COSC 348:
Computing for Bioinformatics

Lecture 3: Introduction concluded

Lubica Benuskova

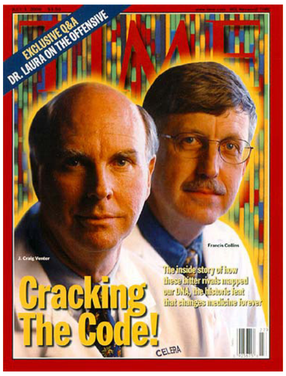
http://www.cs.otago.ac.nz/cosc348/

Genome projects

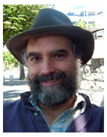
- **Genome sequencing** means determining the order of bases in DNA. There are about 6,600 completed and 30,000 projects according to Genomes OnLine Database (http://www.genomesonline.org).
- In a **shotgun sequencing** project, all the DNA from an organism is first fractured into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read 1000 bases at a time.
- A **genome assembly algorithm** works by taking all the pieces of DNA and aligning them to one another, and detecting all places where two of the short sequences, or reads, overlap. These overlapping reads can be merged, and the process continues until the whole DNA is assembled.

Human genome project (HGP): 1990-2003

- Publicly funded project was initially headed by James D. Watson. He was replaced by Francis Collins in 1993. (http://www.genome.gov/)
- In parallel, sequencing was performed at a privately funded Celera Genomics Corporation led by J. Craig Venter.
- Most of the sequencing was performed at universities and research centers in the USA, U.K., Japan, France, Germany, China, India, Canada, and New Zealand.


TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
Cracking The Code!
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
CELERA

Role of computer science in HGP



- Computer scientist and a PhD student in Biology at the University of California, Santa Cruz, Jim Kent in May 2000, wrote a program that allowed the publicly funded HGP to assemble and publish the human genome database (~ 3 billion of letters).
- Close race: Kent's first assembly of the human genome was released on 22 June. Celera finished its assembly on 25 June, and the dual results were announced at the White House on June 26. On July 7, the human genome data were made publicly available on the WWW and everyone can use them.
- His efforts were motivated also out of concern that the data might be made proprietary via patents by Celera Genomics.

Questions

- How many bases are there in the whole DNA (genome) for different organisms?
- What is the order (sequence) of these bases in all species?
  - It is estimated there could be anywhere from 5 million to 100 million species on the planet, but science has only identified about 2 million…
  - We know genomes of about 2500 species of organisms
- What is the *gene content* of the known genomes?
- What do all these genes do?

How many genes and what are they for?

Species	Number of genes
Mycoplasma genitalium (bacterium)	500
Streptococcus pneumoniae (bacterium)	2,300
Escherichia coli (bacterium)	4,400
Saccharomyces cerevisiae (yeast)	5,800
Drosophila melanogaster (fruit fly)	13,700
Caenorhabditis elegans (roundworm)	19,000
***Homo sapiens* (human)**	**20,500**
Sea urchin	23,300
Arabidopsis thaliana (plant)	25,500
Mus musculus (mouse)	29,000
Oryza sativa (rice)	50,000

Source: Watson JD, et al (2004). *Molecular Biology of the Gene*, 5th ed., Pearson Benjamin Cummings (Cold Spring Harbor Laboratory Press).

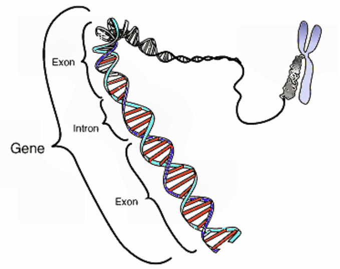
The study found humans have less than twice the number of genes as a fruit fly. Of course, the data could be off a bit. It was done by researchers who had only less than twice the number of genes as a fruit fly...

Gene density and complexity of organisms

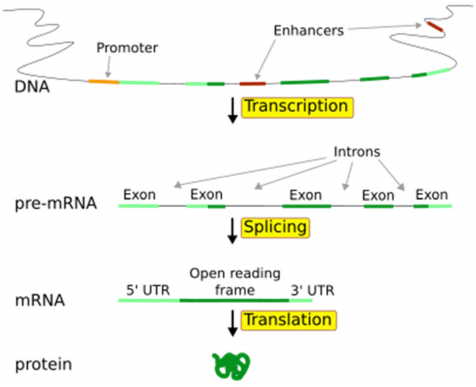
- Current estimates place the human genome at just under 3 billion base pairs and about 20,000–25,000 genes.
- The gene density of a genome is a measure of the number of genes per *million base pairs* (called a *megabase*, Mb).
- Prokaryotic genomes (e.g. in bacteria) have *much higher* gene densities than eukaryotes (e.g., in animals).
- The gene density of the human genome is only 12–15 genes/Mb.
- However, the density or number of genes are not a good measure of organism's complexity, because things are more complicated…

Exons and introns

- In eukaryotic cells (including us) genes contain 2 types of regions
  - *Exons:* the regions *encoding* proteins
  - *Introns:* Regions that do not encode proteins.
- Introns are removed from the messenger RNA in a process called **splicing**.

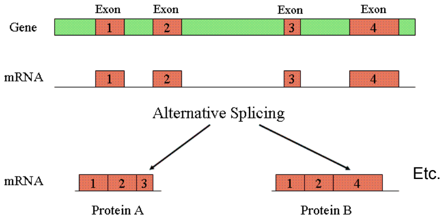


Splicing: exons are combined into Open Reading Frame (ORF)



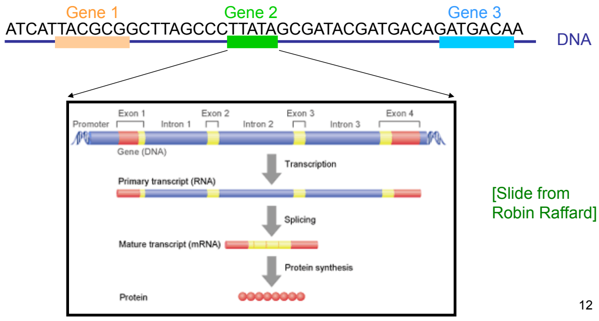
Alternative splicing

- A **single gene** can encode **multiple proteins**, which are produced through the creation of different arrangements of exons through *alternative splicing*.
- The different forms of the resulting mRNA are called *transcript variants*, *splice variants*, or *isoforms.*



Coding and non-coding DNA

DNA consists of genes (protein coding sequences) separated by non-coding sequences regions.

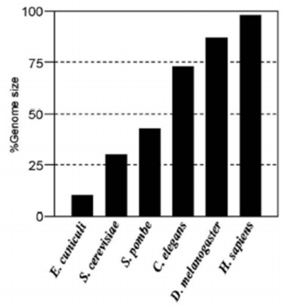


[Slide from Robin Raffard]

Proportion of “junk” or non-coding DNA

- Protein-coding DNA makes up barely 2% of the human genome!
- The rest 98% is the non-coding DNA.
- What is the role of non-coding DNA?

% of “junk” DNA in different species



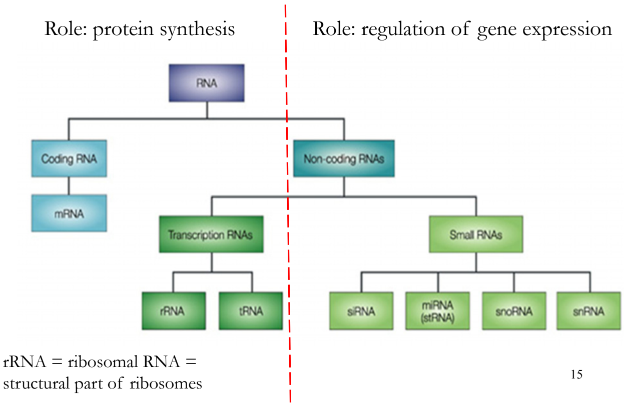
What is the role of non-coding DNA?

- The 98% of DNA has many important functions like
  - *Retrotransposons* (42% of DNA), move in the genome by being transcribed to RNA and then back to DNA by reverse transcriptase; (can drive evolution by causing mutations)
  - Other *transposones*, mobile DNA pieces, “jumping genes”; (can cause mutations, too)
  - Non-functional remains of ancient genes, known as *pseudogenes*; (they were needed in the past and may be needed in the future)
  - DNA producing different types of RNAs that have important regulatory functions upon gene expression and which can regulate which protein will be produced by alternative splicing.
  - Reservoir of sequences from which potentially advantageous new genes can emerge?

Variety of RNAs

Role: protein synthesis

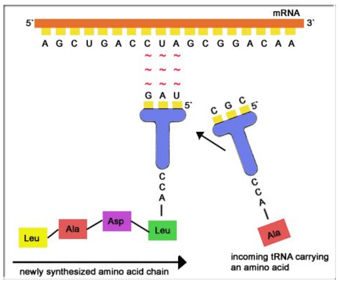
Role: regulation of gene expression



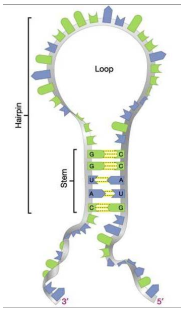
rRNA = ribosomal RNA = structural part of ribosomes

Transfer RNA (tRNA)

- There are as many tRNAs as there are codons.
- Each RNA triplet carries one amino acid.
- tRNAS attach to mRNA according to corresponding letter triplets and when the AAs get into the vicinity of each other, they form a chemical bond.


mRNA
5'
3'
A G C U G A C C U A G C G G A C A A
G A U
5'
C G C
5'
C
C
A
C
C
A
Leu
Ala
Asp
Leu
Ala
newly synthesized amino acid chain
incoming tRNA carrying an amino acid

Small regulatory RNAs

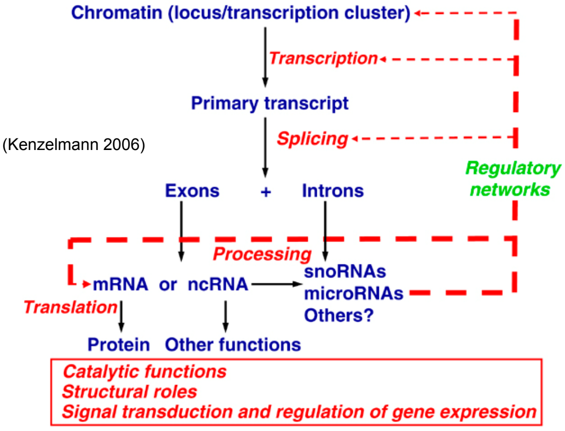


- siRNA = small interfering RNA (silencing RNA)
  - interferes with the expression of genes
- miRNA = microRNA
  - target mRNAs, usually resulting in gene silencing
- snoRNA = small nucleolar RNA
  - guide chemical modifications of ribosomal RNAs (rRNAs) and other RNA genes (tRNA and snRNAs)
- snRNA = small nuclear RNA
  - involved in a variety of important processes such as RNA splicing

Central dogma revisited

- Crick 1958: “DNA makes RNA makes protein, and proteins make us.”
- Petsko 2000: “DNA makes RNA and RNA makes protein, but sometimes RNA makes DNA and other times RNA makes RNA, which makes proteins different from what they would be if only DNA made the RNA, and once upon a time RNA made protein, probably, but no-one knows for certain.”
- Mattick 2006: “Things are even more complicated & regulatory RNAs represent the major output of the genomes of humans.”

Revised definition of gene and flow of genetic information



Chromatin (locus/transcription cluster)

Transcription

Primary transcript

(Kenzelmann 2006)

Splicing

Regulatory networks

Exons + Introns

Processing

mRNA or ncRNA

snoRNAs
microRNAs
Others?

Translation

Protein Other functions

Catalytic functions
Structural roles
Signal transduction and regulation of gene expression

Regulation of gene transcription: basic concepts

- **Transcription factor** (TF) is a protein or RNA that binds to specific DNA domains and controls gene transcription into RNA.

- An **activator** is a TF that increases gene transcription, thus leading to the gene upregulation (activation, promotion). The activator binds to a DNA segment known as *enhancer*. Activator may increase transcription by itself, or may operate through one or more **co-activators**.

- A **repressor** is a TF that decreases gene transcription, thus leading to the gene downregulation (repression, suppression). Repressor proteins attach to a DNA segment known as the *operator*. If an inducer, a molecule that initiates gene expression, is present, then it can interact with the repressor protein and detach it from the operator. There exist also **co-repressor** TFs.

- Inducers function by disabling repressor proteins. Some inducers are modulated by activators.

Regulation of transcription: what is it good for?

- **Basal transcription regulation**. General transcription factors (GTFs) are necessary for transcription to occur.
  - The most common GTFs are TFIIA(B, D, E, F, and H).

- **Cell cycle control.** Many TFs (especially oncogene and tumor suppressor proteins) help regulate cell differentiation, division and apoptosis (programmed cell death).
  - Examples of proto-oncogenes: RAS, WNT, MYC, ERK and TRK.

- **Development.** In response to intra- and extracelluar stimuli, TFs turn on/off the transcription of the appropriate genes, which in turn allow for changes in cell morphology, differentiation and function.
  - Examples: the Hox TF family (body pattern formation).

Regulation of transcription: what is it good for?

- **Response to intercellular signals**. Cells can communicate with each other by releasing molecules that produce signalling molecular chain cascades within other cells. External molecular signals activate TFs in the target cell.
  - Example: formation of organs in the body.

- **Response to environment**. Not only do transcription factors act downstream of signaling cascades related to biological stimuli, but they can also be downstream of signaling cascades involved in environmental stimuli.
  - Examples: heat shock factor (HSF) which upregulates genes necessary for survival at higher temperatures.

Towards artificial life: the synthetic genomics

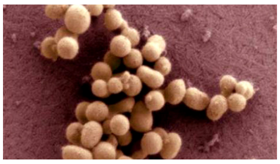
- "Craig Venter creates synthetic life form" says the news headline from 20 may 2010.

- Venter's team led by Daniel Gibson implanted a synthetic copy of the 1-million base genome of *M. mycoides* into *M. capricolum*.

- *M. capricolum* started to make proteins like *M. mycoides* and divide.

- Synthetic genome contained true genes and also strings of bases that coded the names of people involved in the project and some famous quotes.

The project cost $40m and took 20 scientists over 10 years of work…



Towards artificial life: the synthetic genomics

- The deed of scientists from the J. Craig Venter Institute (JCVI) is "a defining moment in the history of biology and biotechnology"

- Synthetic Biology is (http://syntheticbiology.org/)
  - A) the design and construction of new biological parts, devices, and systems, and
  - B) the re-design of existing, natural biological systems for useful purposes.

- Scientists are focused on developing synthetic organisms able to produce various kinds of biological products, including medicines, renewable fuels and agents that remove harmful waste from nature.

- Synthetic genomics holds great promise for the future as well as unknown dangers…