MSc Bioinformatics

Course

Fundamentals of Bioinformatics

Lecture 2: Genome Principles and Evolution

Centre for Integrative Bioinformatics VU (IBIVU)
Faculty of Exact Sciences / Faculty of Earth and Life Sciences
http://ibi.vu.nl, heringa@few.vu.nl, 87649 (Heringa), Room P1.28
OH BRAD, THEY SAY THERE'S DNA IN MY BODY!

WHO CARES, DARLING, ...WHO CARES...
DNA makes RNA makes Protein

*From gene to function*

DNA makes RNA makes Protein.
Cell complexity – metabolic pathways
Central Dogma of Molecular Biology

Replication is carried out by DNA polymerase
Transcription is carried out by RNA polymerase (II)
Translation is performed on ribosomes

Transcription + Translation = Expression

Going against the central dogma:
Reverse transcriptase copies RNA into DNA
DNA replication

1. Double-stranded DNA unwinds.

2. The junction of the unwound molecules is a replication fork.

3. A new strand is formed by pairing complementary bases with the old strand.

4. Two identical molecules are made. Each has one new and one old DNA strand.
Enzymes in DNA replication

- Helicase unwinds parental double helix
- Binding proteins stabilize separate strands
- Primase adds short primer to template strand
- DNA polymerase binds nucleotides to form new strands
- Exonuclease removes RNA primer and inserts the correct bases
- Ligase joins Okazaki fragments and seals other nicks in sugar-phosphate backbone
DNA replication is not perfect

- Despite the proofreading capacity of DNA polymerase, wrong bases are incorporated and nucleotides are skipped or inserted!
- Frequency of mistakes: after corrections (‘repair’) ~ 1 in $10^9$ bases per replication round
- Due to DNA replication mistakes and DNA rearrangements, genomes in different individuals become different
- Mutations (Selection: negative or positive, or neutral)
- Genetic variation is source of Evolution (e.g. Speciation)
A gene codes for a protein

Transcription + Translation = Expression
DNA makes mRNA makes Protein

Translation happens within the ribosome

transcription + translation = expression
Gene expression is depending on a Transcription Factor (TF) binding a Transcription Factor Binding Site (TFBS) and a polymerase (Pol II in eukaryotes).

(TF and polymerase are proteins)
Figure 6 Conservation in the GAL1–GAL10 intergenic region. Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation. Asterisks denote conserved positions in the multiple alignment. Blue arrows denote the start and transcriptional orientation of the flanking ORFs. Experimentally validated factor-binding footprints are boxed and labelled according to the bound factor. Stretches of conserved nucleotides are underlined. Nucleotides matching the published Gal4 motif are shown in red. The fourth experimentally validated site differs: it shows a longer footprint and a non-standard consensus motif (bold). This variant motif is also conserved across all four species. Scer, S. cerevisiae; Spar, S. paradoxus; Smik, S. mikatae; Sbay, S. bayanus.
Gene Expression

- Transcription factors (TF) are essential for transcription initialisation (‘genes need TFs’)
- TFs bind to DNA-motifs called TF binding sites
- Transcription is done by polymerase type II (eukaryotes)
- mRNA must then move from nucleus to ribosomes (extranuclear) for translation
- In eukaryotes there can be many TF-binding sites upstream of an ORF (Open Reading Frame) which together regulate transcription
- Nucleosomes (chromatin structures composed of histones) are structures round of which DNA coils. This blocks access of TFs
Bacterial transcription initiation

Schematic representation of elements involved in bacterial transcription initiation. RNA polymerase binds to the promoter region, which initiates transcription through interaction with transcription factors binding at different sites. **Abbreviations:** TSS, transcription start site; ORF, reading frame; pol, polymerase; TF, transcription factor; TF site or TFBS, transcription factor binding site.
Three examples of DNA binding protein families

<table>
<thead>
<tr>
<th>Name</th>
<th>Helix-loop-helix (Myc type)</th>
<th>Cys-His zinc finger</th>
<th>Leucine zipper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>[DENSTAP]-K-[LIVMWAGN]-[FYWCPHKR]-[LIVT]-[LIV]-x(2)-[STAV]-[LIVMSTAC]-x-[VMFYH]-[LIVMTA]-{P}-{P}-{LIVMSR}</td>
<td>C-x(2,4)-C-x(3)-[LIVMFYWVC]-x(8)-H-x(3,5)-H</td>
<td>L-x(6)-L-x(6)-L-x(6)-L</td>
</tr>
<tr>
<td>Structure</td>
<td><img src="image1" alt="Helix-loop-helix" /></td>
<td><img src="image2" alt="Cys-His zinc finger" /></td>
<td><img src="image3" alt="Leucine zipper" /></td>
</tr>
<tr>
<td>Function</td>
<td>DNA Binding</td>
<td>DNA Binding</td>
<td>DNA Binding</td>
</tr>
<tr>
<td>Example</td>
<td><img src="image1" alt="Example1" /></td>
<td><img src="image2" alt="Example2" /></td>
<td><img src="image3" alt="Example3" /></td>
</tr>
</tbody>
</table>

**Fig. 2.14.** A motif dictionary of DNA binding proteins, showing the relationships between sequence motifs, structural motifs, and functional properties.
434 Cro protein complex (phage)
PDB: 3CRO
Zinc finger DNA recognition (Drosophila) PDB: 2DRP

..YRCVKVC$R$VY THISNFCRHY VTSH...

Co-ordinate (interact with) the zinc ion (white spheres in figure)
Leucine zipper  
(yeast)  
PDB: 1YSA

..RARKLQRMKQVLEDKVEELLSKNYHELNEVARL...
Eukaryotic transcription initiation

Schematic diagram of an eukaryotic promoter with transcription factors and RNA polymerase bound to the promoter.

Abbreviations: Inr, initiator sequence; ORF, reading frame; pol, polymerase; TF, transcription factor; TF site or TFBS, transcription factor binding site
Transcription factor interaction can be complex

Hematopoietic (blood) gene regulatory network consisting of 11 interacting transcription factors

The currently known blood stem cell regulatory network model based on comprehensive cis-regulatory information. Blue arrows: activation, Red arrows: inhibition
DNA is packaged and protected

- DNA winds around histone proteins (nucleosomes).
- Other proteins wind DNA into more tightly packed form, the chromosome.
- Unwinding portions of the chromosome is important for mitosis, replication and making RNA.
Transcription Factors (eukaryotes)

Transcription Factor (TF): protein that binds to DNA and to a polymerase (Pol II)

Polymerase: complex protein that transcribes DNA into mRNA

Transcription factor – polymerase interaction sets off gene transcription…

Nucleosomes (chromatin structures composed of histones) are structures round which DNA coils. This blocks access of TFs

… many TFBSs are possible upstream of a gene
Translation: from mRNA to protein

- Translation from 4 nucleotides in (DNA and) RNA to 20 different amino acids in proteins.

- Translation uses **codons**, i.e. groups of three nucleotides. This gives $4^3 = 64$ possibilities.

- The 64 codons encode 20 amino acid types, a start codon and stop codons.

- The encoding is given by a so-called codon table, linking each possible codon to its translated product.

- The codon table is redundant, i.e. different codons can encode the same amino acid (Trp has 1 codon, Leu has 6 codons, there are 3 stop codons).

- Translation involves mRNA (template), tRNA (amino acid carrier) and ribosome (protein generating giant enzyme).
Encoding proteins – the translation process uses the codon table

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td></td>
<td>UUU</td>
<td>UCU</td>
<td>UAU</td>
<td>UGU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UUC</td>
<td>UCC</td>
<td>UAC</td>
<td>UGC</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Leu</td>
<td>Ser</td>
<td>Cys</td>
<td>Stop</td>
</tr>
<tr>
<td>U</td>
<td></td>
<td>UUU</td>
<td>UCU</td>
<td>UAU</td>
<td>UGU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UUC</td>
<td>UCC</td>
<td>UAC</td>
<td>UGC</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>CUU</td>
<td>CCU</td>
<td>CAU</td>
<td>CGU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUC</td>
<td>CCC</td>
<td>CAC</td>
<td>CGC</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>AUU</td>
<td>ACU</td>
<td>AAU</td>
<td>AGU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUC</td>
<td>ACC</td>
<td>AAC</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>AUG</td>
<td></td>
<td>AUA</td>
<td>ACA</td>
<td>AAA</td>
<td>AGA</td>
</tr>
<tr>
<td></td>
<td>Met</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>GUU</td>
<td>GCU</td>
<td>GAU</td>
<td>GGU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUC</td>
<td>GCC</td>
<td>GAC</td>
<td>GGC</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>GUA</td>
<td>GCA</td>
<td>GAA</td>
<td>GGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUG</td>
<td>GCG</td>
<td>GAG</td>
<td>GGG</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
<td></td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Single Letter Code</td>
<td>DNA codons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>ATT, ATC, ATA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>CTT, CTC, CTA, CTG, TTA, TTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>GTT, GTC, GTA, GTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>TTT, TTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methionine</td>
<td>M,</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Start</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>TGT, TGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>GCT, GCC, GCA, GCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>G</td>
<td>GGT, GGC, GGA, GGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>CCT, CCC, CCA, CCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>ACT, ACC, ACA, ACG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>TCT, TCC, TCA, TCG, AGT, AGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>TAT, TAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>TGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>CAA, CAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>AAT, AAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>CAT, CAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>E</td>
<td>GAA, GAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>D</td>
<td>GAT, GAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>AAA, AAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>CGT, CGC, CGA, CGG, AGA, AGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stop codons</strong></td>
<td><strong>Stop</strong></td>
<td>TAA, TAG, TGA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Frequency of Amino Acids in the Swiss-Prot database

6. AMINO ACID COMPOSITION
6.1 Composition in percent for the complete database

Ala (A) 8.28  Gln (Q) 3.94  Leu (L) 9.67  Ser (S) 6.49
Arg (R) 5.54  Glu (E) 6.77  Lys (K) 5.86  Thr (T) 5.32
Asn (N) 4.05  Gly (G) 7.09  Met (M) 2.43  Trp (W) 1.07
Asp (D) 5.45  His (H) 2.27  Phe (F) 3.86  Tyr (Y) 2.91
Cys (C) 1.36  Ile (I) 5.99  Pro (P) 4.68  Val (V) 6.88

Release 57.12 of 15-Dec-09 of UniProtKB/Swiss-Prot contains 513877 sequence entries, comprising 180750753 amino acids abstracted from 185524 references.
An artist’s impression of translation
3D structures of RNA: transfer-RNA structures

- Secondary structure of tRNA (cloverleaf)
- Tertiary structure of tRNA
Prokaryote gene

- Prokaryotes, which include bacteria and Archaea, have relatively small genomes with sizes ranging from 0.5 to 10Mbp (1Mbp=10^6 bp).

- The gene density in the genomes is high, with more than 90% of a genome sequence containing coding sequence.

- There are very few repetitive sequences. Each prokaryotic gene is composed of a single contiguous stretch of open reading frame (ORF) coding for a single protein or RNA with no interruptions within a gene (no splicing).
The majority of genes have a start codon ATG (or AUG in mRNA) coding for methionine. Occasionally, GTG and TTG are used as alternative start codons, but methionine is still the actual amino acid inserted at the first position.

Because there may be multiple ATG, GTG, or TGT codons in a frame, the presence of these codons at the beginning of the frame does not necessarily give a clear indication of the translation initiation site.

To help gene identification, other features associated with translation are used: One such feature is the **ribosomal binding site (RBS)**, also called the **Shine-Delgarno sequence**, a stretch of purine-rich sequence complementary to 16S rRNA in the ribosome. It is located immediately downstream of the transcription initiation site and slightly upstream of the translation start codon. In many bacteria, it has a consensus motif of AGGAGGT. Identification of the ribosome binding site can help locate the start codon.
Prokaryote gene

- At the end of the protein coding region is a stop codon that causes translation to stop. There are three possible stop codons, identification of which is straightforward.
- Many prokaryotic genes are transcribed together as one operon. The end of the operon is characterized by a transcription termination signal called \( \rho \)-independent terminator. The terminator sequence has a distinct stem-loop secondary structure followed by a string of Ts. Identification of the terminator site, in conjunction with promoter site identification, can sometimes help in gene prediction.
A piece of double stranded DNA:

DNA direction is from 5’ to 3’
So, there are six possibilities to make a protein from an unknown piece of DNA, only one of which might be a natural protein.

Frameshifts can occur (e.g. by deletion of 1 or 2 mRNA bases) leading to greatly altered protein sequences.
(Protein coding) gene prediction

- **ab-initio based approaches**
  - Predicts genes based on the given sequence alone.
  - Relies on two main features:
    - Gene signals (start & stop codons, intron splice signals, TFBS’s, ribosomal binding sites, poly-A sites)
    - Gene content (statistical description of gene content which is very different from non-coding regions)

- **homology based approaches**
  - Predictions are based on significant matches of the query sequence with sequences of known genes (db searching)

- **consensus approaches**

→ Helps us to estimate the number of genes (think about the estimated number of human genes: estimates have gone down from 200,000 → 20,000)
Gene prediction in prokaryotes...

ATG

Translation start

Coding region

Stop

Transcription terminator

RBS

AGGAGGT

Shine-Delgarno sequence

**Figure 8.1:** Structure of a typical prokaryotic gene structure. *Abbreviation:* RBS, ribosome binding site.

**Figure 8.2:** Coding frame detection of a bacterial gene using either the GC bias or the TESTCODE method. Both result in similar identification of a reading frame (*dashed arrows*).
Prokaryote gene prediction
how to predict an ORF by hand

- Perform conceptual translation in all six possible frames, three frames forward and three frames reverse. Because a stop codon occurs in about every twenty codons by chance in a noncoding region, a frame longer than 30 codons without interruption by stop codons is suggestive of a gene coding region (threshold is normally set even higher at 50 or 60 codons).
Identifying (annotating) human genes, i.e. finding what they are and what they do, is a difficult problem.

- First, the gene should be delineated on the genome

  - Gene finding methods should be able to tell a gene region from a non-gene region
  - Start, stop codons, further compositional differences

- Then, a putative function should be found for the gene located
Eukaryotes have spliced genes ...

- **Promoter**: involved in transcription initiation (TF/RNAPol-binding sites)
- **TSS**: transcription start site
- **UTRs**: un-translated regions (important for translational control)
- **Exons** will be spliced together by removal of the **Introns**
- **Poly-adenylation site** important for transcription termination (but also: mRNA stability, export mRNA from nucleus etc.)
DNA makes mRNA makes Protein
DNA makes RNA makes Protein

... yet another picture to appreciate the above statement
Eukaryotic mRNA

Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as mature RNA after posttranscriptional processing. 

*Abbreviations: UTR, untranslated region; poly-A, polyadenylation.*
Eukaryotic mRNA

The nascent transcript from a eukaryotic gene is modified in three different ways before becoming a mature mRNA for protein translation.

1. The first is **capping** at the 5’ end of the transcript, which involves methylation at the initial residue of the RNA.

2. The second modification is **polyadenylation**, which is the addition of a stretch of As (~250) at the 3’ end of the RNA. This process is controlled by a poly-A signal, a conserved motif slightly downstream of a coding region with a consensus CAATAAA(T/C).
3. The third event is **splicing**, which is the process of removing introns and joining exons.
   - It involves a large RNA-protein complex called **spliceosome**.
   - The reaction requires intermolecular interactions between a pair of nucleotides at each end of an intron and the RNA component of the spliceosome.
   - Some eukaryotic genes can have their transcripts spliced and joined in different ways to generate more than one transcript per gene. This is the phenomenon of **alternative splicing**, which is a major mechanism for generating functional diversity in eukaryotic cells.
Most introns start from the sequence **GU** and end with the sequence **AG** (in the 5' to 3' direction). They are referred to as the **splice donor** and **splice acceptor** site, respectively. However, these are not sufficient to signal the presence of an intron. Another important sequence is called the **branch site** located 20 - 50 bases upstream of the acceptor site. The consensus sequence of the branch site is "CU(A/G)A(C/U)", where A is conserved in all genes.

In over 60% of cases, the exon sequence is (A/C)AG at the donor site, and G at the acceptor site.

![Schematic overview](image)
The detailed splicing mechanism is quite complex, involving five snRNAs (small nuclear RNAs) and their associated proteins. These ribonucleoproteins form a large (60S) complex, called **spliceosome**. Then, after a two-step enzymatic reaction, the intron is removed and two neighboring exons are joined together. The branch point A residue plays a critical role in the enzymatic reaction.
Alternative splicing

A single gene can be transcribed into many different mRNAs, leading to different proteins.

Schematic cutoff from three splicing structures in the murine hyaluronidase gene. Directionality of transcription from 5' to 3' is shown from left to right. Exons and introns are not drawn to scale.
Alternative splicing provides a first idea of the problems we can expect when aligning protein sequences.

For example, if we need to compare different splice forms of divergent proteins in multiple organisms.
Wrapping up

- DNA as an information carrying molecule
- DNA makes RNA makes protein
  - The molecular machinery as an information device
  - Information on DNA around (upstream and downstream) the protein coding region (all together called ‘gene’)
  - How much can DNA itself influence the encoding processes?
  - Replication, transcription, expression, polymerase, epi-genomics (nucleosomes), TF, TFBS, translation, mRNA, tRNA, ribosome, 6-frame translation, frameshift, 3rd codon position and GC richness, splicing, pre-mRNA, mature mRNA, spliceosome, splice donor/acceptor sites, alternative splicing
    - What informational aspects are playing a role in all of these terms or processes?