Analysis of Microarray Data with Methods from Machine Learning and Network Theory

Summer Lecture 2015

Prof. Dr. A. B. Cremers
Dr. Jörg Zimmermann

Introduction to DNA Microarrays

- A technology that is reshaping molecular biology
- "Industrialization" of biological data acquisition
- Shifting the focus from data acquisition to data analysis

Current and Potential Uses and Applications

- Molecular diagnosis of diseases
- Disease characterization
- Target Identification
- Pathway Mapping
- Prediction of Drug Efficacy
- Toxicology
- Personalized Medicine
- ...

Types of Microarrays

Microarray has become a general term, there are many types now:

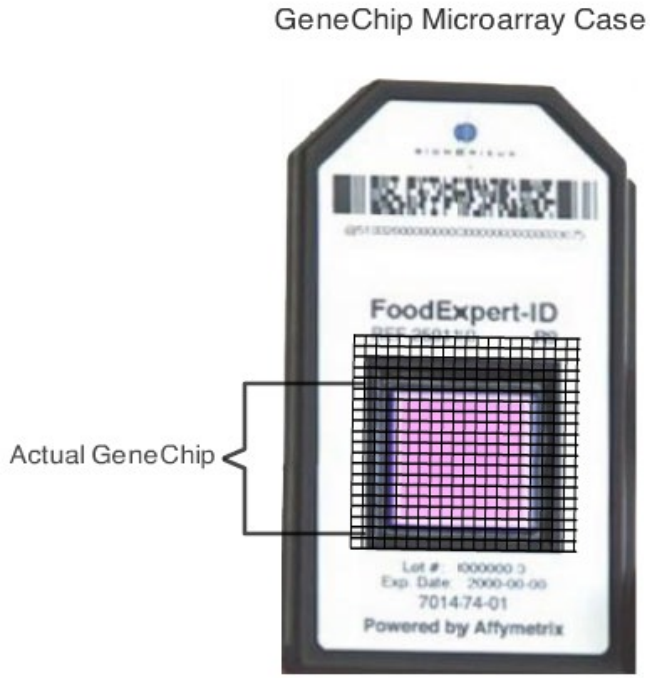
- DNA Microarrays
- Protein Microarrays
- Tissue Microarrays
- ...

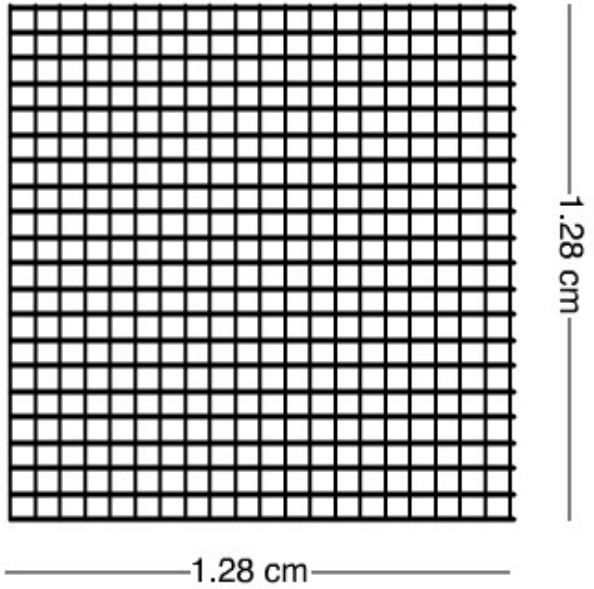
We will discuss DNA Microarrays

What is a DNA Microarray?

- A grid of DNA spots (probes) on a substrate used to detect complementary sequences
- The substrate can be plastic, glass, silicon
- RNA/DNA of interest is labelled and hybridizes with the probes on the array
- Hybridization with probes is detected optically (laser-induced fluorescence)

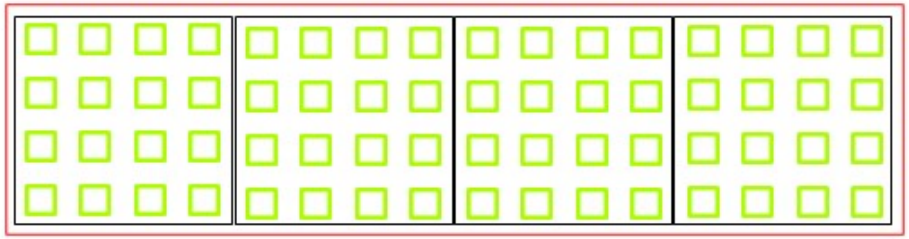
Structure of a DNA MicroarrayChip

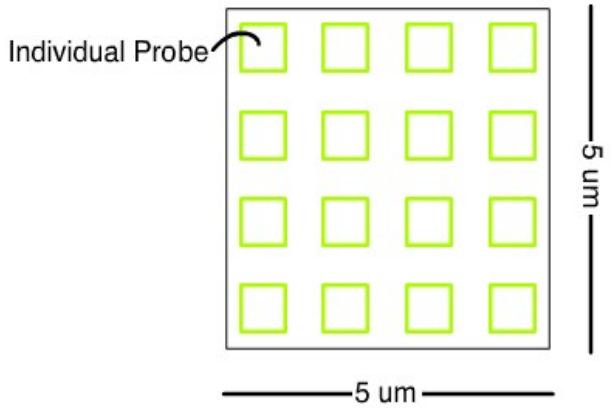




Gene Expresion Profiling

mRNA levels are averaged over a population of probe spots:





Central Assumption of DNA Microarrays

- The level of a given mRNA is positively correlated with the expression of the associated gene:

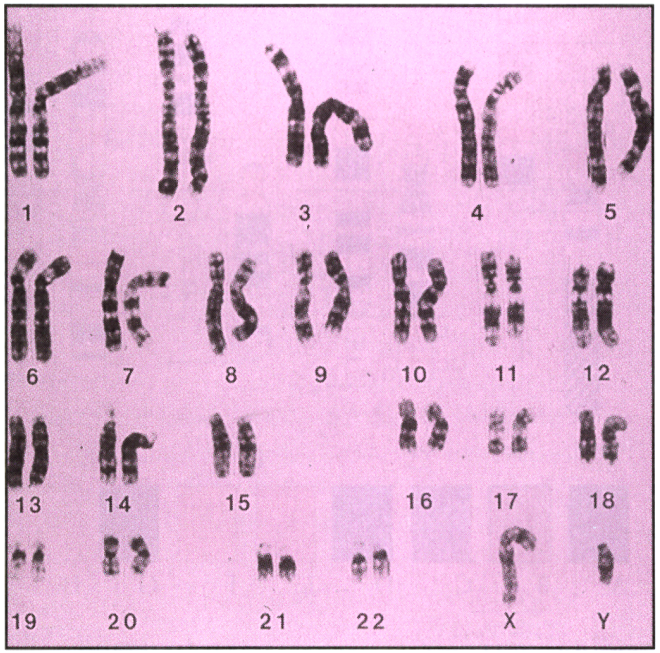
  Higher mRNA levels mean higher protein expression, lower mRNA means lower protein expression

- Other factors are comparatively small:

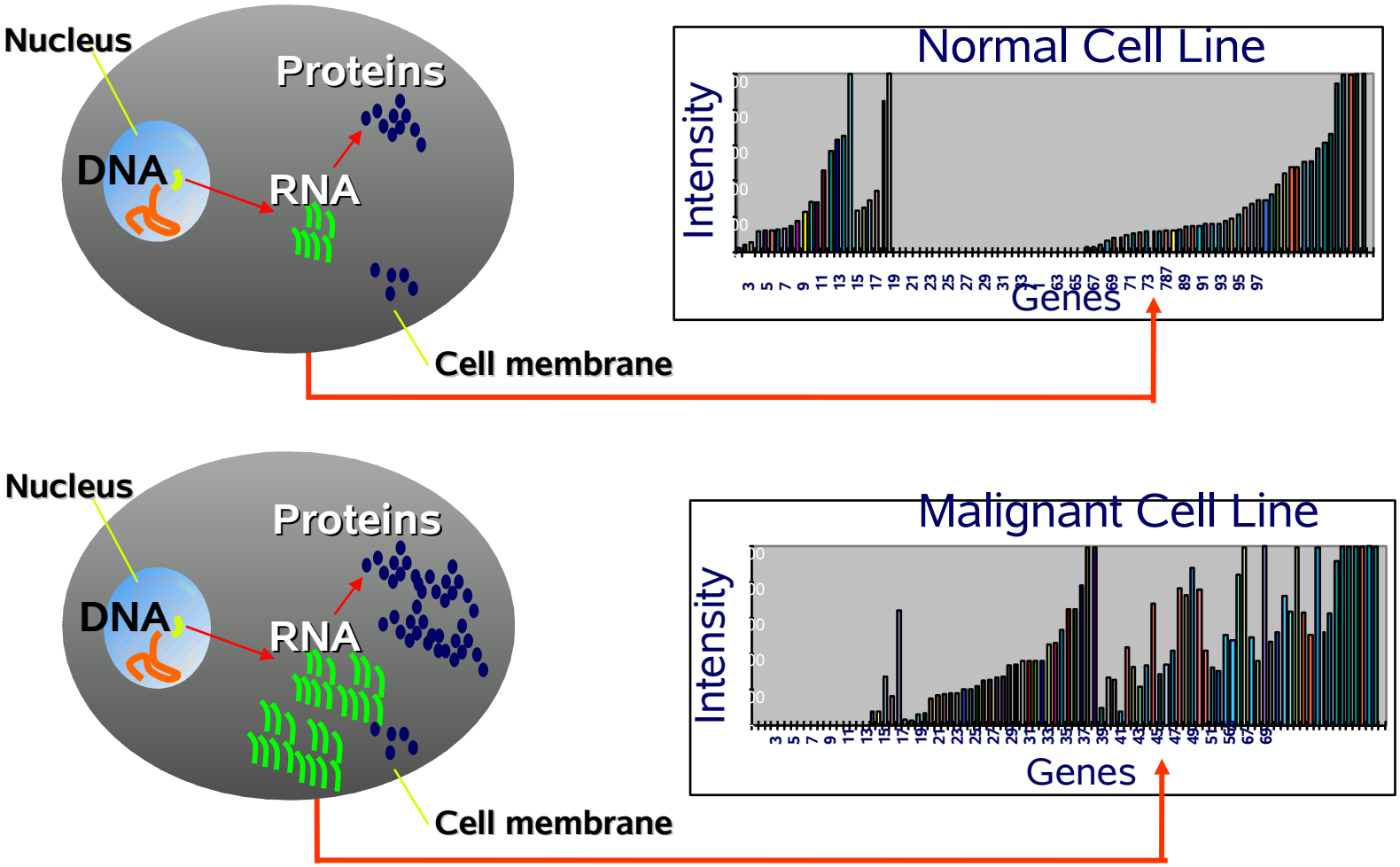
  protein degradation, mRNA degradation, codon preference, translation rates, translation lag, ...

The Human Genome

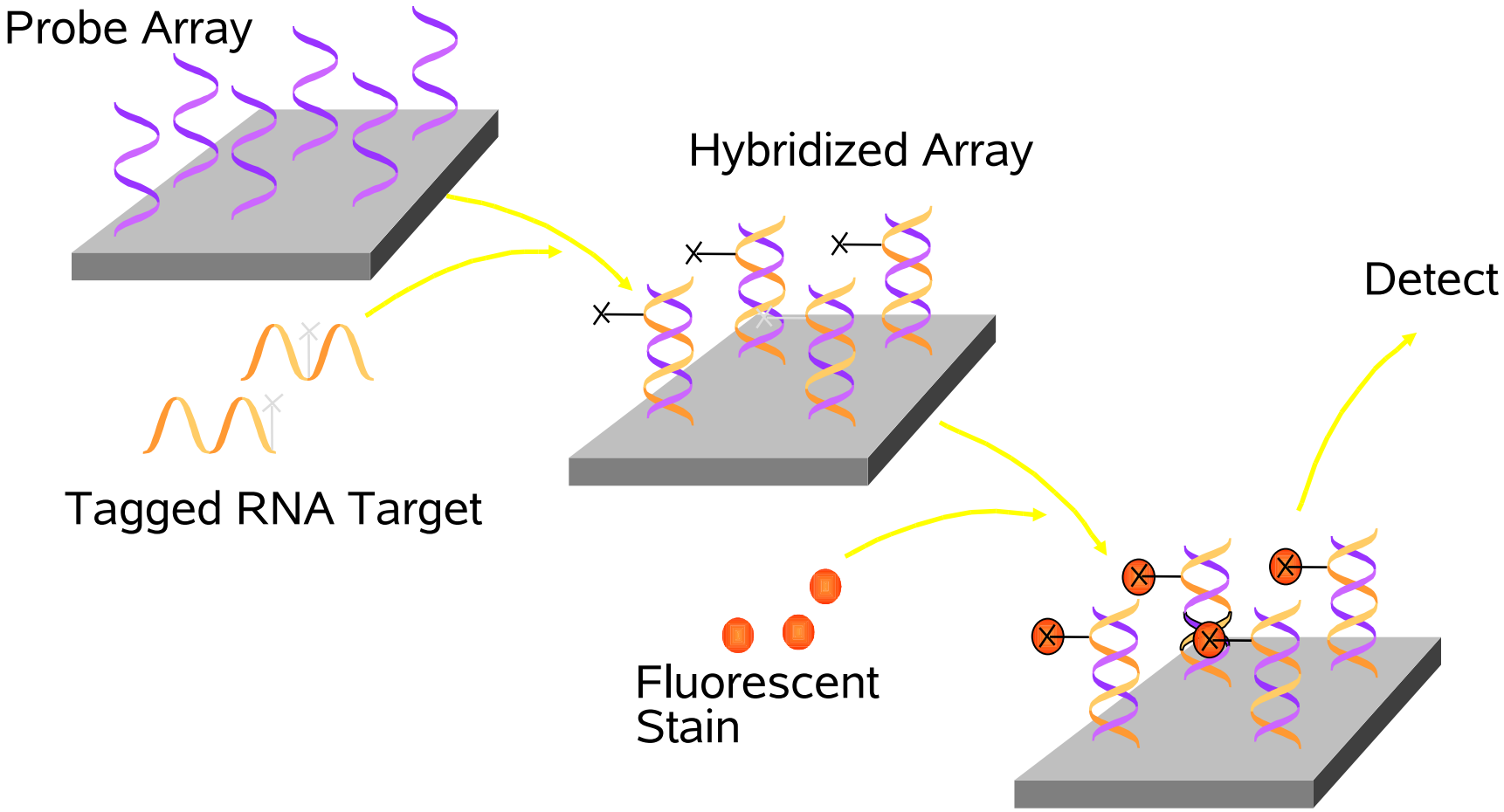
- Only ~1.5% of genome (genetic material) encodes proteins
- ca. 20,000 genes
- Millions of differences between any two people, but overall 99.9% are the same



Different Cells - different Expression profiles



The Process leading to Expression Data

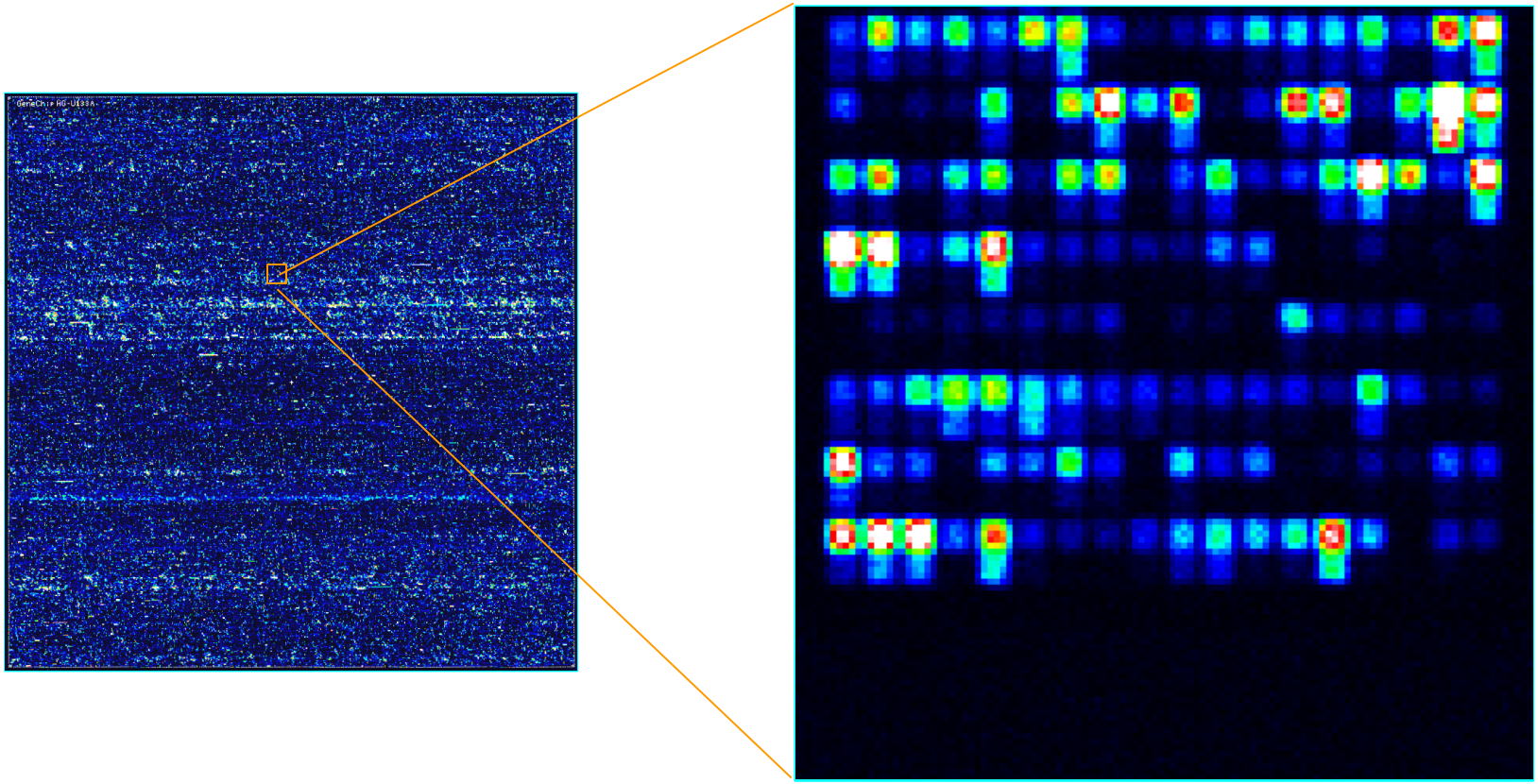


Estimation of Gene Expression Indexes

- DNA Chips consist of several 100000 probes
- A pobe consists of short DNA sequences
- Target mRNA binds to the corresponding probes by hybridization
- 16 probes of same DNA sequences form a probe set
- Laser scan of hypridized DNA Chip results in a CEL-file

Laser Image of hybridized Array

Laser scanning of chip results in reflection intensity values for all probes



The CEL-file format (Header)

[CEL]
Version=3

[HEADER]
Cols=712
Rows=712
TotalX=712
TotalY=712
OffsetX=0
OffsetY=0
GridCornerUL=215 233
GridCornerUR=4476 257
GridCornerLR=4460 4527
GridCornerLL=198 4503
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[0..46132] CL20030502101AA:CLS=4733 RWS=4733 XIN=3 YIN=3 VE=17
  2.0 05/02/03 13:39:35 HG-U133A.1sq 6
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004

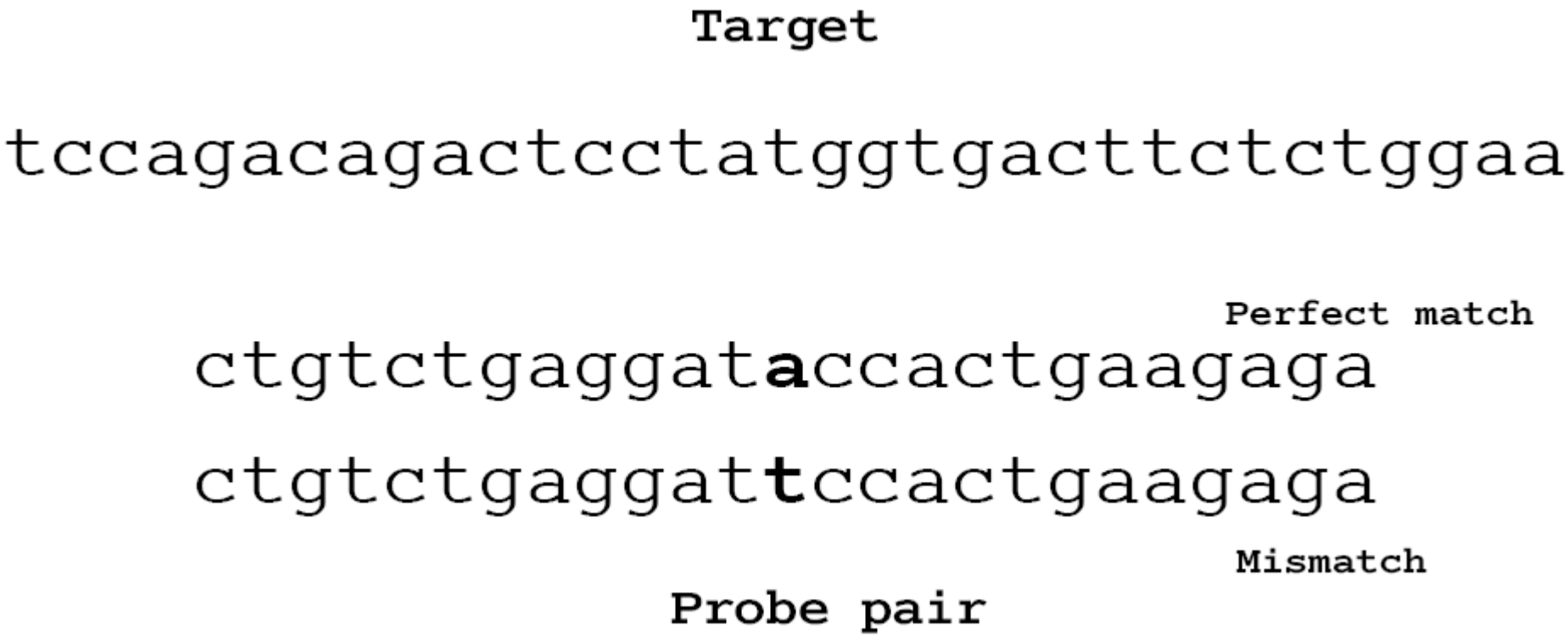
The CEL-file format (Body)

[INTENSITY]
NumberCells=506944

CellHeader=X	Y	MEAN	STDV	NPIXELS
0	0	135.0	26.1	16
1	0	7654.5	675.6	16
2	0	156.5	25.3	16
3	0	7813.4	821.6	16
4	0	105.5	19.8	16
5	0	137.0	24.1	16

.
.
.
.

Perfect Match and Mismatch



Perfect Match and Mismatch

- Perfect match is a 25 nucleotide probe which hybridizes perfectly with target.

- Mismatch has one false nucleotide: reduced hybridization rate

- Mismatch serves as a reference for the inference of gene expression indexes from scanned intensity values!

Introduction to Inferential Statistics

Goal: Find a Hypothesis which is a good explanation for your data and quantify your uncertainty

There are two basic types of Uncertainty:

Data Uncertainty:	Hypothesis Uncertainty:
Contingency (Event risk)	Plausibility (Model risk)

If data contingency is modelled by probability, we are in the realm of classical statistics, if both contingency of data and plausibility of hypotheses is modelled by probability, it is called Bayesian statistics (there are other approaches to modeling uncertainty, this is an active area of research, especially in Machine Learning and Artificial Intelligence).