

Beyond Genomics:

Detecting Codes and Signals
in the Cellular Transcriptome

Brendan J. Frey

University of Toronto

Purpose of my talk

To identify **aspects of bioinformatics** in which **attendees of ISIT** may be able to make **significant contributions**

Beyond Genomics:

Detecting Codes and Signals
in the Cellular Transcriptome

Brendan J. Frey

University of Toronto

The Genome

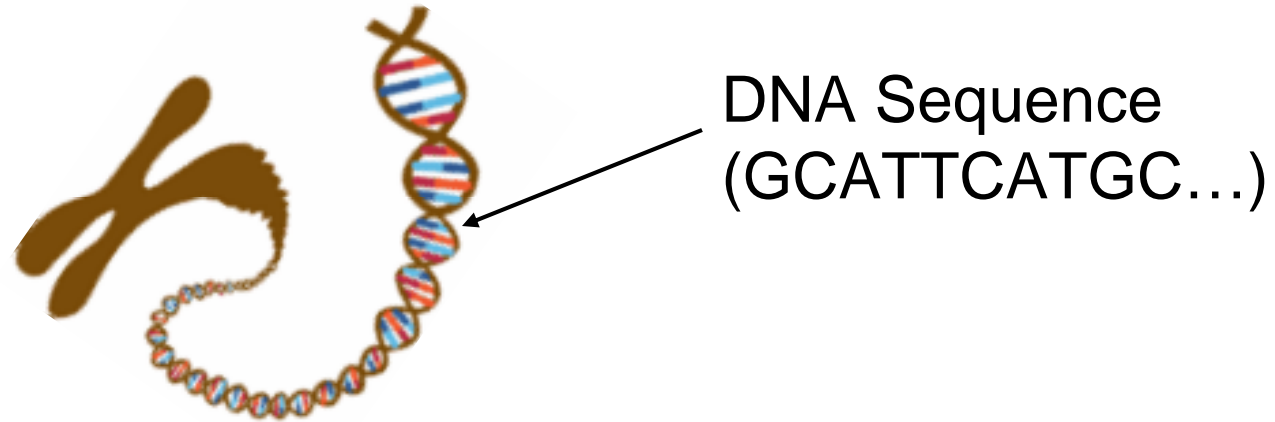
Starting point: Discrete biological sequences

- Symbols are Bases: G, C, A, T

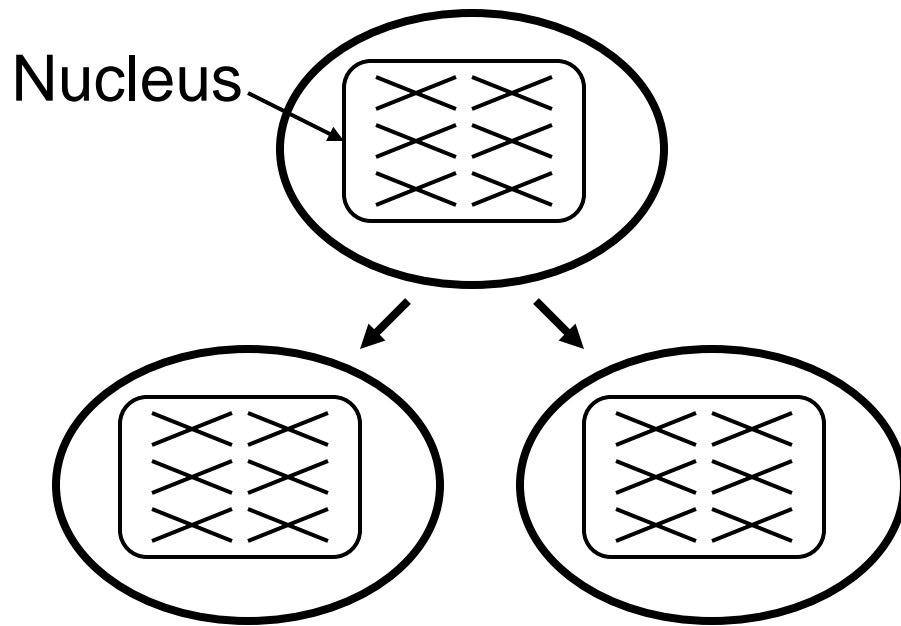
RED indicates a definition
that you should remember

- Examples of biological sequences
 - Genes
 - DNA
 - Chromosomes
 - Proteins
 - Peptides
 - RNA
 - Viruses
 - HIV

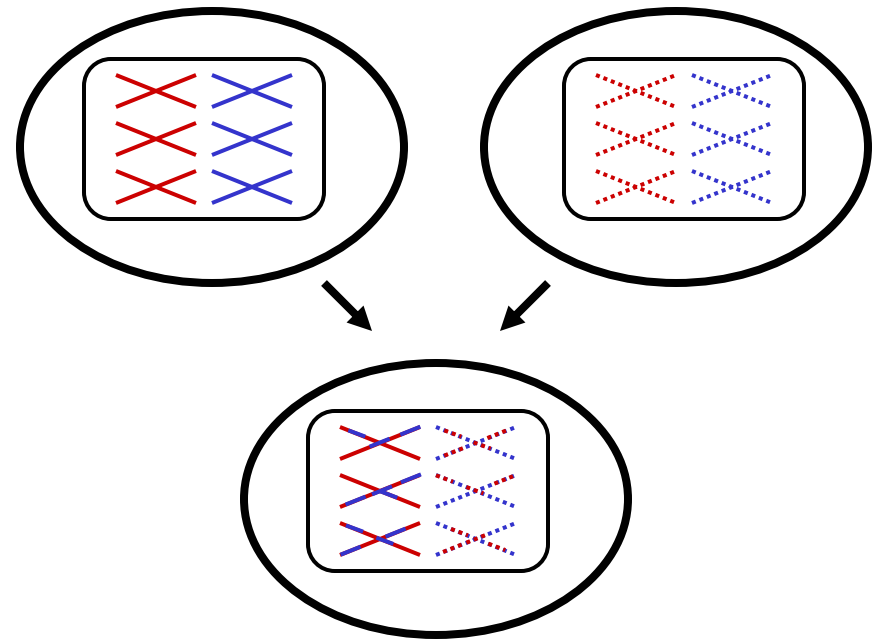
Chromosomes: Inherited DNA sequence



Cell replication



Sexual cell reproduction



The genome

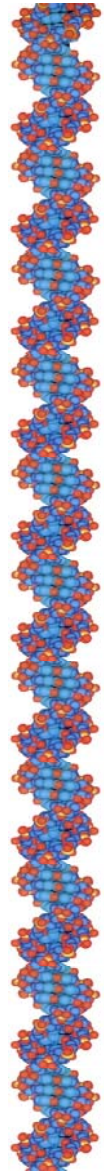
- Genome: Chromosomal DNA sequence from an organism or species
- Examples

<u>Genome</u>	<u>Length (bases)</u>
Human	3,000 million (750MB)
Mouse	2,600 million
Fly	100 million
Yeast	13 million

Genes

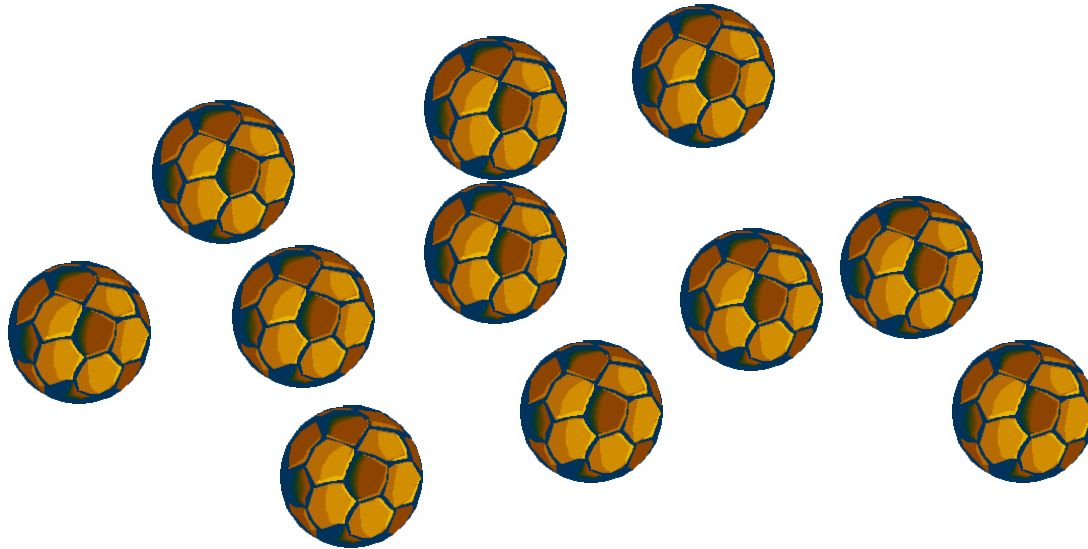
- A gene is a subsequence of the genome that encodes a functioning bio-molecule
- The library of known genes
 - Comprises only 1% of genome sequence
 - Increases in diversity every year
 - Is probably far from complete

The Transcriptome

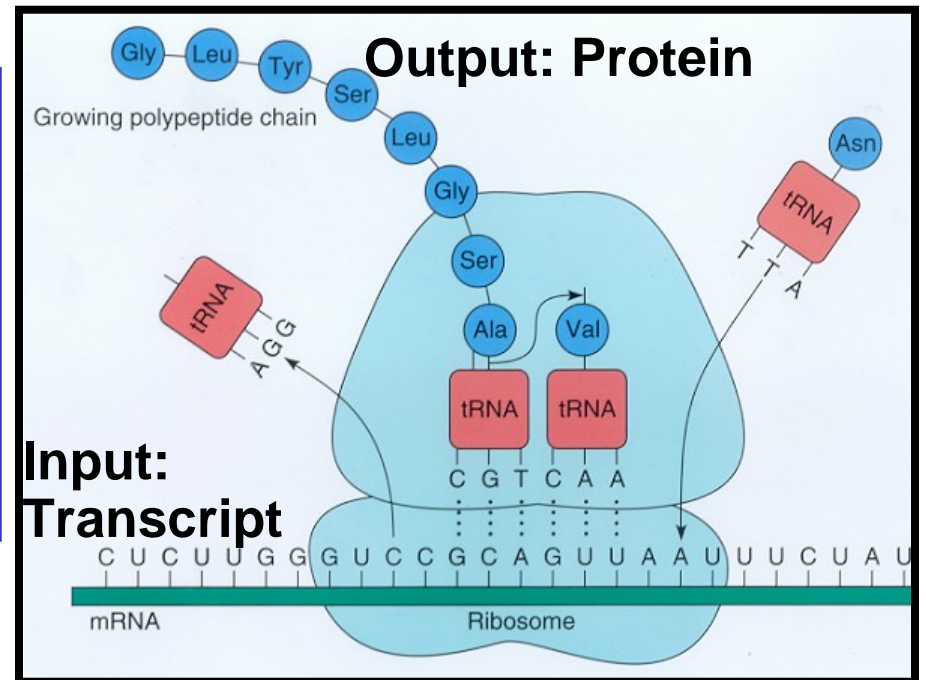
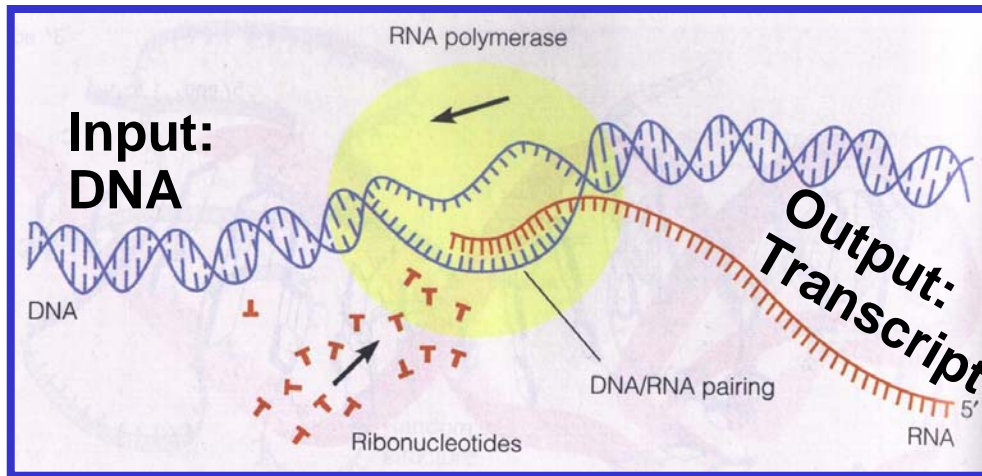
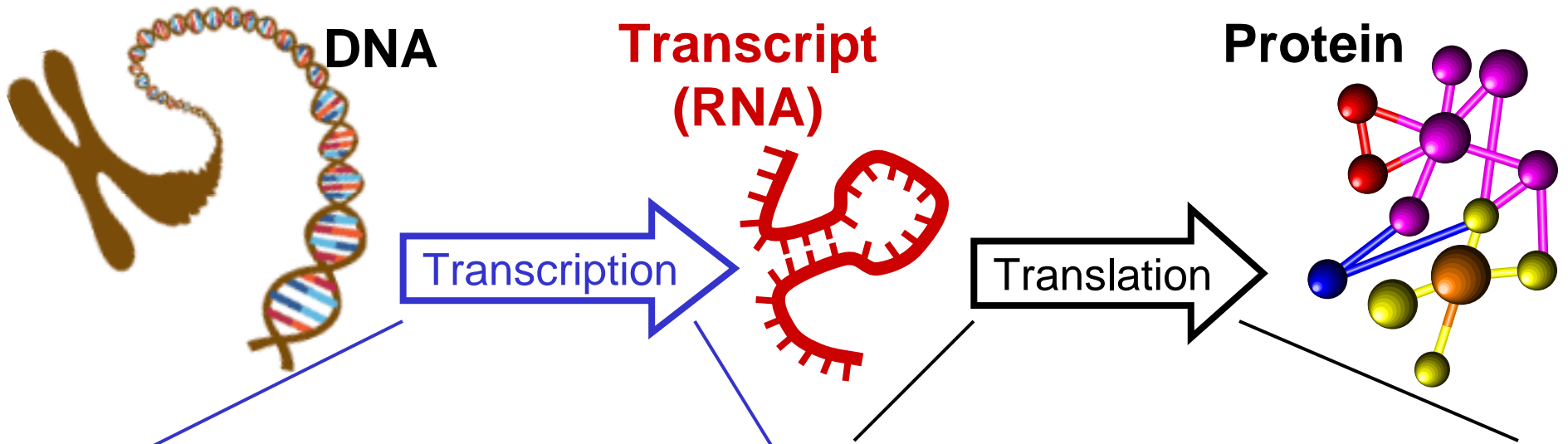


Genome: The digital backbone
of molecular biology

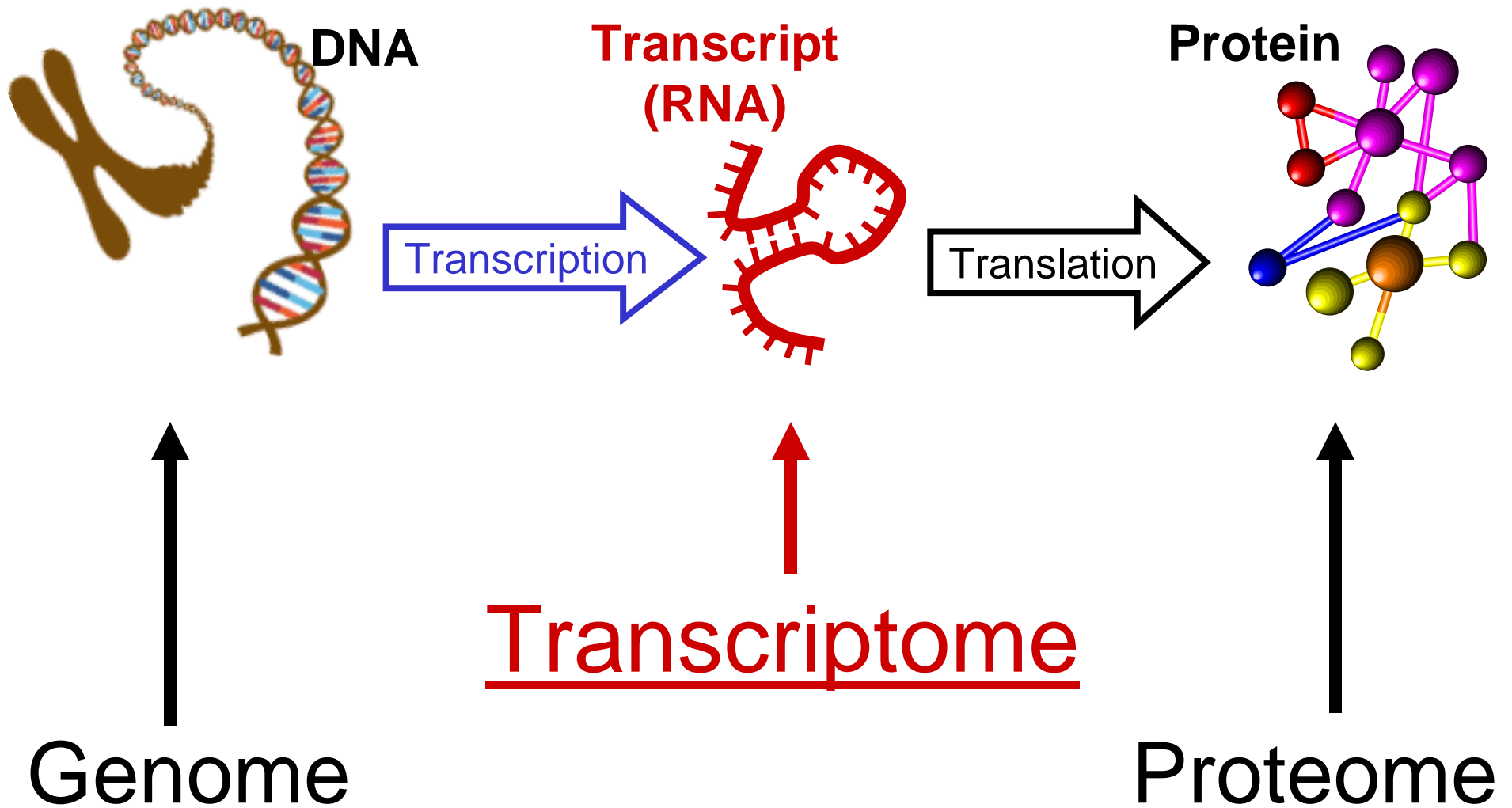
Transcripts: Perform functions
encoded in the genome



Traditional genes

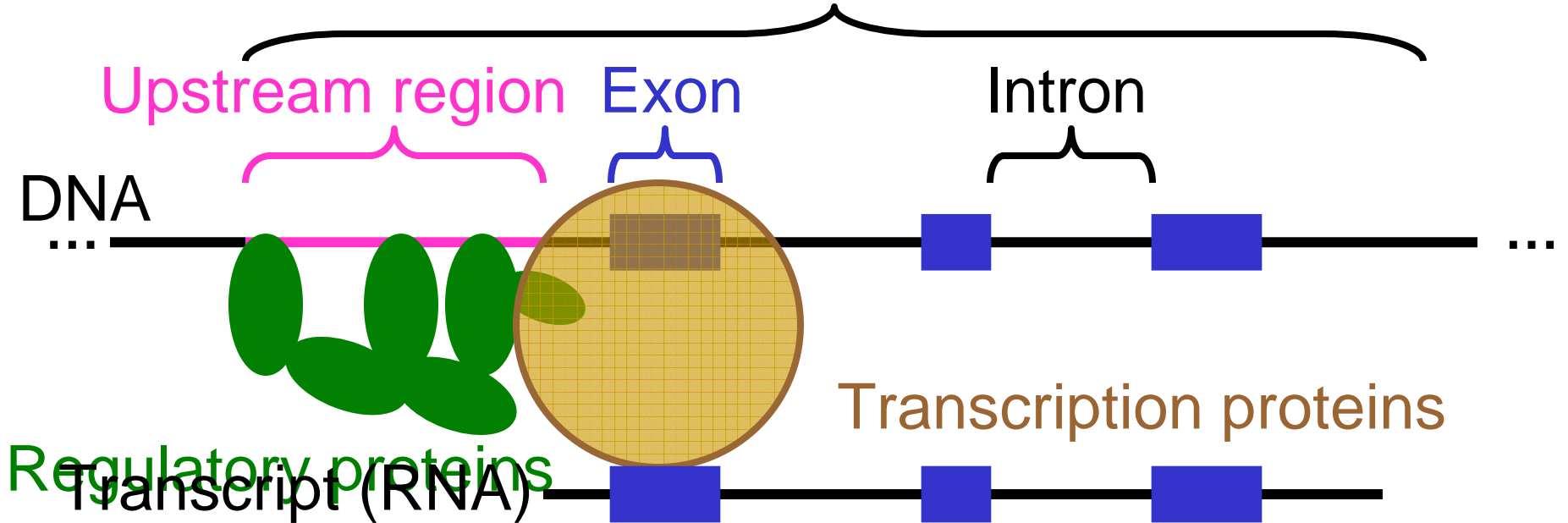


Traditional genes



Transcription

Gene

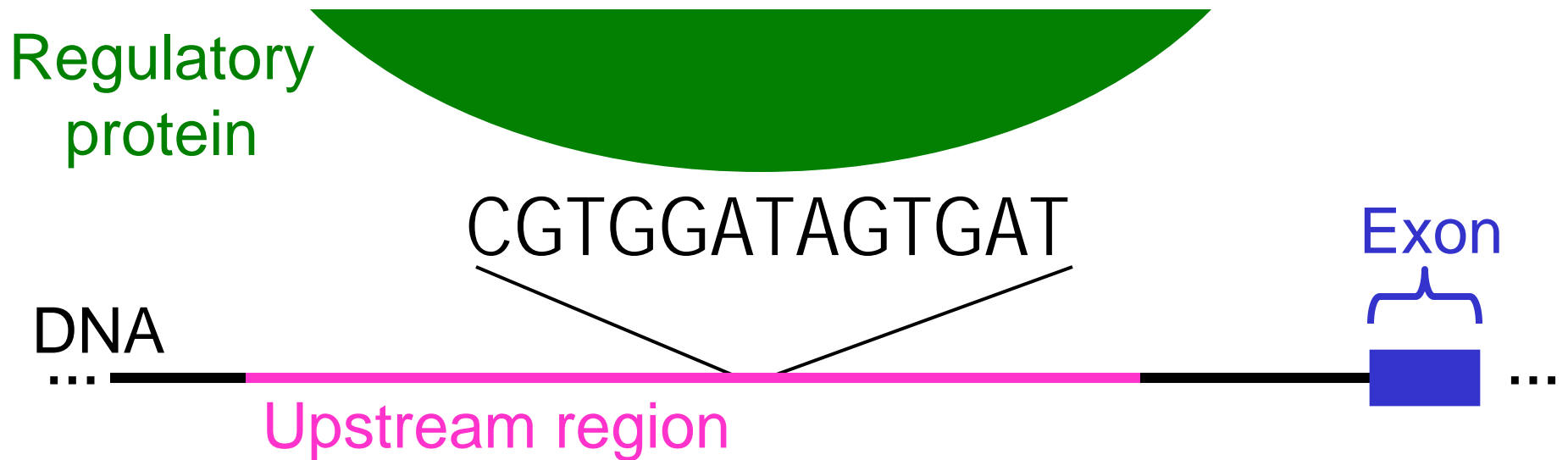


Transcription



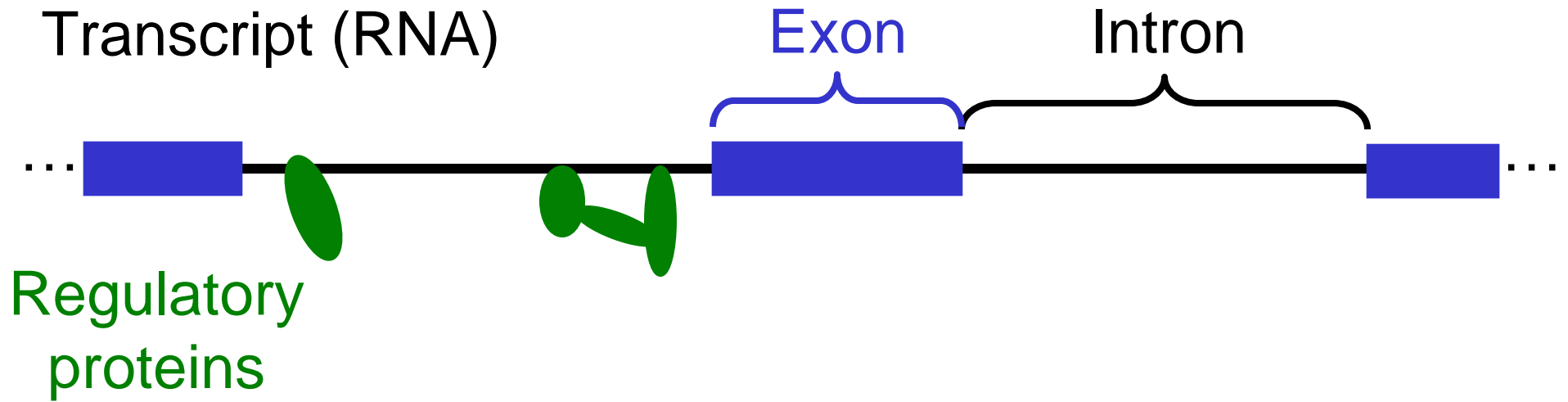
Transcription

- Codewords in the **upstream region** bind to corresponding **regulatory proteins**

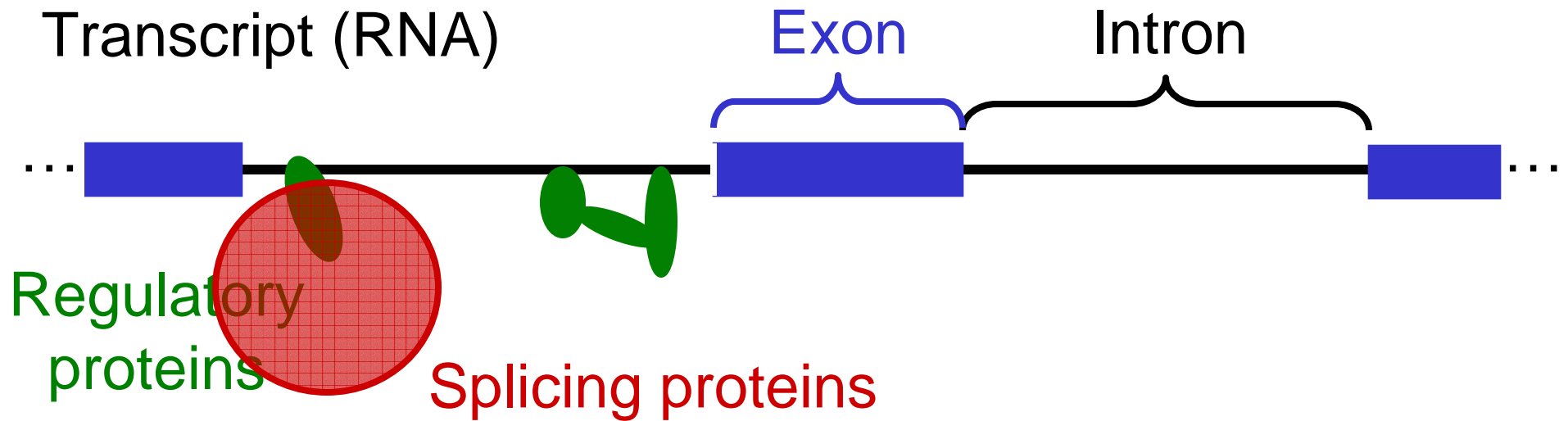


- Code: Set of regulatory codewords
- Signals: Concentrations of regulatory proteins and the output transcript

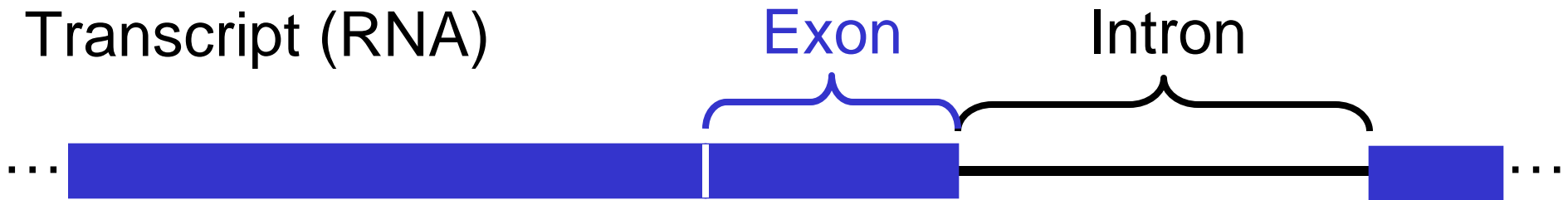
Splicing of transcripts



Splicing of transcripts

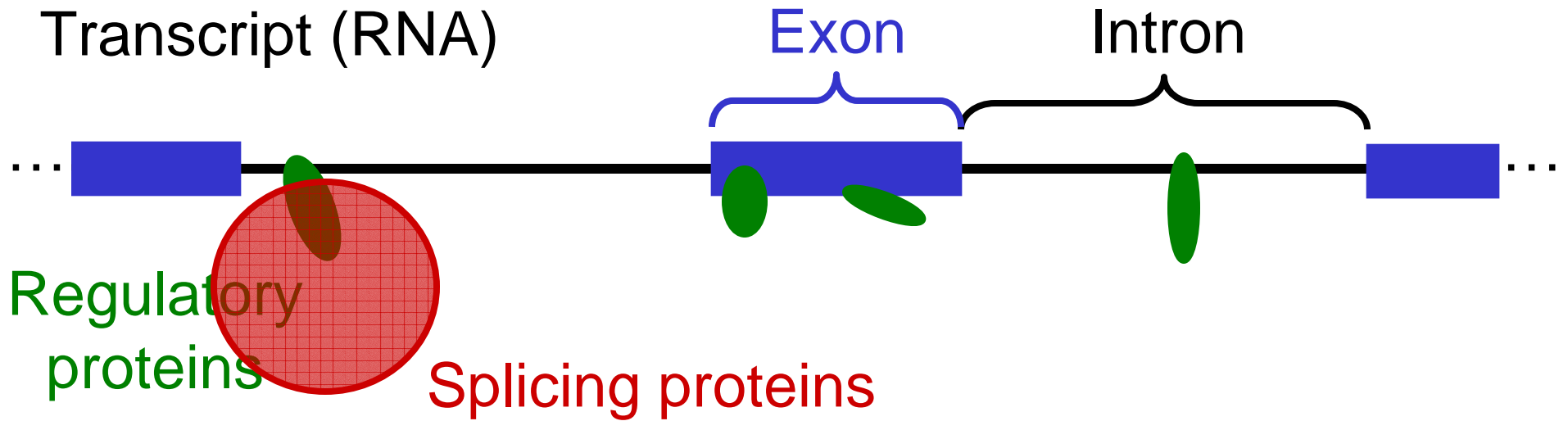


Splicing of transcripts

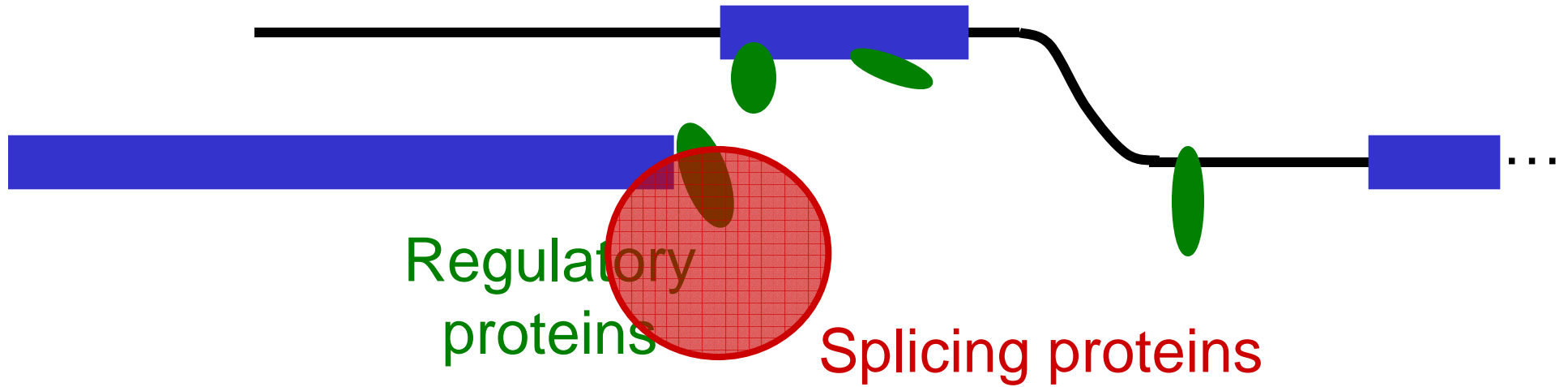


- The intron is spliced out
- However, splicing may occur quite differently...

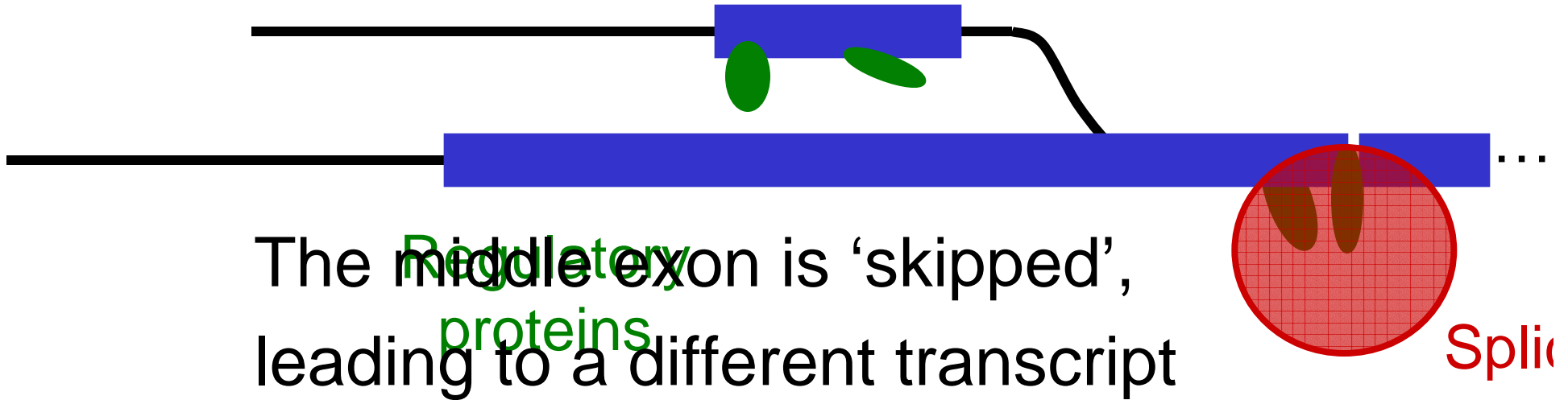
Splicing of transcripts



Splicing of transcripts

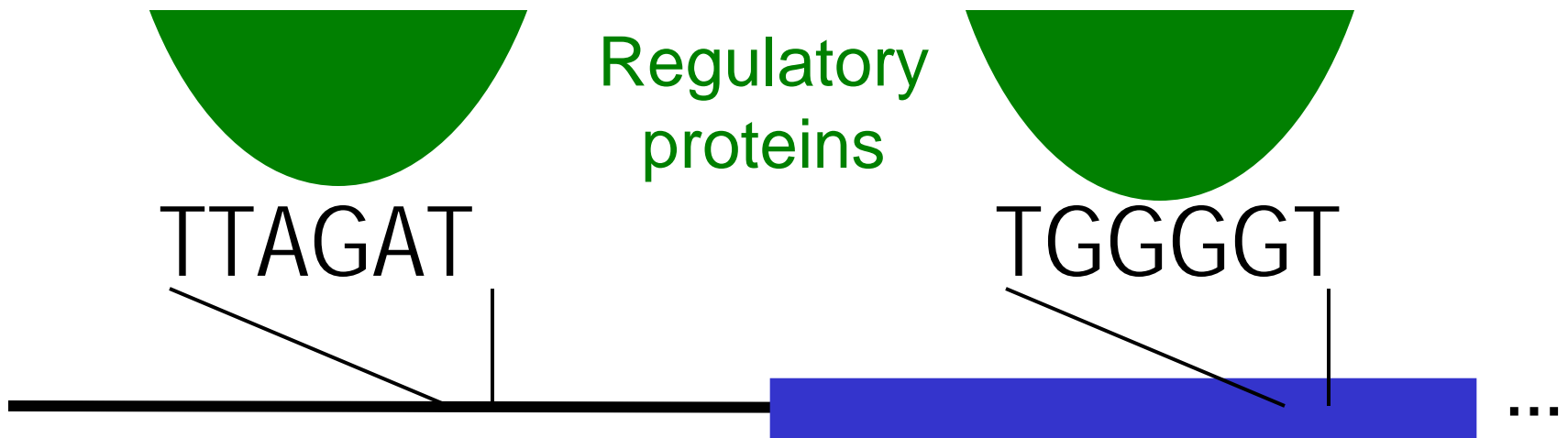


Splicing of transcripts



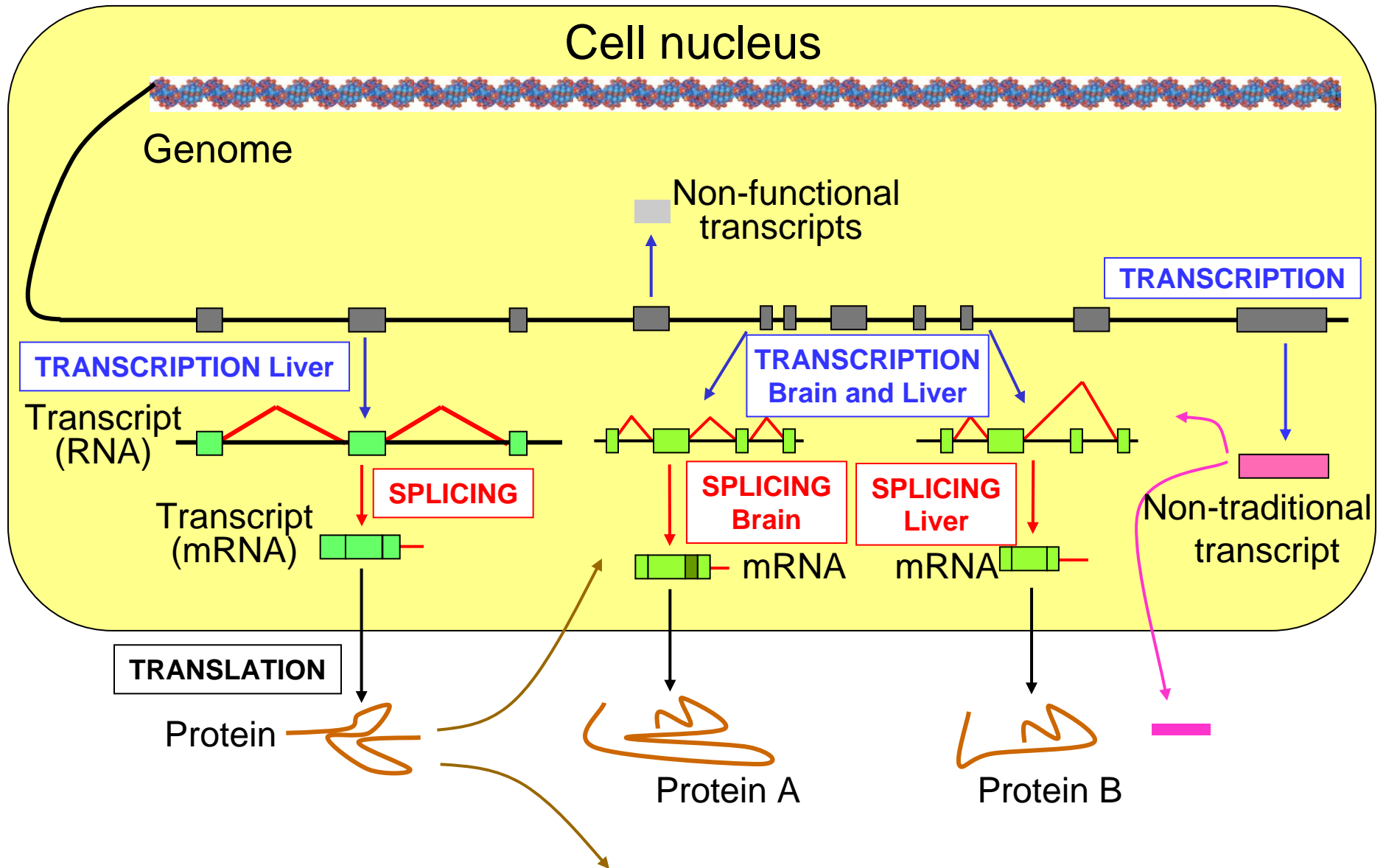
Splicing of transcripts

- Codewords in the introns and exons bind to corresponding **regulatory proteins**

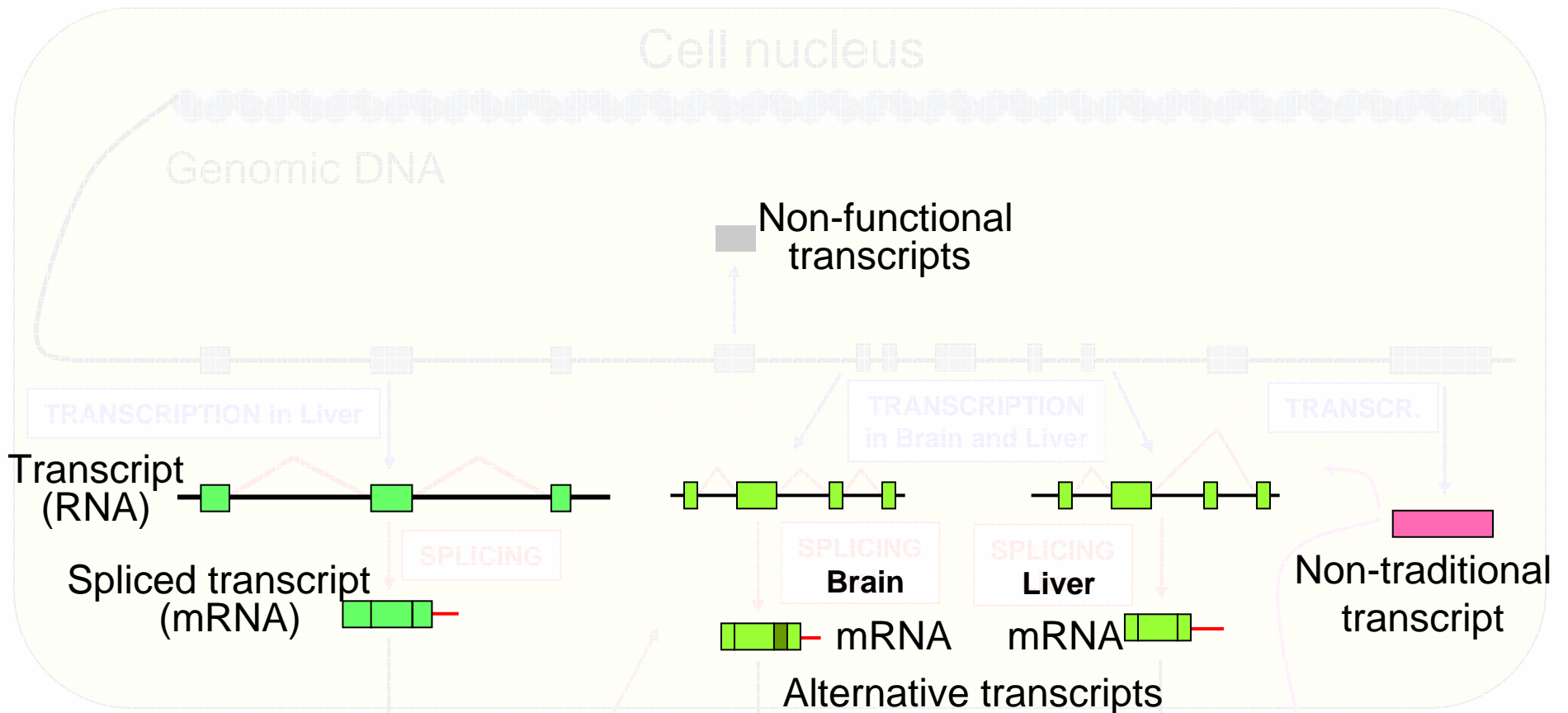


- Code: Set of regulatory codewords
- Signals: Concentrations of regulatory proteins and different spliced transcripts

The modern transcriptome



The modern transcriptome



... it turns out to be surprising in many ways

The Resources

Your collaborators can do lab work...

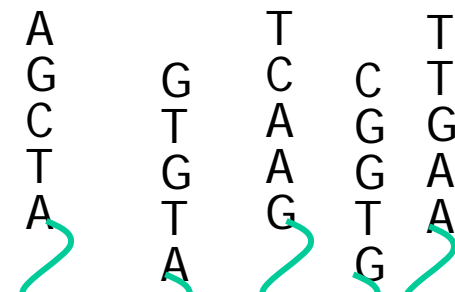
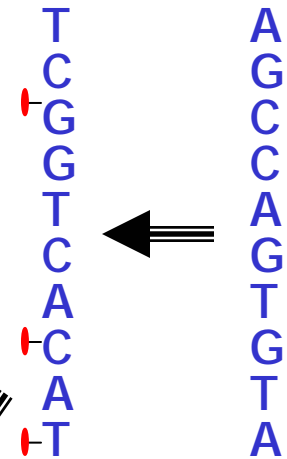
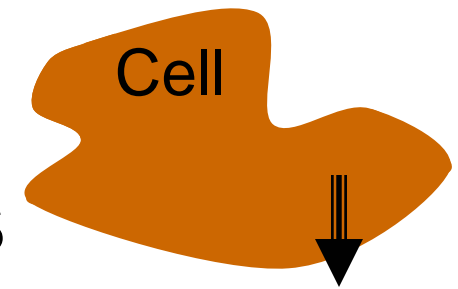
- Sequencing: Snag an actual transcript and figure out its sequence
- Microarrays: Find out if your predicted transcript fragment is expressed in a tissue sample
- Mass spectrometry: Find out if a protein is present in a sample

Databases

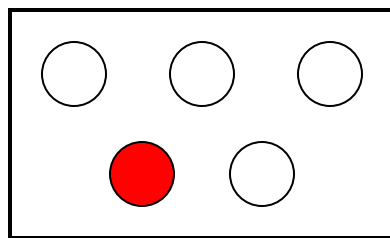
- Genomes
- Genome annotations
- Libraries of observed transcript fragments
- Microarray datasets containing measured concentrations of transcripts
- ...

Measuring transcript concentrations using microarrays

1. Fabricate microarray with probes
2. Extract transcripts from cell
3. Add florescent tag
4. Hybridize tagged sequence to microarray
5. Excite florescent tag with laser and measure intensity

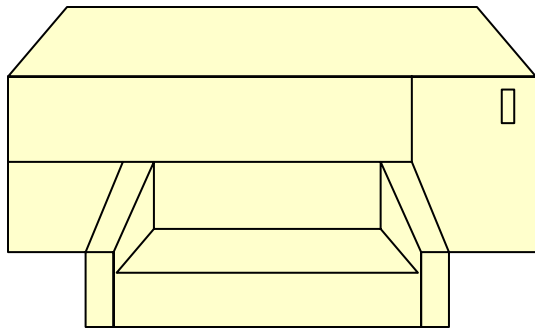


probes

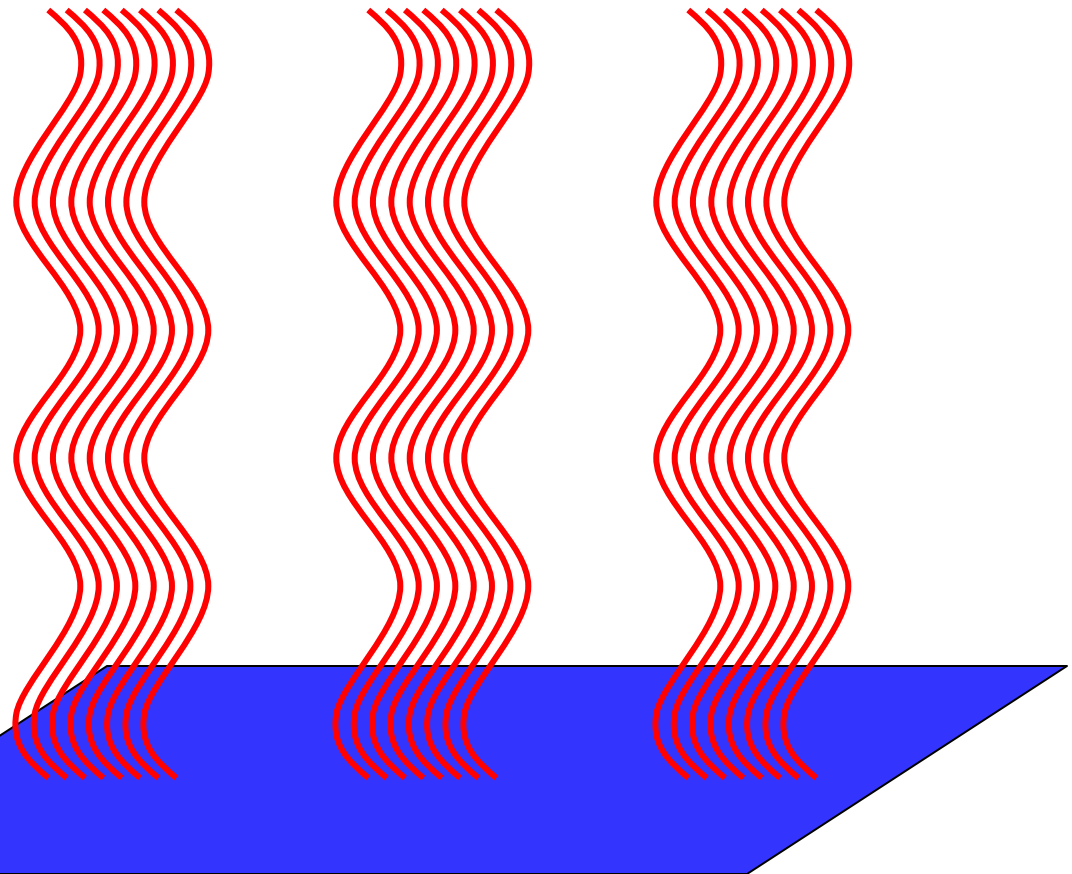


Inkjet printer technology

Hughes et al, Nature Biotech 2001



Print nucleic acid sequences using inkjet printer



Then and now...

- First microarrays (late 1990s)
 - ‘Cancer chips’, ‘gene chips’, ...
 - 5,000-10,000 probes per slide
 - Noisy
- Current microarrays
 - ‘Sub-gene resolution’
 - 200,000 probes per slide
 - Low noise
 - Multi-chip designs are cost effective

The Case Study:
Discovering protein-making transcripts
using factor graphs

BJ Frey, ..., TR Hughes
Nature Genetics, September 2005

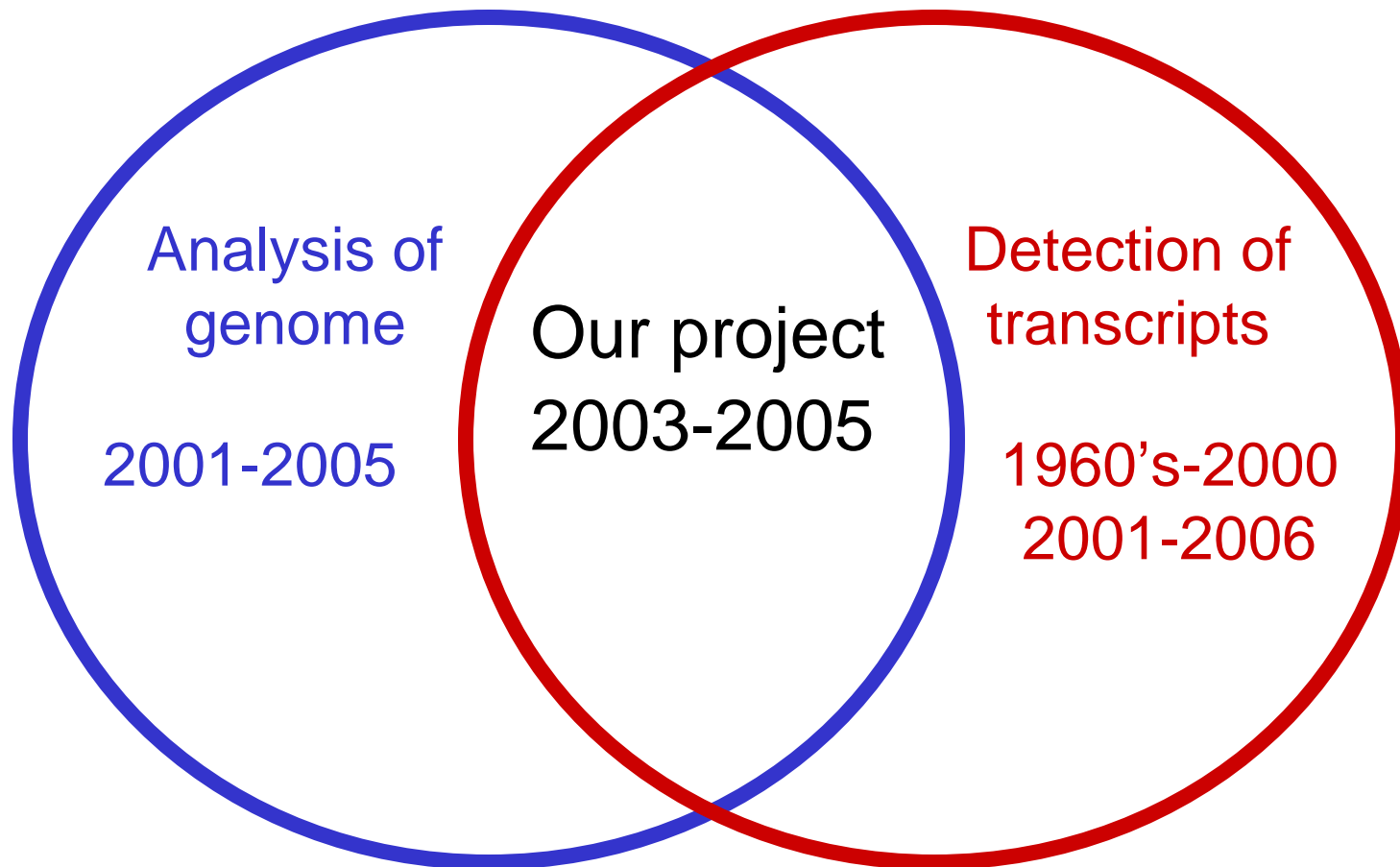
Controversy about the gene library

Despite Frey et al's impressive computational **reconstruction of gene structure**, we argue that this **does not prove the complexity of the transcriptome**

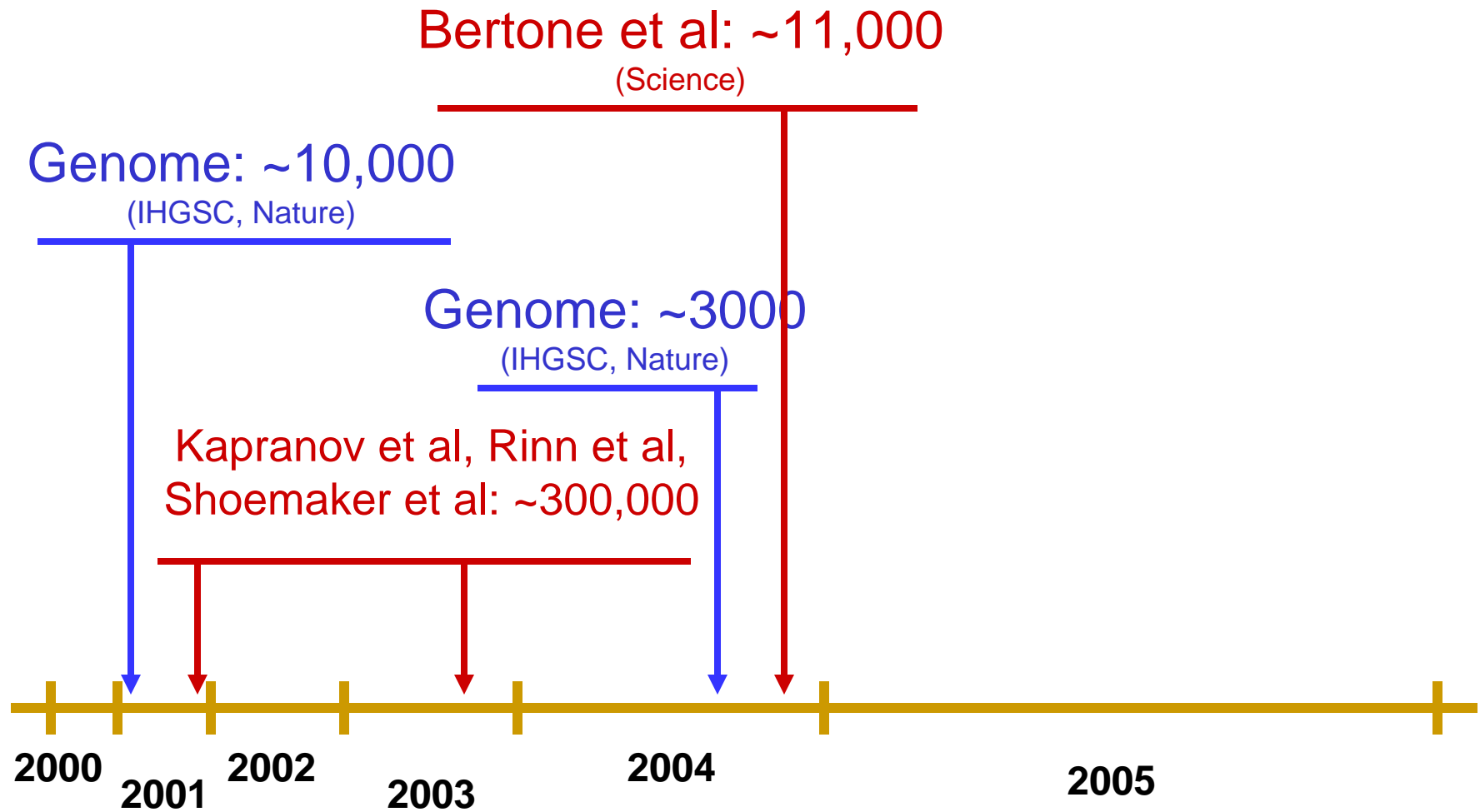
– FANTOM/RIKEN Consortium
Science, March **2006**

How it all started...

Research on the transcriptome

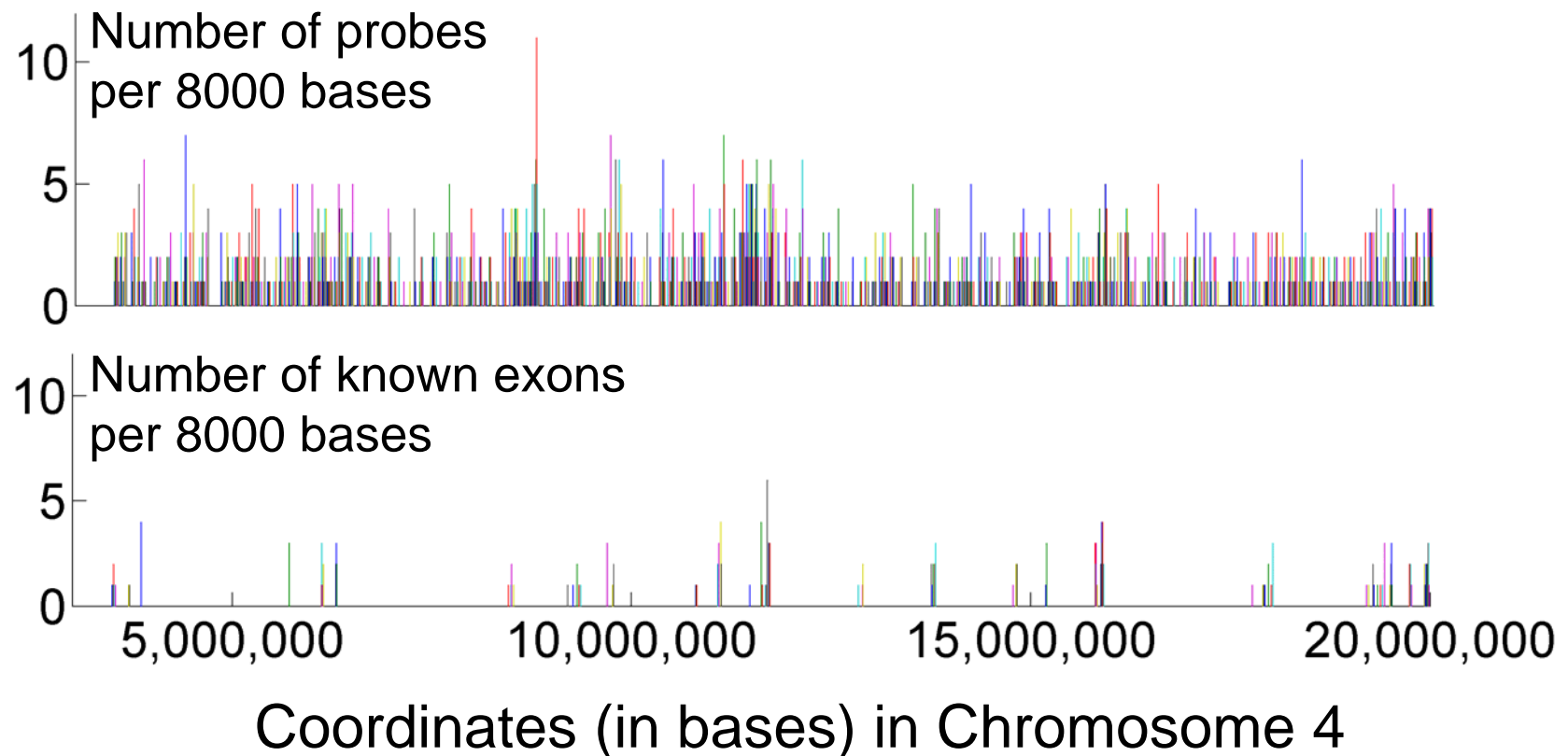


Estimates of number of undiscovered genes



Our microarrays

- Our genome analysis highlighted 1 million possible exons (~180,000 already known)
- We designed one 60-base probe for each possible exon



Our samples (37 tissues)

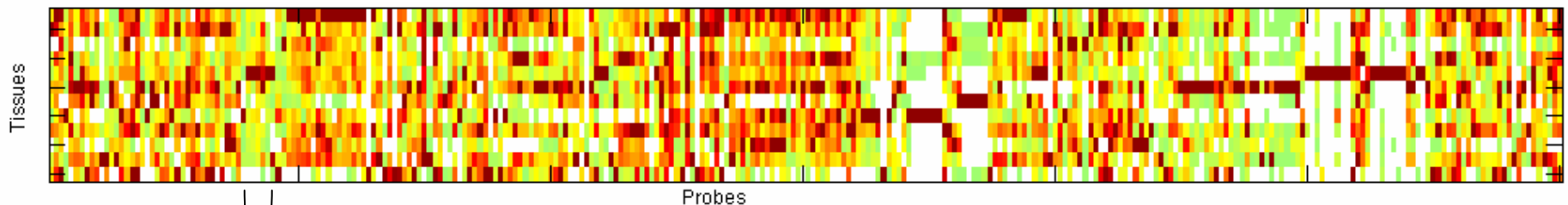


Twelve pools of mouse mRNA

Pool	Composition (mRNA per array hybridization)
1	Heart (2 μg), Skeletal muscle (2 μg)
2	Liver (2 μg)
3	Whole brain (1.5 μg), Cerebellum (0.48 μg), Olfactory bulb (0.15 μg)
4	Colon (0.96 μg), Intestine (1.04 μg)
5	Testis (3 μg), Epididymis (0.4 μg)
6	Femur (0.9 μg), Knee (0.4 μg), Calvaria (0.06 μg), Teeth+mandible (1.3 μg), Teeth (0.4 μg)
7	15d Embryo (1.3 μg), 12.5d Embryo (12.5 μg), 9.5d Embryo (0.3 μg), 14.5d Embryo head (0.25 μg), ES cells (0.24 μg)
8	Digit (1.3 μg), Tongue (0.6 μg), Trachea (0.15 μg)
9	Pancreas (1 μg), Mammary gland (0.9 μg), Adrenal gland (0.25 μg), Prostate gland (0.25 μg)
10	Salivary gland (1.26 μg), Lymph node (0.74 μg)
11	12.5d Placenta (1.15 μg), 9.5d Placenta (0.5 μg), 15d Placenta (0.35 μg)
12	Lung (1 μg), Kidney (1 μg), Adipose (1 μg), Bladder (0.05 μg)

Signal: The data (small part of the data from Chromosome 4)

Each column is an expression profile



Example of a transcript

Code:

A 'vector repetition code with deletions'



The transcript model

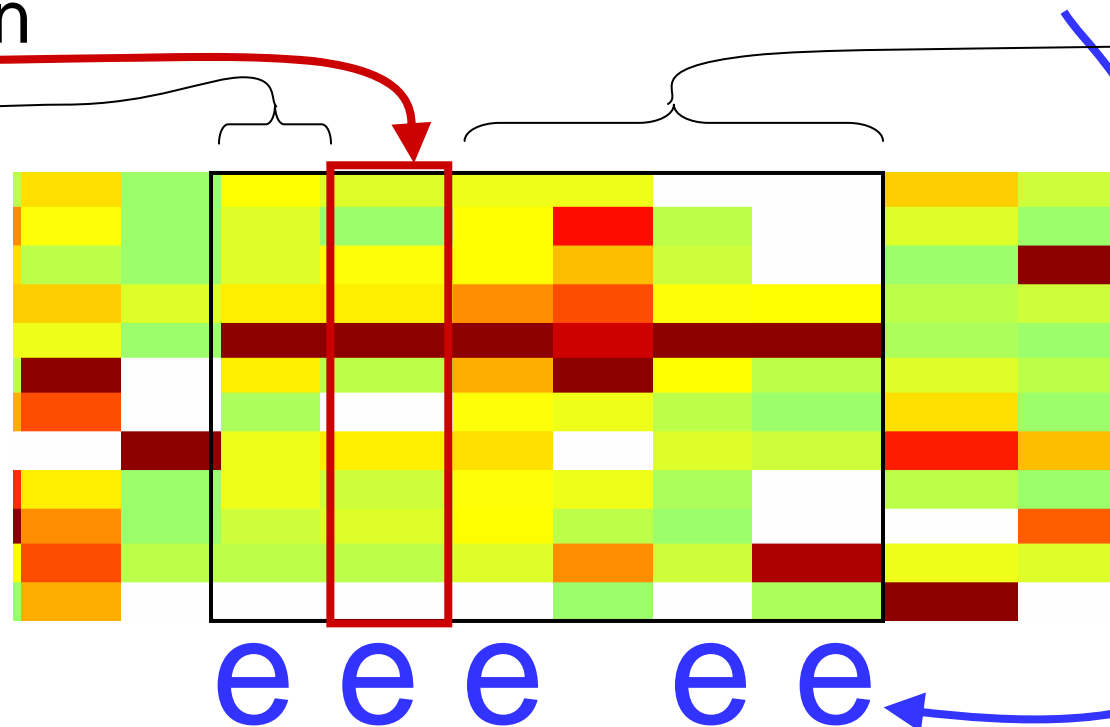
Each transcript is modeled using

A prototype expression profile

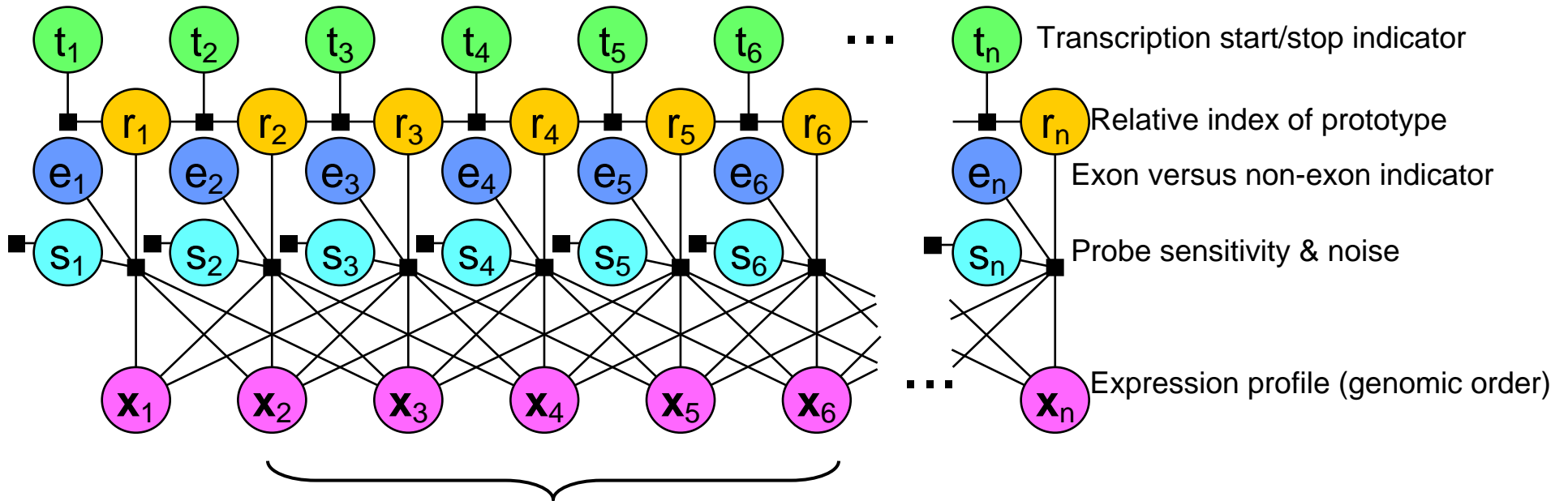
probes before prototype (eg, 1)

probes after prototype (eg, 4)

Flag indicating whether each probe corresponds to an exon



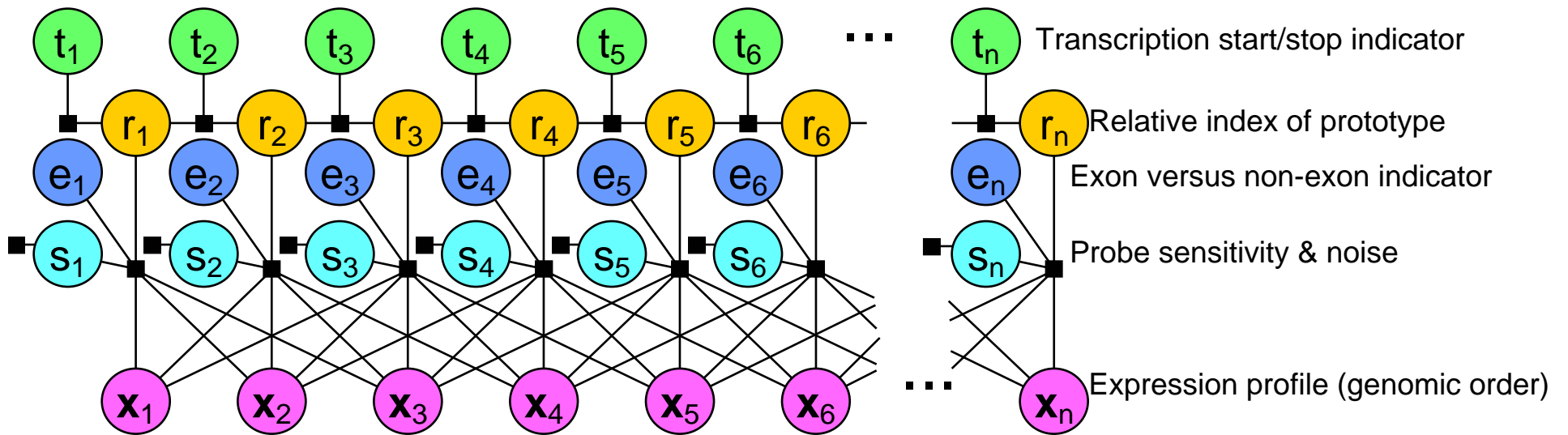
The factor graph



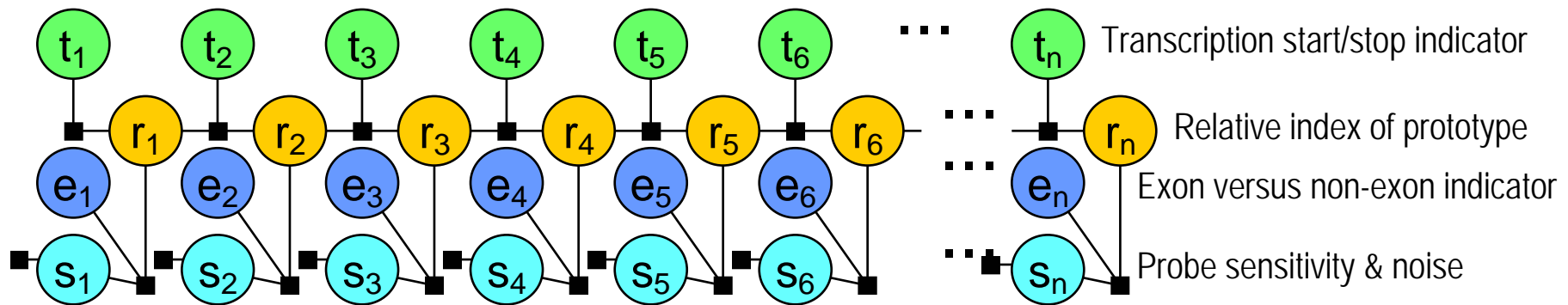
The prototype for x_i is x_{i+r_i} , $r_i \in \{-W, \dots, W\}$. We use $W=100$

ONLY 1 FREE PARAMETER:
 \mathcal{K} , probability of starting a transcript

After expression data (x) is observed, the factor graph becomes a tree



After expression data (x) is observed, the factor graph becomes a tree



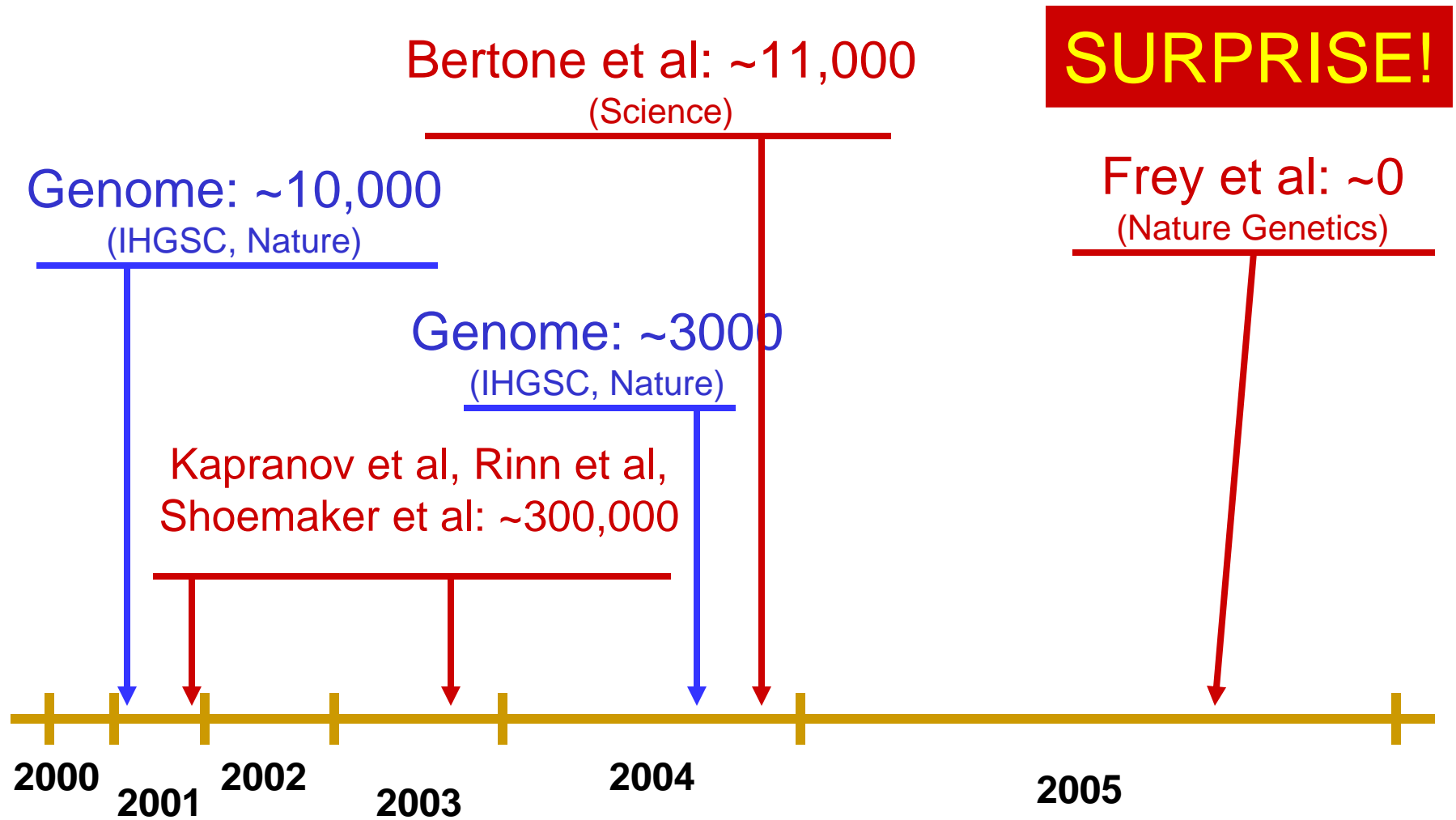
Computation: The max-product algorithm performs **exact inference and learning**.

Summary of results *

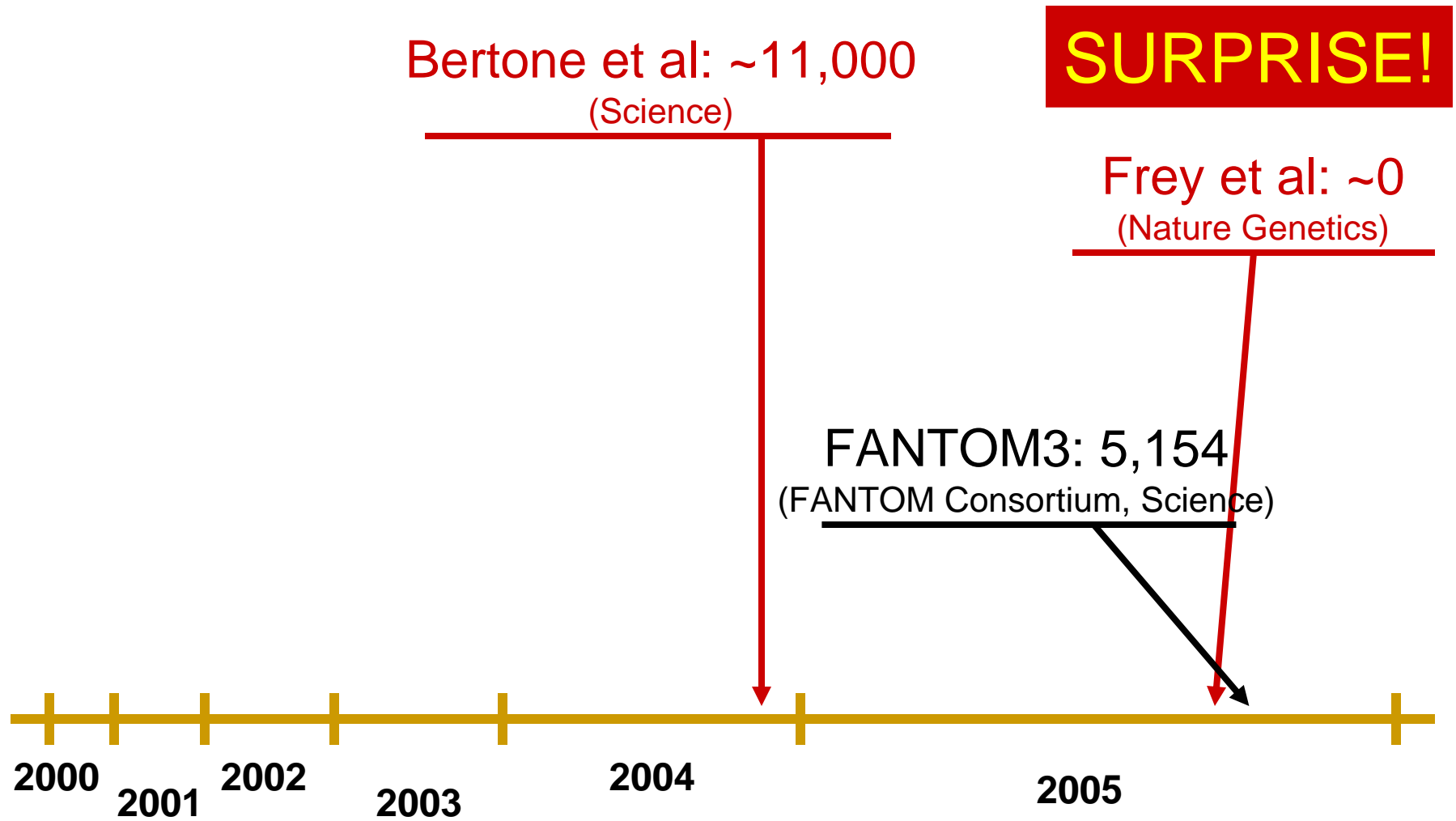
- 10 X more sensitive than other transcript-based methods
- Detected 155,839 exons
- Predicted ~30,000 new exons
- Reconciled discrepancies in thousands of known transcripts

* Exon false positive rate: 2.7%

Revisiting Estimates of number of undiscovered genes



Contentious results



... [We discovered] new mouse protein-coding transcripts, including **5,154 encoding previously-unidentified proteins** ...

– FANTOM/RIKEN Consortium
Science, Sep 2005

We wondered: Are these really new genes?

... we found that **2917 of the FANTOM proteins** are in fact splice isoforms of **known transcripts** ...

- Frey et al
Science, March 2006

... the **number of new protein-coding genes** found by us has been **revised from 5154 to 2222...**

- FANTOM/RIKEN Consortium
Science, March 2006

Last word...

... the number of completely new protein-coding genes discovered by the FANTOM consortium is at most in the hundreds...

– Frey et al
Science, March 2006

The Closing Remarks

Open problems

- Producing genome-wide libraries of functioning transcripts, including
 - Alternatively-spliced transcripts
 - Transcripts that don't make proteins
- Understanding functions of transcripts
- Developing models of how transcription and alternative splicing are regulated
- Developing models of gene interactions
 - 'Genetic networks'

Should you work in computational biology?

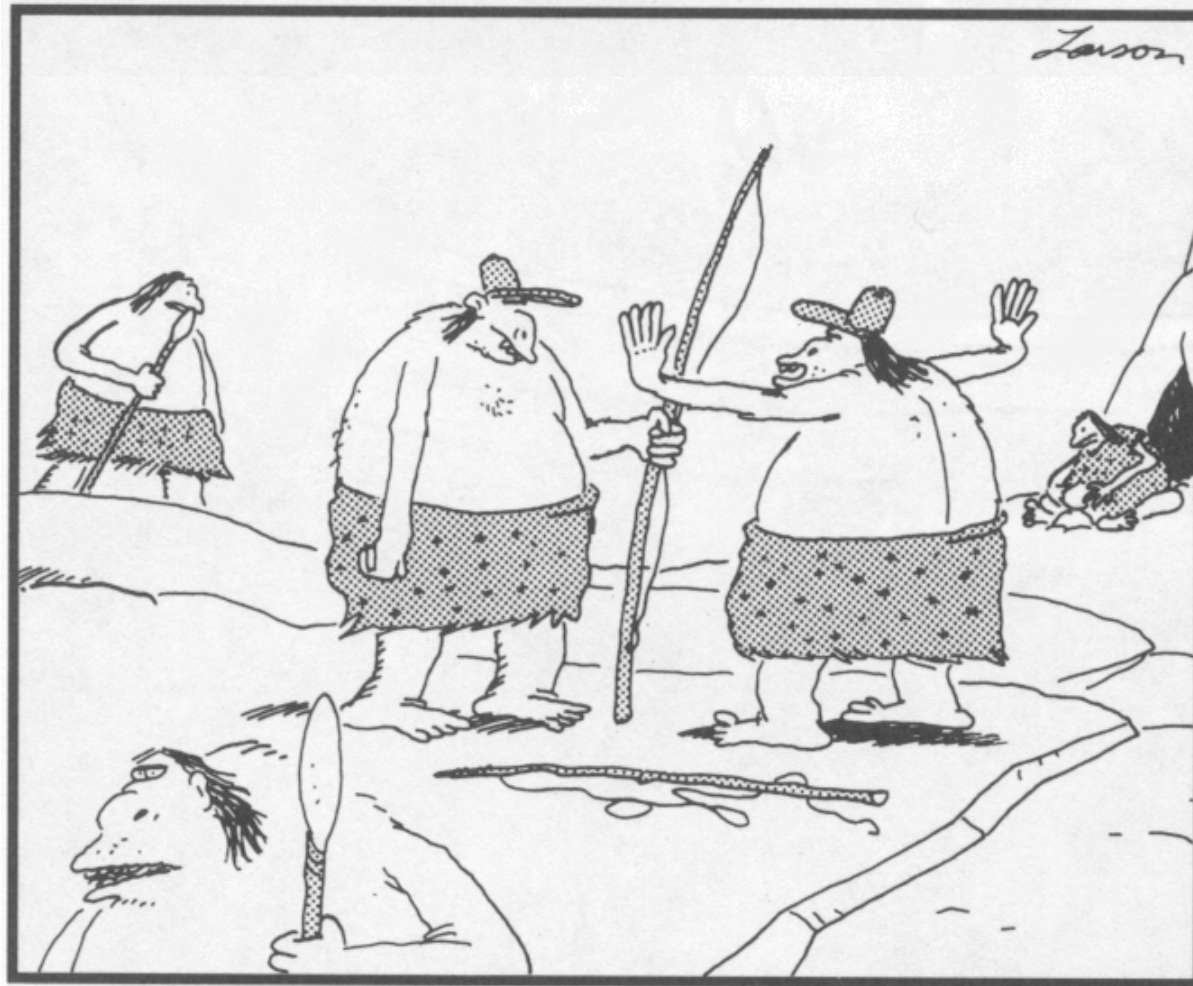
Pluses

- A major scientific frontier
- Potential for high impact on society

Minuses

- Mostly a collection of facts
- Mechanisms are complex and beyond our control
- Lacking a mathematical framework

Remember, communication theory also once lacked a mathematical framework...



"Ok, Zorg, lets try using a *prefix code*"

Should you work in computational biology?

Pluses

- A major scientific frontier
- Potential for high impact on society
- **Lacking a mathematical framework**

Minuses

- Mostly a collection of facts
- Mechanisms are complex and beyond our control

How do you enter this field?

- Hire a tutor (ie, student or postdoc)
- Hire a programmer
- Get involved in several ‘winner’ projects
- Be prepared to drop ‘loser’ projects
- Build mutually-beneficial collaborations
- How long will it take?

For more information...

- As of Friday July 14, 2006:
<http://www.psi.toronto.edu/isit2006.html>
 - These slides
 - Pointers to helpful papers, databases, etc

Acknowledgements

- Frey Group

- Quaid D Morris (postdoc)
- Leo Lee (postdoc)
- Yoseph Barash (postdoc)
- Ofer Shai (PhD)
- Inmar Givoni (PhD)
- Jim Huang (PhD)
- Marc Robinson (programmer)

Genomics Collaborators

- Hughes' Lab
- Blencowe's Lab
- Emili's Lab
- Boone's Lab

Medical Collaborators:
E Sat, J Rossant, BG
Bruneau, JE Aubin