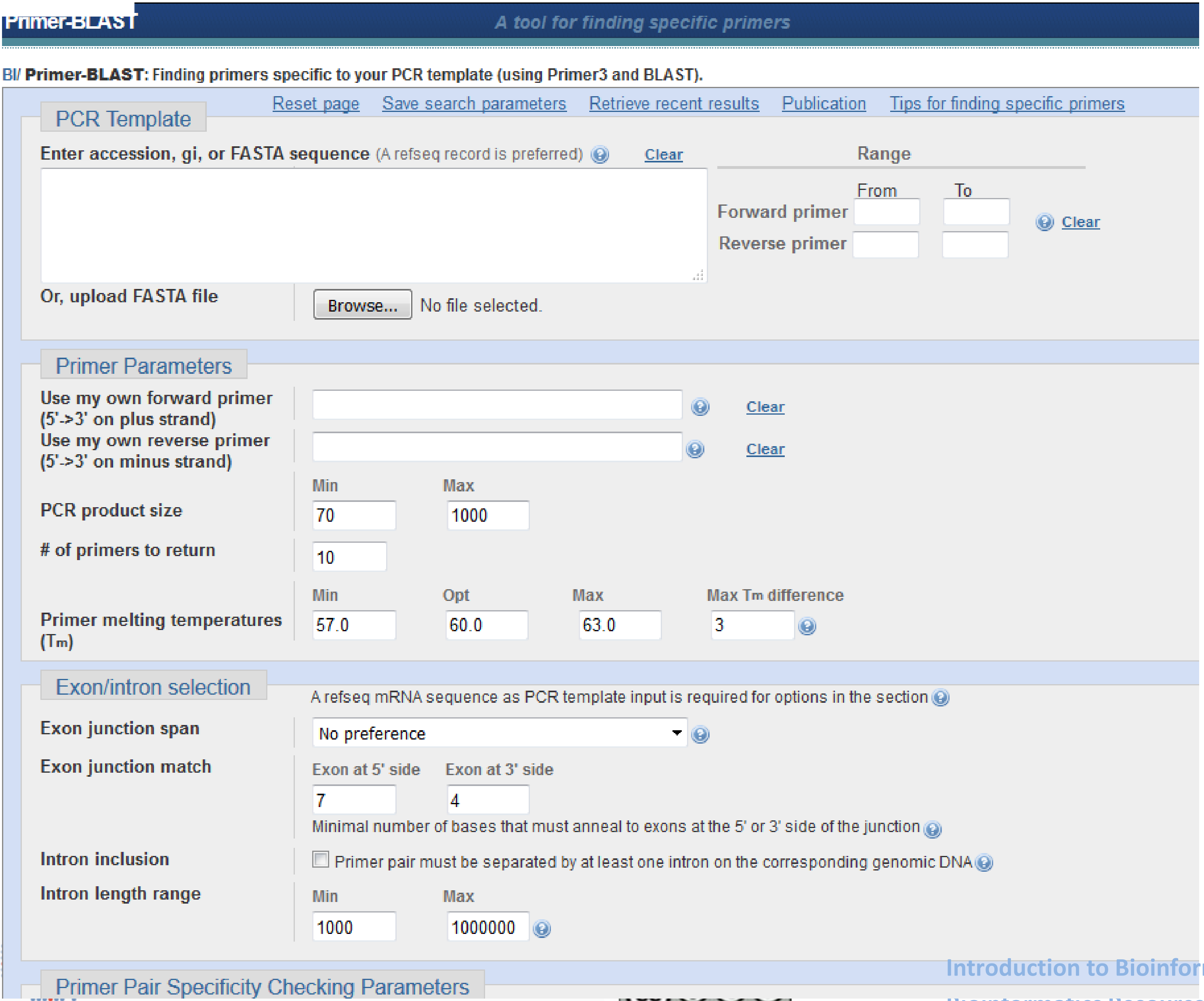
- Can engineer restriction sites
- Primers should be similar length and Tm
- Should amplify only required piece from genome



H3ABioNet
Pan African Bioinformatics Network for H3Africa

cc BY NC SA

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).

PCR Template

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

Range

	From	To
Forward primer		
Reverse primer		

Clear

Or, upload FASTA file Browse... No file selected.

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

	Min	Opt	Max	Max Tm difference
PCR product size	70		1000	
# of primers to return	10			
Primer melting temperatures (Tm)	57.0	60.0	63.0	3

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span: No preference

Exon junction match: Exon at 5' side 7; Exon at 3' side 4

Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction

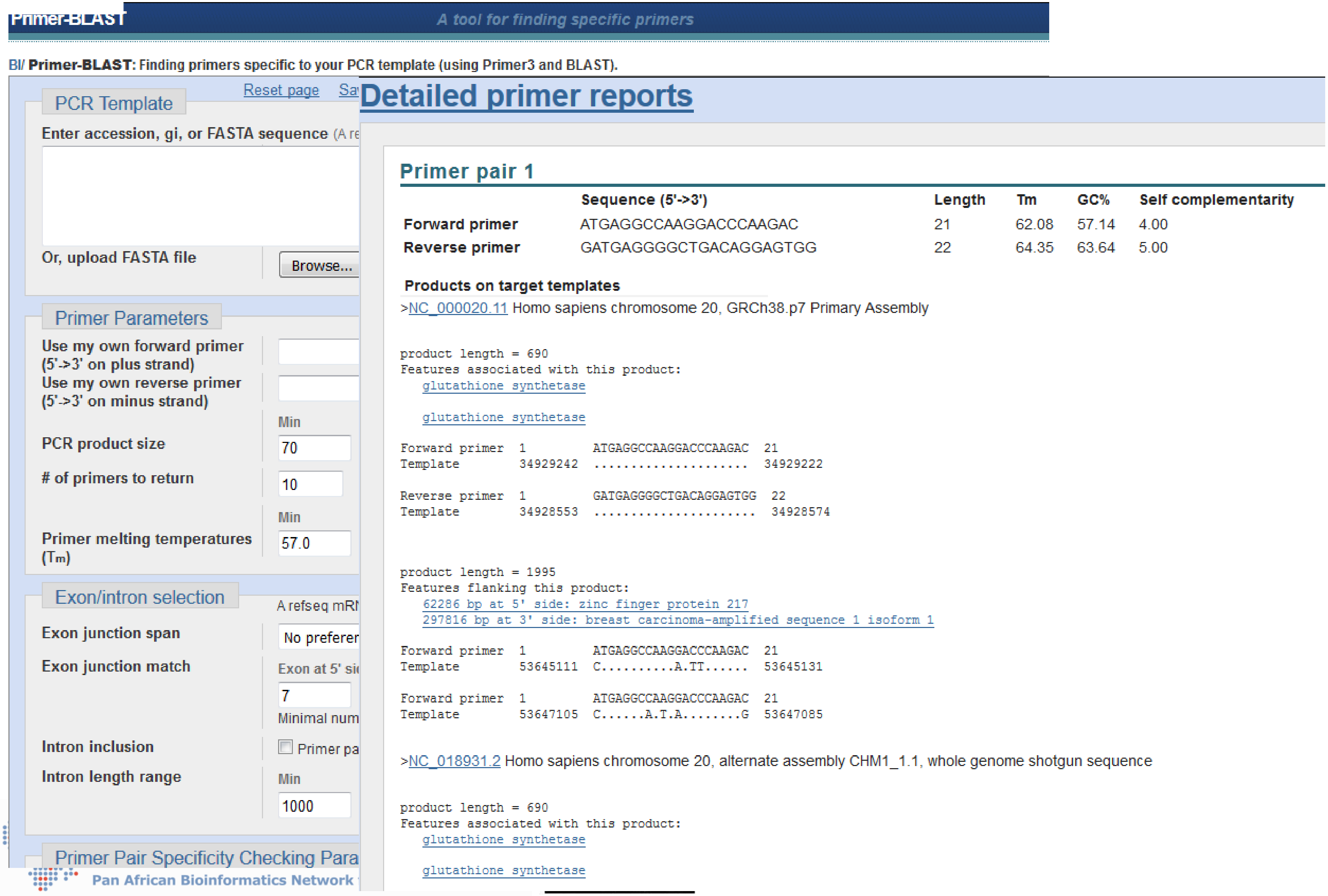
Intron inclusion: □ Primer pair must be separated by at least one intron on the corresponding genomic DNA

Intron length range: Min 1000; Max 1000000

Primer Pair Specificity Checking Parameters

BY NC SA

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 Sa

PCR Template

Enter accession, gi, or FASTA sequence (A re

Or, upload FASTA file Browse...

Primer Parameters

Use my own forward primer (5'->3' on plus strand)

Use my own reverse primer (5'->3' on minus strand)

	Min
PCR product size	70
# of primers to return	10
Primer melting temperatures (Tm)	57.0

Exon/intron selection

A refseq mR

Exon junction span	No preferer
Exon junction match	Exon at 5' si 7 Minimal num
Intron inclusion	□ Primer pa
Intron length range	Min 1000

Primer Pair Specificity Checking Para

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
```

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)

cc BY NC SA

Translation

- Can choose frame if you know it
- Otherwise 6-frame translation:
  - Choose start codon ATG
  - Otherwise lists all codons between stop codons
- Results –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**

cc BY NC SA

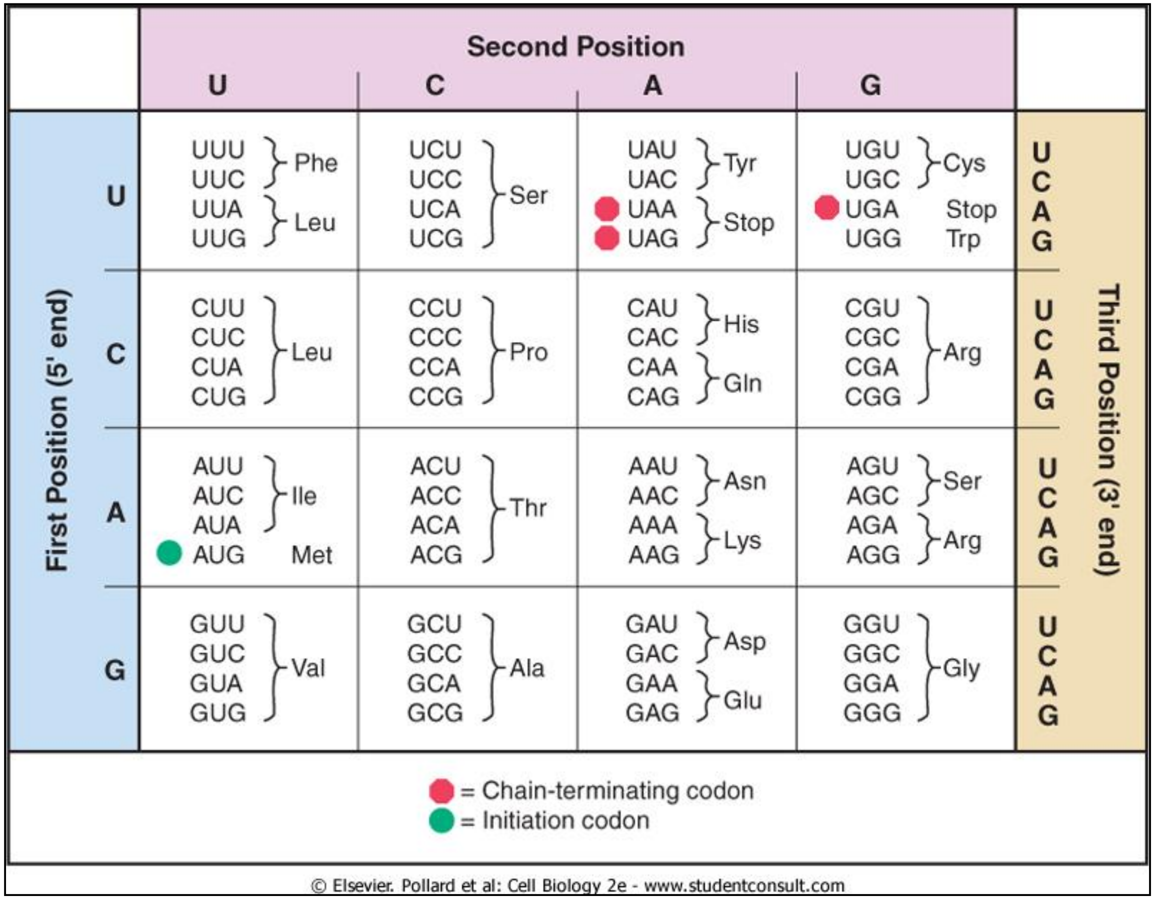
Open reading frame

- 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

BY NC SA

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



First Position (5' end)	Second Position U	 C	 A	 G	Third Position (3' end)
U	UUU UUC } Phe UUA UUG } Leu	UCU UCC UCA UCG } Ser	UAU UAC } Tyr 🔴 UAA 🔴 UAG } Stop	UGU UGC } Cys 🔴 UGA Stop UGG Trp	U C A G
C	CUU CUC CUA CUG } Leu	CCU CCC CCA CCG } Pro	CAU CAC } His CAA CAG } Gln	CGU CGC CGA CGG } Arg	U C A G
A	AUU AUC AUA } Ile 🟢 AUG Met	ACU ACC ACA ACG } Thr	AAU AAC } Asn AAA AAG } Lys	AGU AGC } Ser AGA AGG } Arg	U C A G
G	GUU GUC GUA GUG } Val	GCU GCC GCA GCG } Ala	GAU GAC } Asp GAA GAG } Glu	GGU GGC GGA GGG } Gly	U C A G

🔴 = Chain-terminating codon
🟢 = Initiation codon

© Elsevier. Pollard et al: Cell Biology 2e - www.studentconsult.com

Translating sequences

- 6 possible reading frames, 3 in each direction

Ser Arg Leu

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+1

Reverse complement

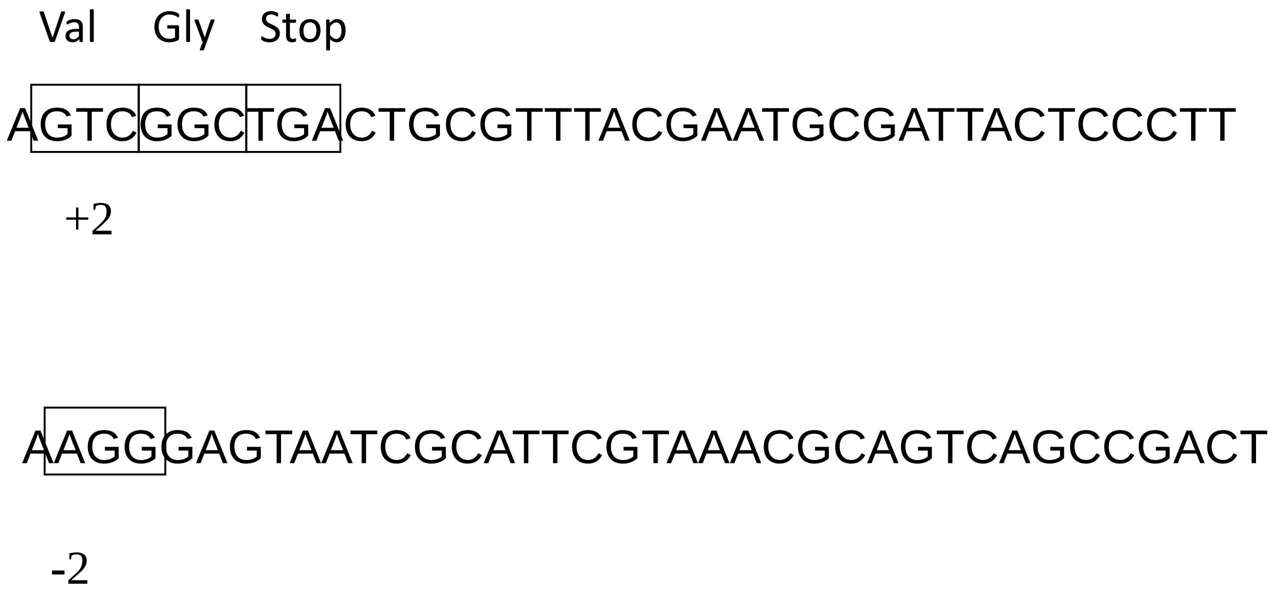
AAGGGAGTAATCGCATTCGTAAACGCAGTCAGCCGACT

-1

cc BY NC SA

Translating sequences

- 6 possible reading frames, 3 in each direction



Val Gly Stop

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+2

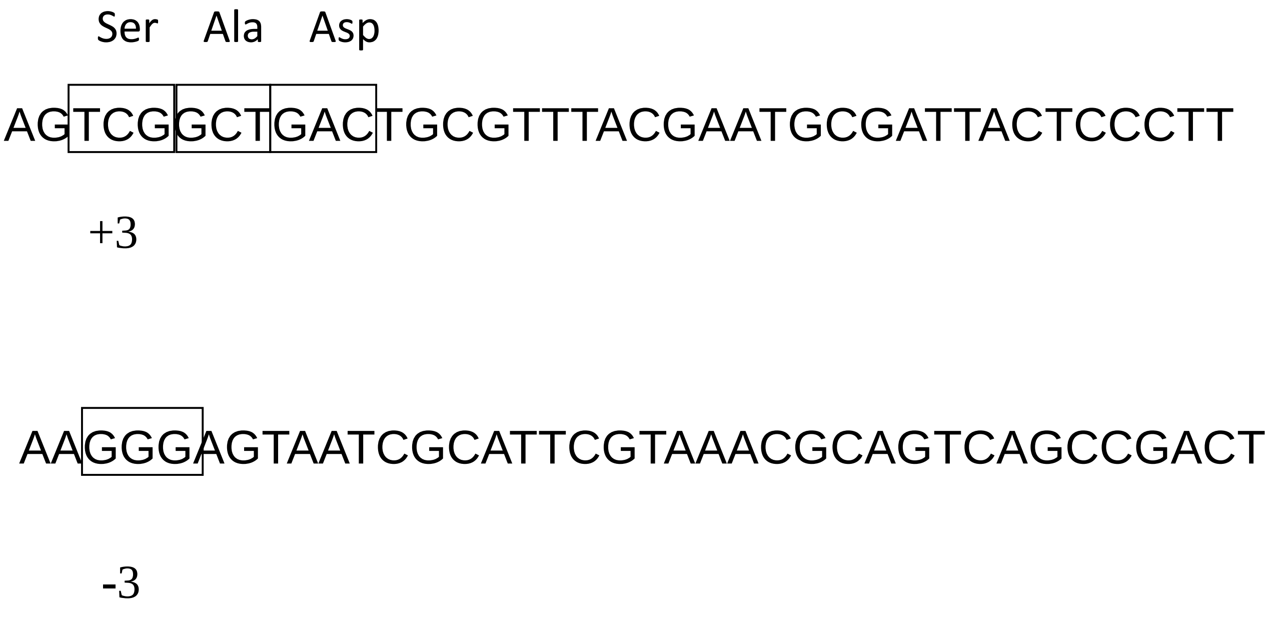
AAGGGAGTAATCGCATTCGTAAACGCAGTCAGCCGACT

-2

cc BY NC SA

Translating sequences

- 6 possible reading frames, 3 in each direction



Ser Ala Asp

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+3

AAGGGAGTAATCGCATTCGTAAACGCAGTCAGCCGACT

-3

Translating sequences

- 6 possible reading frames, 3 in each direction

Arg Leu Thr

AGTCGGCTGACTGCGTTTACGAATGCGATTACTCCCTT

+1

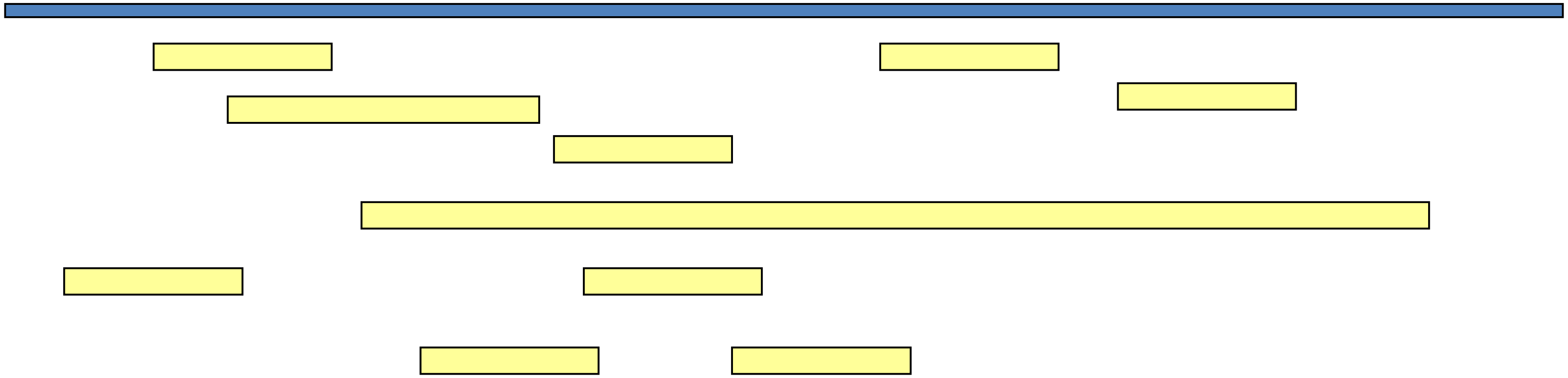
Reverse complement

AAGGGAGTAATCGCATTCGTAAACGCAGTCAGCCGACT

-1

Getting the final protein

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



BY NC SA

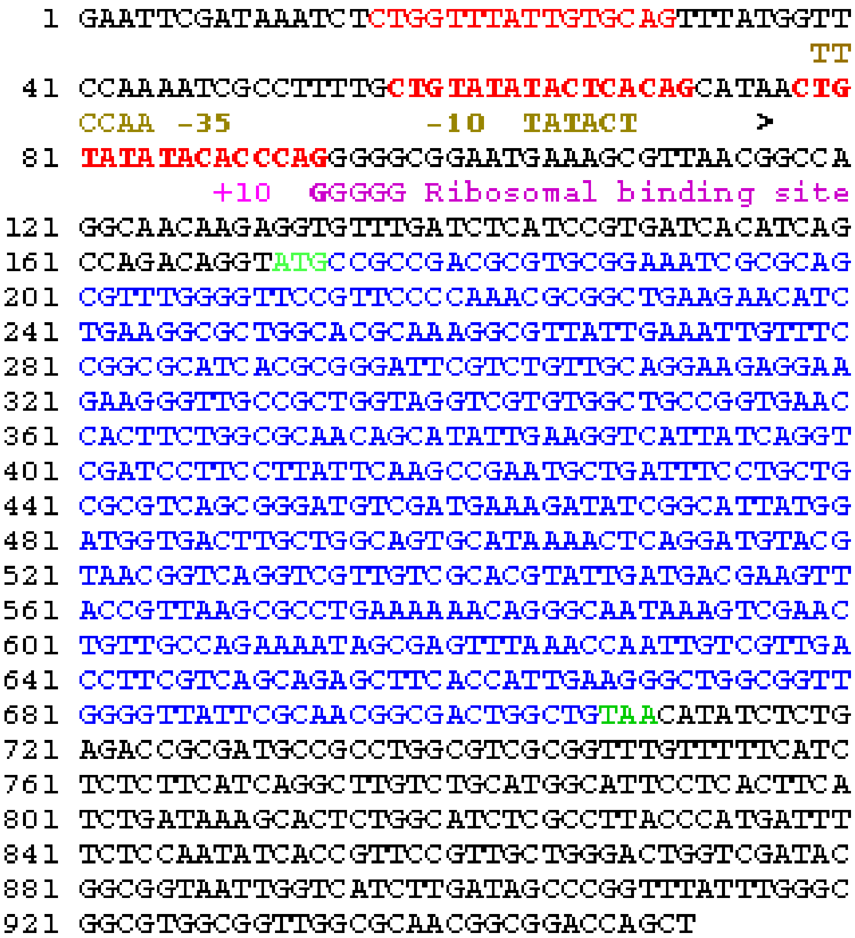
Gene Prediction -bacteria

Promoter

Start codon

CDS

Stop codon

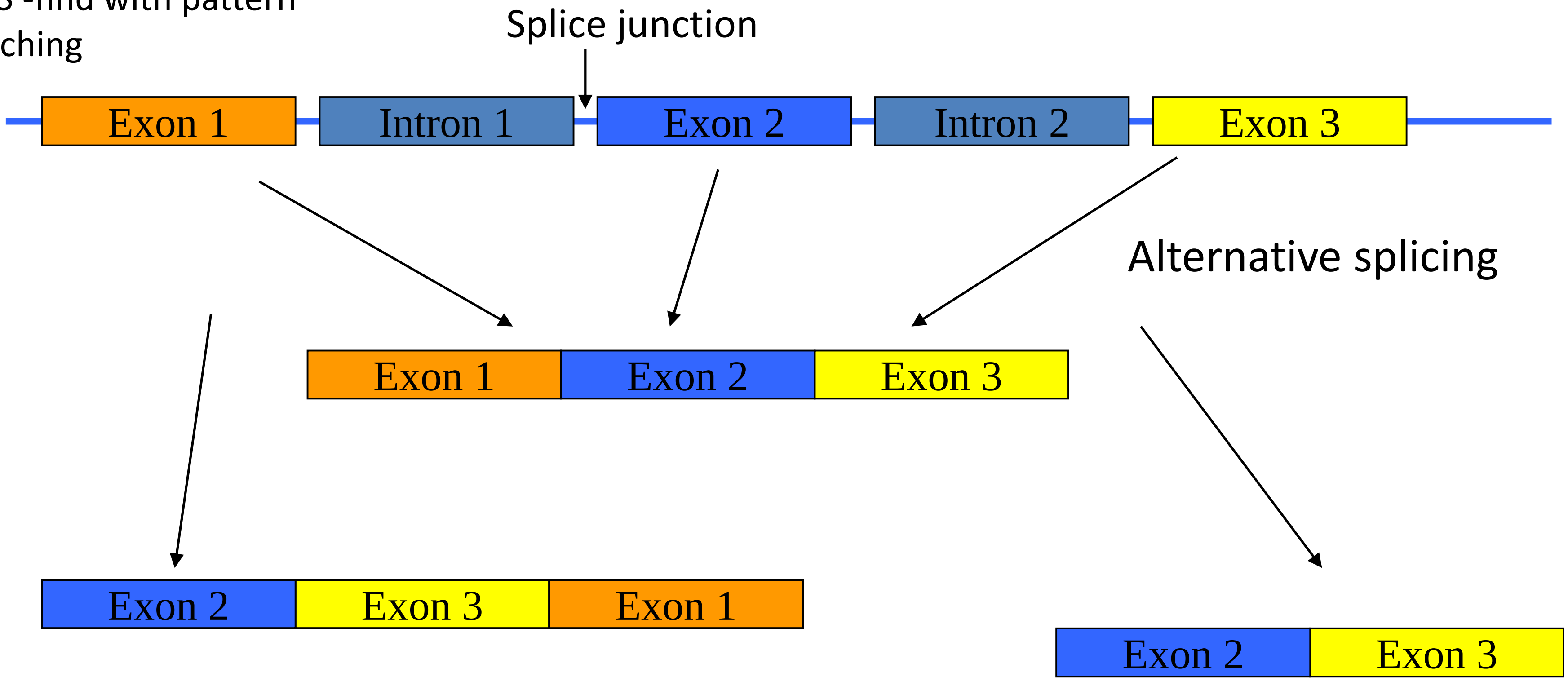


```
   1 GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT
                                           TT
  41 CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG
     CCAA -35          -10  TATACT      >
  81 TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA
            +10  GGGGG Ribosomal binding site
 121 GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG
 161 CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG
 201 CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC
 241 TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC
 281 CGGCGCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAA
 321 GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC
 361 CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT
 401 CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG
 441 CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG
 481 ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG
 521 TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT
 561 ACCGTTAAGCGCCTGAAAAAACAGGGCAATAAAGTCGAAC
 601 TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA
 641 CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT
 681 GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG
 721 AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC
 761 TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA
 801 TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT
 841 TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC
 881 GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC
 921 GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT
```

BY NC SA

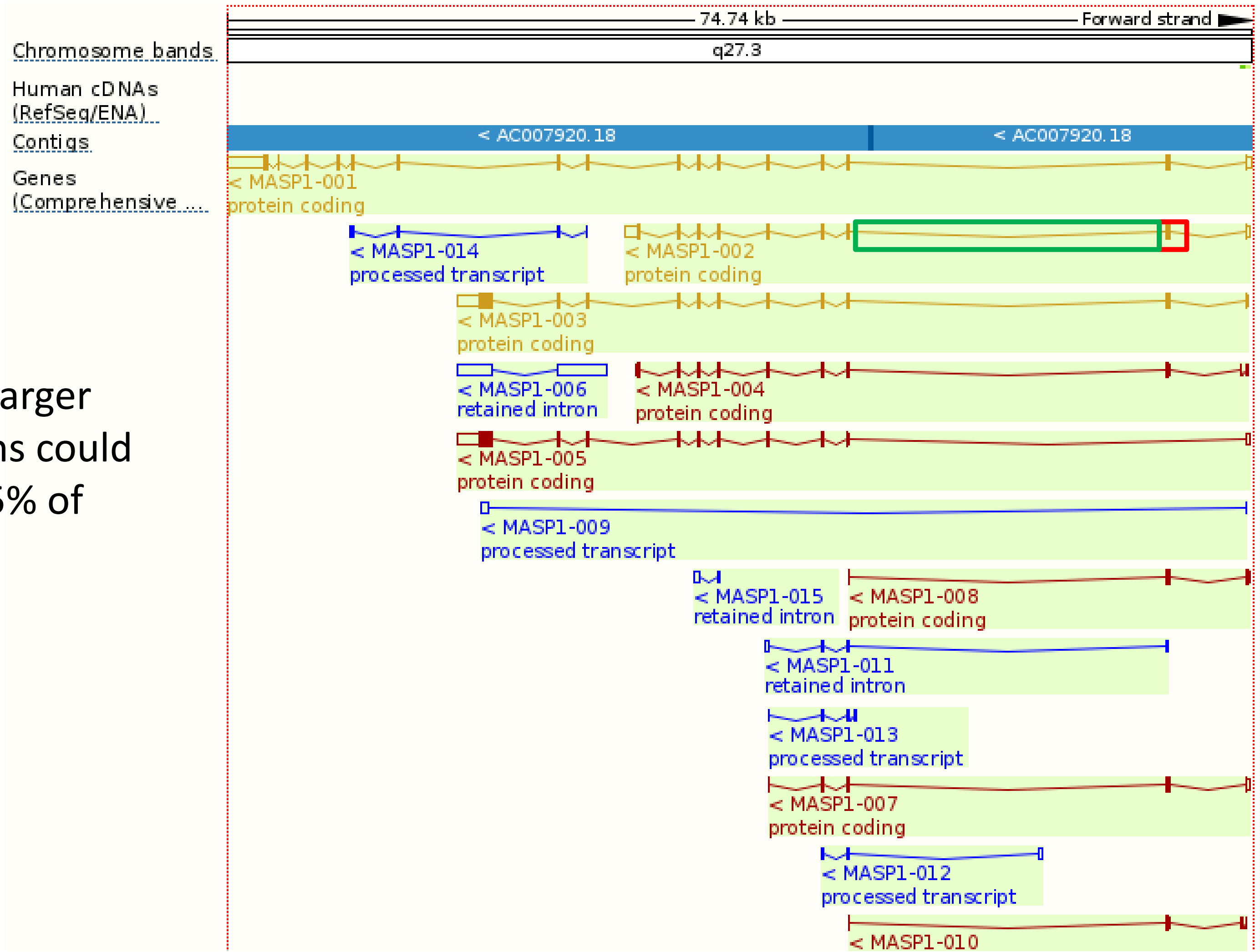
Complex Eukaryotic systems

Promoter region –many TFBS -find with pattern matching



BY NC SA

Human introns and exons



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

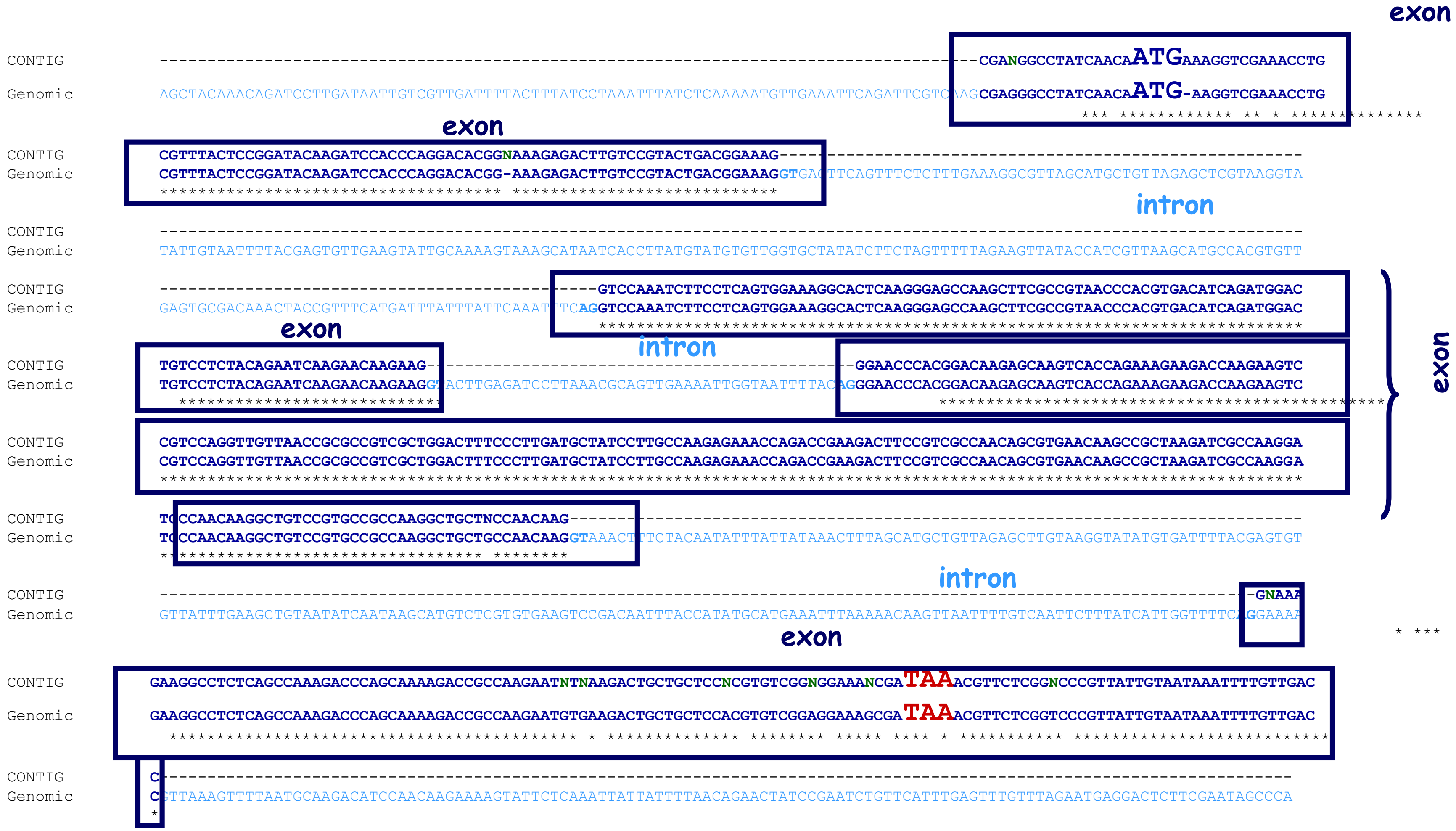
cc BY NC SA

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 info

BY NC SA

Using EST data: mRNA against genomic sequence



```
CONTIG      --------------------------------------------------------------------------------------CGANGGCCTATCAACAATGAAAGGTCGAAACCTG

Genomic     AGCTACAAACAGATCCTTGATAATTGTCGTTGATTTTACTTTATCCTAAATTTATCTCAAAAATGTTGAAATTCAGATTCGTCAAGCGAGGGCCTATCAACAATG-AAGGTCGAAACCTG
                                                                                                       ***  ************  **  *  **************

CONTIG      CGTTTACTCCGGATACAAGATCCACCCAGGACACGGNAAAGAGACTTGTCCGTACTGACGGAAAG-------------------------------------------------------
Genomic     CGTTTACTCCGGATACAAGATCCACCCAGGACACGG-AAAGAGACTTGTCCGTACTGACGGAAAGGTGACTTCAGTTTCTCTTTGAAAGGCGTTAGCATGCTGTTAGAGCTCGTAAGGTA
            ************************************ ****************************

CONTIG      ------------------------------------------------------------------------------------------------------------------------
Genomic     TATTGTAATTTTACGAGTGTTGAAGTATTGCAAAAGTAAAGCATAATCACCTTATGTATGTGTTGGTGCTATATCTTCTAGTTTTTAGAAGTTATACCATCGTTAAGCATGCCACGTGTT

CONTIG      ----------------------------------------------GTCCAAATCTTCCTCAGTGGAAAGGCACTCAAGGGAGCCAAGCTTCGCCGTAACCCACGTGACATCAGATGGAC
Genomic     GAGTGCGACAAACTACCGTTTCATGATTTATTTATTCAAATTTCAGGTCCAAATCTTCCTCAGTGGAAAGGCACTCAAGGGAGCCAAGCTTCGCCGTAACCCACGTGACATCAGATGGAC
                                                          **************************************************************************

CONTIG      TGTCCTCTACAGAATCAAGAACAAGAAG---------------------------------------------GGAACCCACGGACAAGAGCAAGTCACCAGAAAGAAGACCAAGAAGTC
Genomic     TGTCCTCTACAGAATCAAGAACAAGAAGGTACTTGAGATCCTTAAACGCAGTTGAAAATTGGTAATTTTACAGGGAACCCACGGACAAGAGCAAGTCACCAGAAAGAAGACCAAGAAGTC
              **************************                                                  ************************************************

CONTIG      CGTCCAGGTTGTTAACCGCGCCGTCGCTGGACTTTCCCTTGATGCTATCCTTGCCAAGAGAAACCAGACCGAAGACTTCCGTCGCCAACAGCGTGAACAAGCCGCTAAGATCGCCAAGGA
Genomic     CGTCCAGGTTGTTAACCGCGCCGTCGCTGGACTTTCCCTTGATGCTATCCTTGCCAAGAGAAACCAGACCGAAGACTTCCGTCGCCAACAGCGTGAACAAGCCGCTAAGATCGCCAAGGA
            ************************************************************************************************************************

CONTIG      TCCCAACAAGGCTGTCCGTGCCGCCAAGGCTGCTNCCAACAAG-----------------------------------------------------------------------------
Genomic     TCCCAACAAGGCTGTCCGTGCCGCCAAGGCTGCTGCCAACAAGGTAAACTTTCTACAATATTTATTATAAACTTTAGCATGCTGTTAGAGCTTGTAAGGTATATGTGATTTTACGAGTGT
            ********************************** ********

CONTIG      -------------------------------------------------------------------------------------------------------------------GNAAA
Genomic     GTTATTTGAAGCTGTAATATCAATAAGCATGTCTCGTGTGAAGTCCGACAATTTACCATATGCATGAAATTTAAAAACAAGTTAATTTTGTCAATTCTTTATCATTGGTTTTCAGGAAAA
                                                                                                                                          *  ***

CONTIG      GAAGGCCTCTCAGCCAAAGACCCAGCAAAAGACCGCCAAGAATNTNAAGACTGCTGCTCCNCGTGTCGGNGGAAANCGATAAACGTTCTCGGNCCCGTTATTGTAATAAATTTTGTTGAC

Genomic     GAAGGCCTCTCAGCCAAAGACCCAGCAAAAGACCGCCAAGAATGTGAAGACTGCTGCTCCACGTGTCGGAGGAAAGCGATAAACGTTCTCGGTCCCGTTATTGTAATAAATTTTGTTGAC
              ****************************************** * ************** ******** ***** **** * *********** *************************

CONTIG      C-----------------------------------------------------------------------------------------------------------------------
Genomic     CGTTAAAGTTTTAATGCAAGACATCCAACAAGAAAAGTATTCTCAAATTATTATTTTAACAGAACTATCCGAATCTGTTCATTTGAGTTTGTTTAGAATGAGGACTCTTCGAATAGCCCA
            *
```

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/

cc BY NC SA

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)
- BPROM -prediction of bacterial promoters, etc.

BY NC SA

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

cc BY NC SA

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

cc BY NC SA