Computational Biology
Machine Learning Approaches to Biology

Computational Science Colloquium

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Bioinformatics and Mathematics

Hardy's 1908 paper:

To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...

Hardy–Weinberg principle: both allele and genotype frequencies in a population remain constant from generation to generation

Computational Biology ⊂ Bioinformatics

http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html:

N ational
C enter for
B iotechnology
I nformation

Bioinformatics:
the field of bioinformatics has evolved such that the most pressing task now involves the *analysis and interpretation* of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures

Computational biology:
actual process of analyzing and interpreting data

Overview

- virus drug targets
- single nucleotide polymorphism: schizophrenia
- microarray preprocessing
- classification: therapy outcome from gene expression profiles
- feature selection → gene selection
- diagnosis based on peptide arrays
- copy number variations for complex diseases (bipolar)
- sequence classification: recurrent neural networks
- next generation sequencing
- nucleosome positions
- WE NEED HELP

Coiled Coil Proteins: Virus Drug Target

CHARITÉ

Berlin,
Germany

Group:
Molecular Libraries
Dr. Volkmer

Peptide arrays

Coiled Coil Proteins: Virus Drug Target

University of BRISTOL

School of Chemistry
Prof. Dek Woolfson

U.S. Human Cases of Swine Flu Infection

(As of April 28, 2009, 11:00 AM ET)

State	# of laboratory confirmed cases
California	10 cases
Kansas	2 cases
New York City	45 cases
Ohio	1 case
Texas	6 cases
TOTAL COUNT	**64 cases**

Coiled Coil Proteins: Virus Drug Target

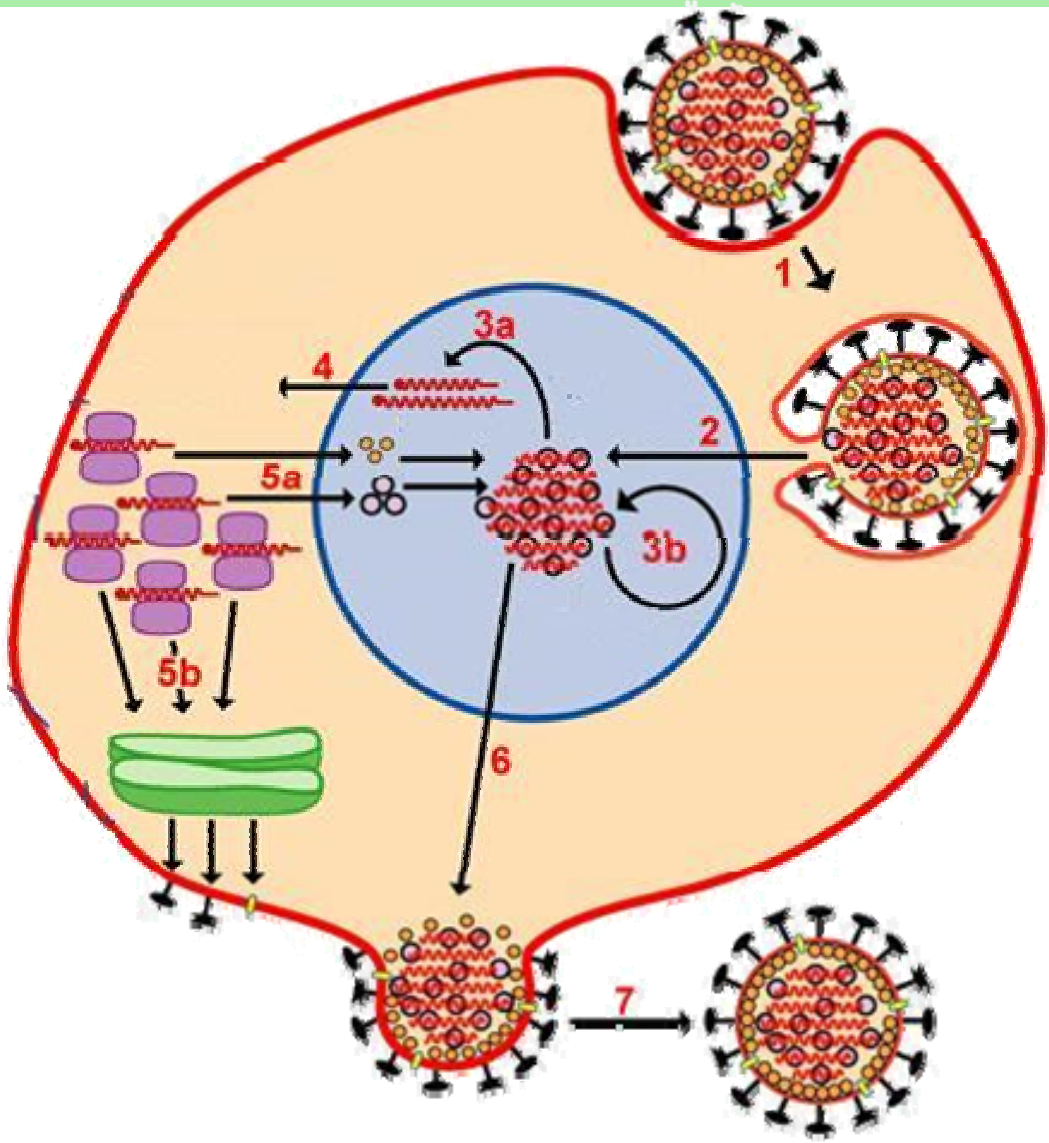
Welt

28.04.2009 um 23:33 Uhr

Schweinegrippe: Erster Fall in Wien bestätigt

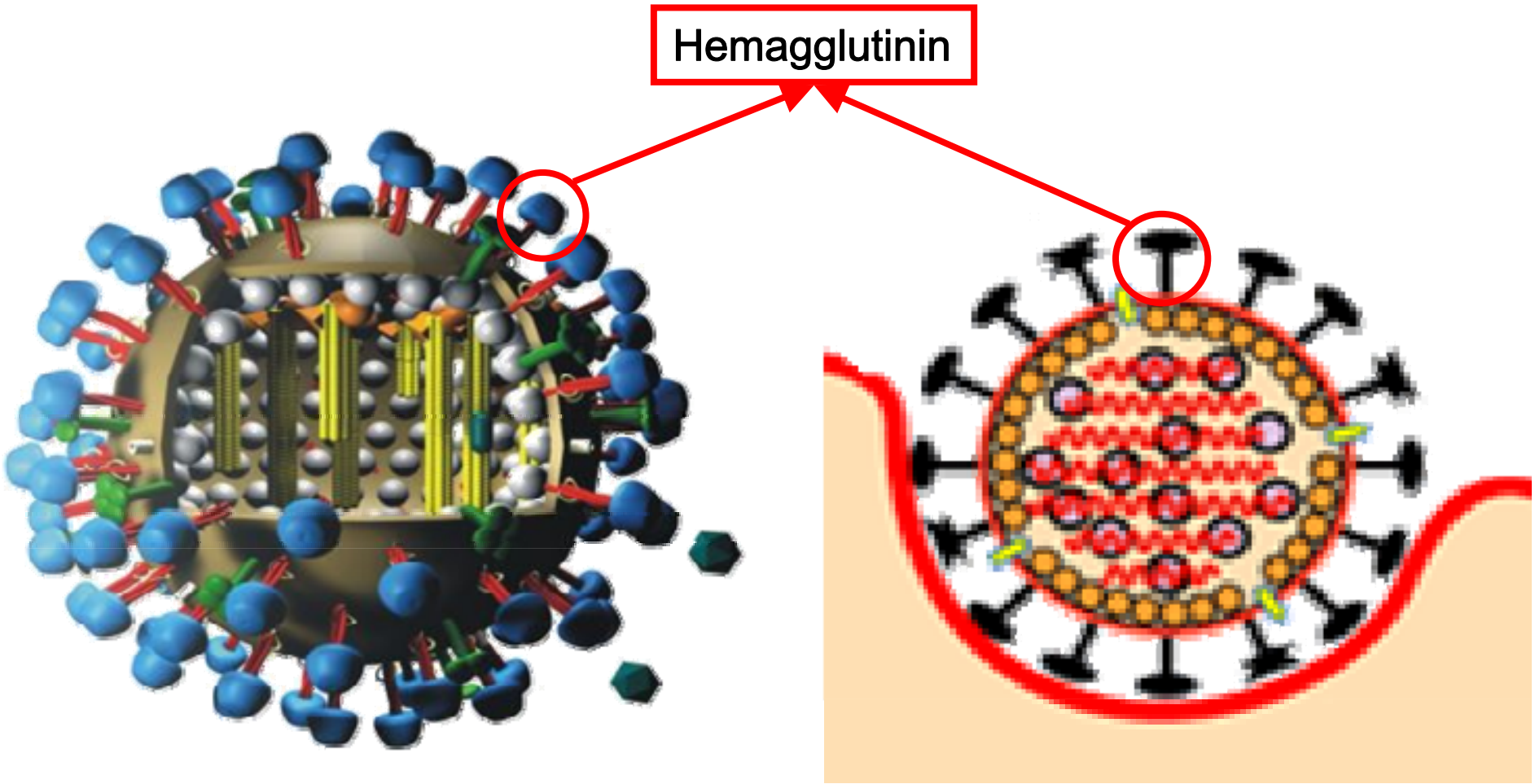
Im Krankenhaus in Steyr (OÖ) wurde heute Vormittag eine junge Frau eingeliefert, die ersten Angaben zufolge die Symptome des Schweinegrippe-Virus aufweist. Die 28-Jährige war vor kurzem von einem Mexiko-Urlaub heimgekehrt und hat sich offensichtlich dort infiziert. Die mexikanischen Behörden versuchen indes, die Herkunft des Virus ausfindig zu machen - die Spur führt zu einer Schweinefarm im Norden des Landes.

Coiled Coil Proteins: Virus Drug Target



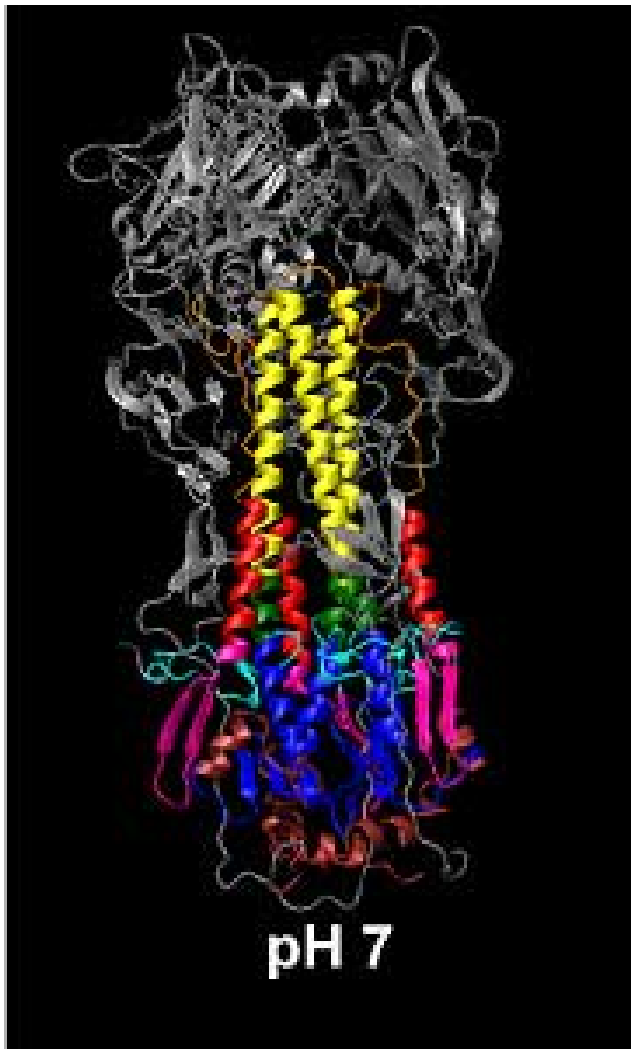
Coiled Coil Proteins: Virus Drug Target

Virus (H5N1, bird flu)

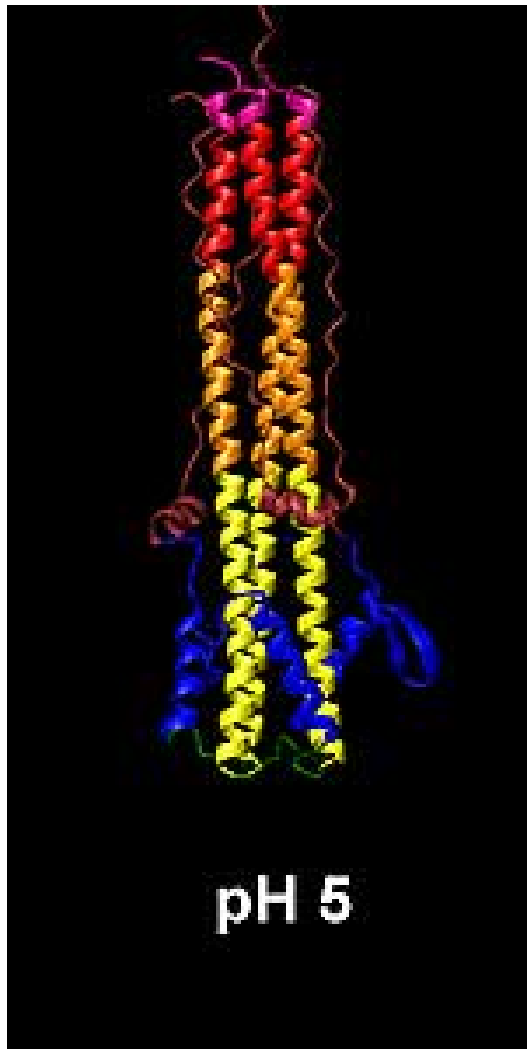


Coiled Coil Proteins: Virus Drug Target

out-side the cell



Hemagglutinin



in-side the cell

Coiled Coil Proteins: Virus Drug Target

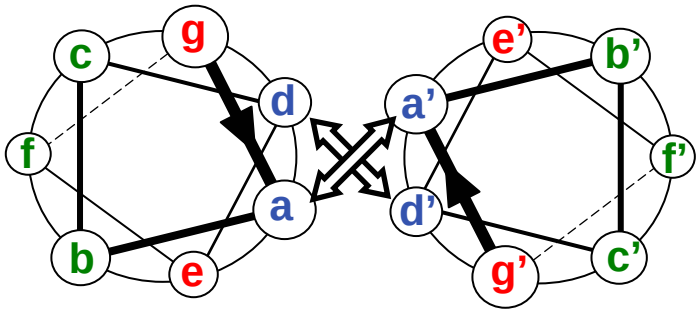
Coiled-coil domain prediction

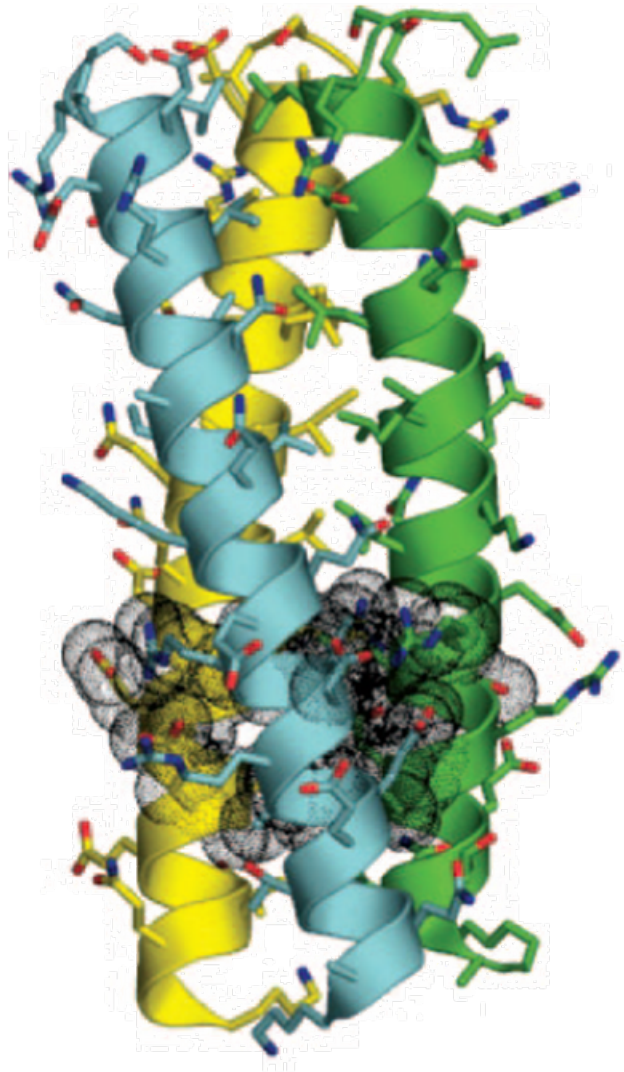
a-b-c-d-e-f-g

(a) and (d) → hydrophobic amino acids
→ interaction between helices

(e) and (g) → charged amino acids
→ inter- and intra-helical salt bridges

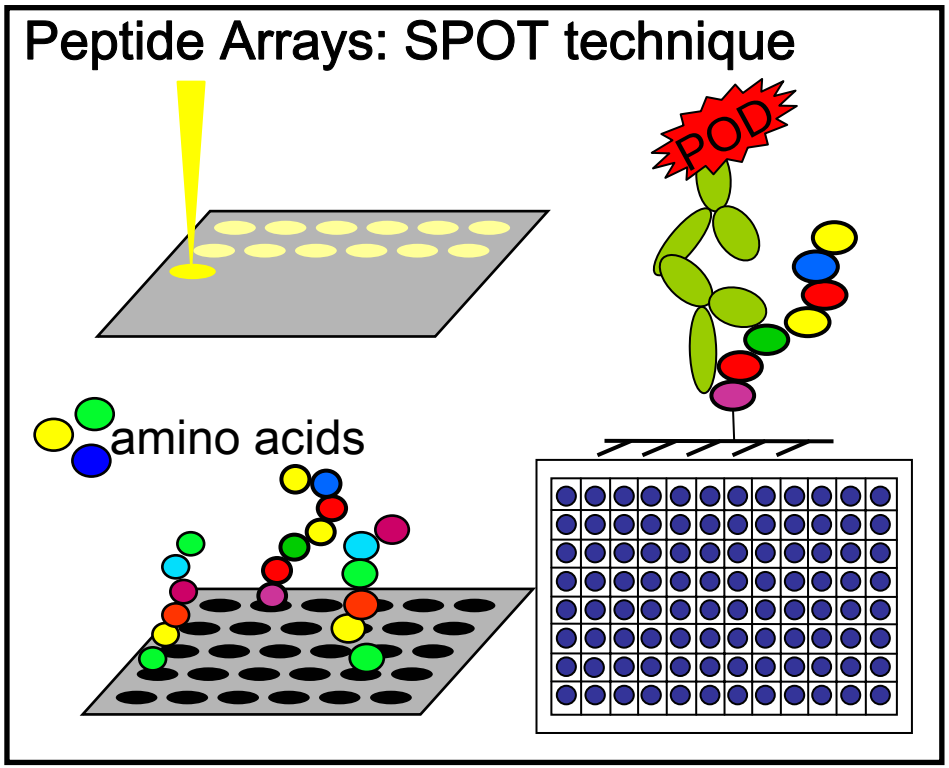
(b), (c), and (f) → polar amino acids
→ outer surface

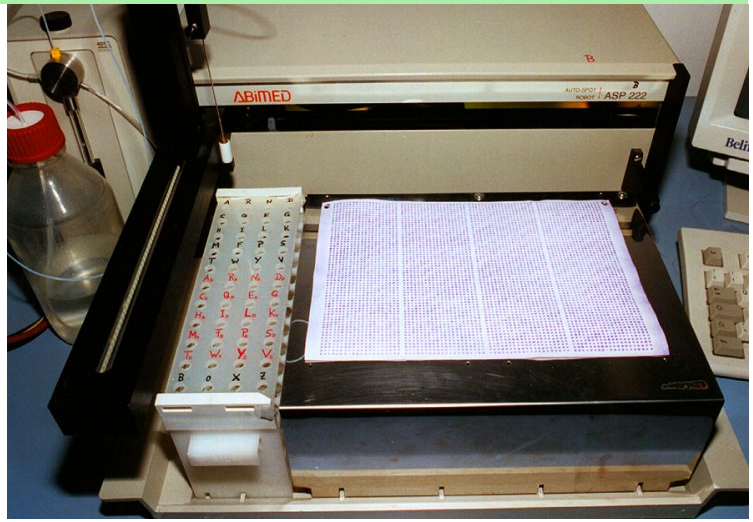


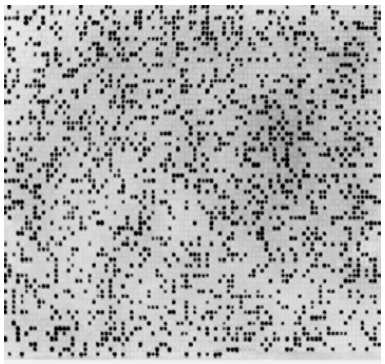


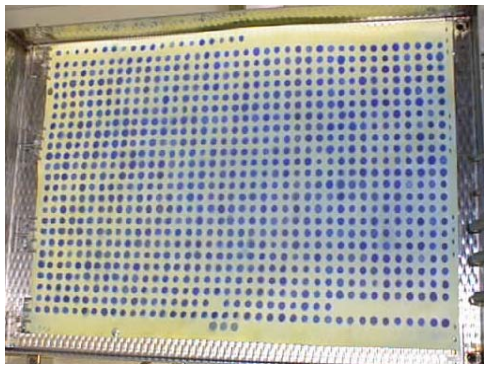
RMKQLEDKVEELLSKNYHLENEVARLKKLVG
abcdefgabcdefgabcdefgabcdefgab

Coiled Coil Proteins: Virus Drug Target









Single Nucleotide Polymorphism

CHARITÉ

Berlin,
Germany

Group:
Dept. Psychiatry and Psychotherapy
Prof. Andreas Heinz

Group:
Dept. Nephrology and Internal Intensive Care
Prof. Petra Reinke

Single Nucleotide Polymorphism

Cleveland Clinic

Cleveland, Ohio,
United States

Group: Experimental Haematology and Hematopoiesis
Taussig Cancer Center
Prof. Dr. J. Maciejewski

Single Nucleotide Polymorphism

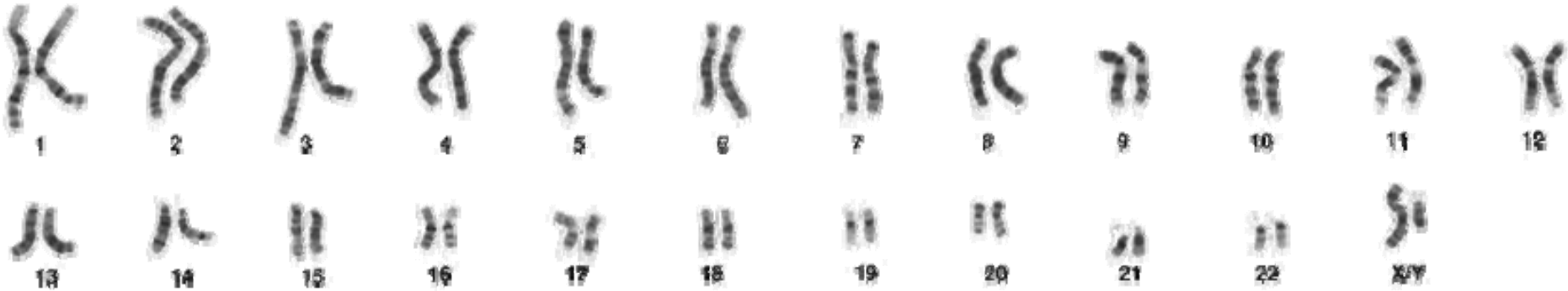
single nucleotide polymorphism
(SNP - pronounced snip)

➥Variation in the DNA at the same position in at least 1% of the population

➥SNPs occur all 100 to 300 base pairs

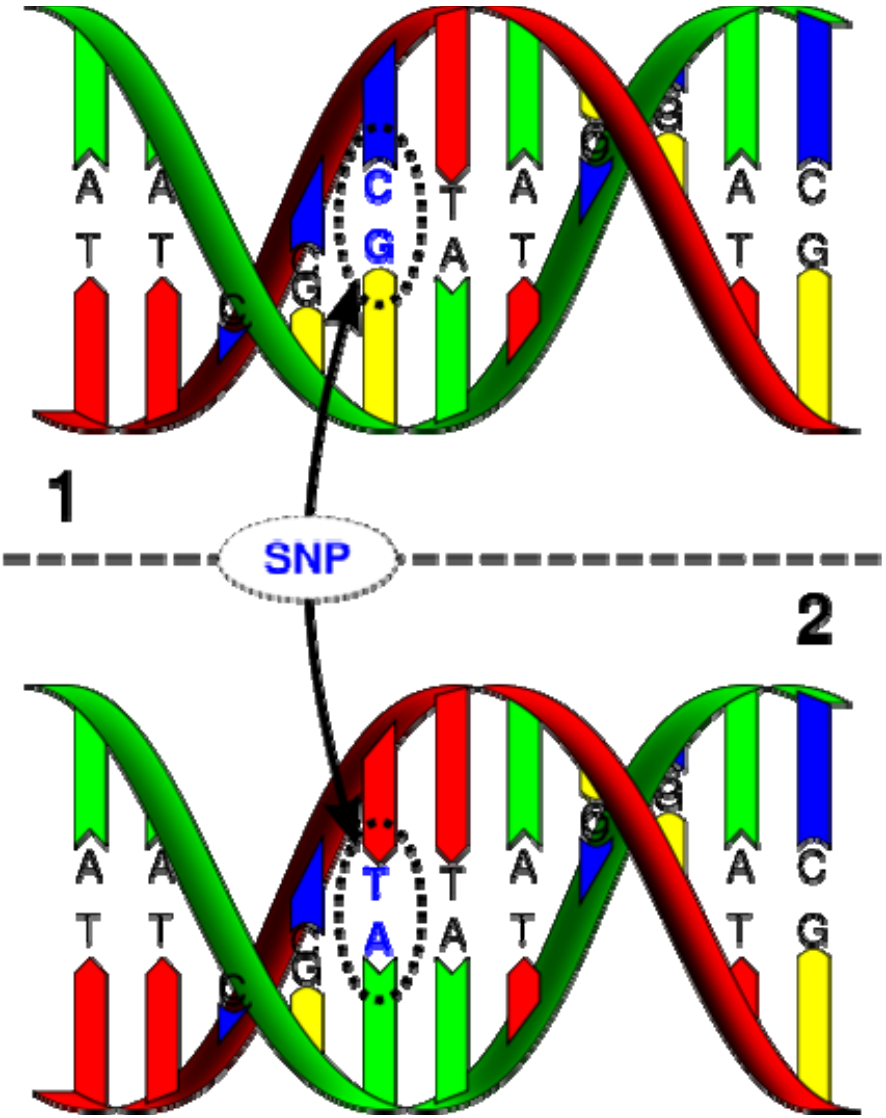
➥Current research relate diseases to SNPs

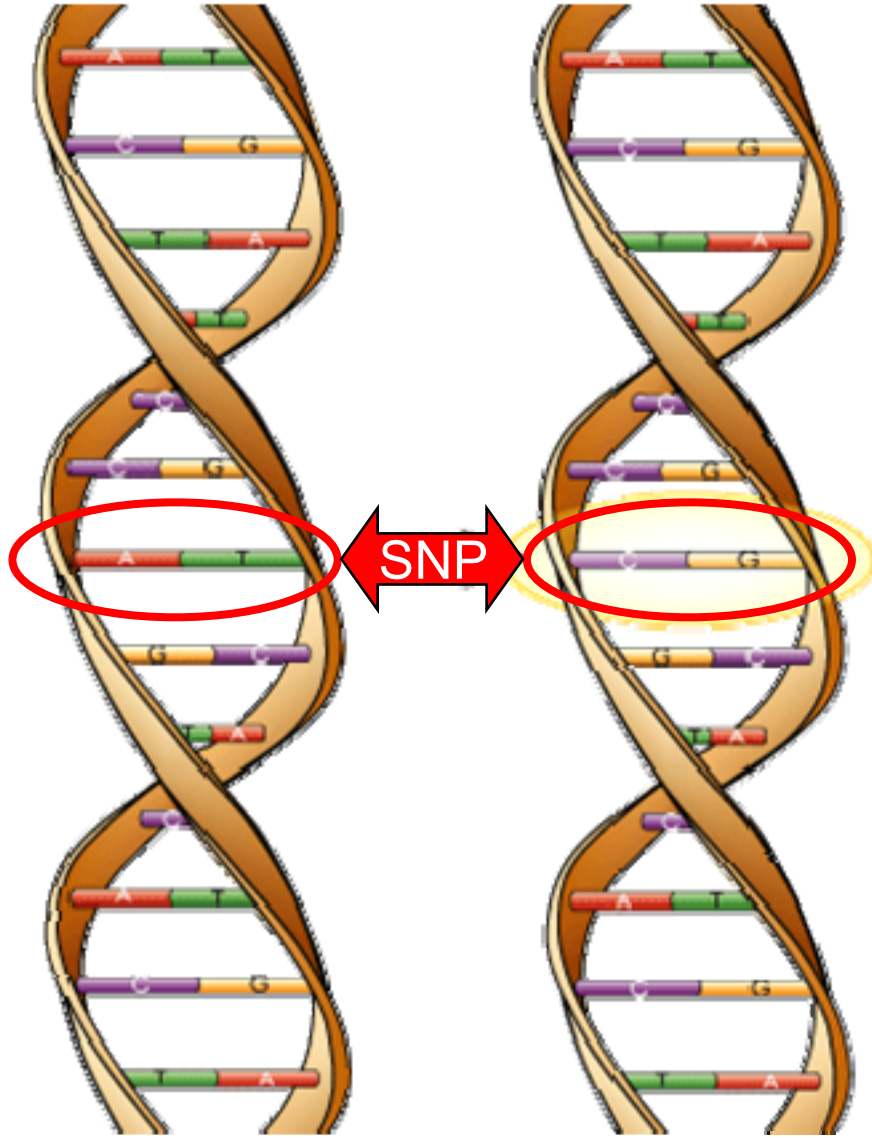
➥Each position exist twice in the human genome: heterozygous



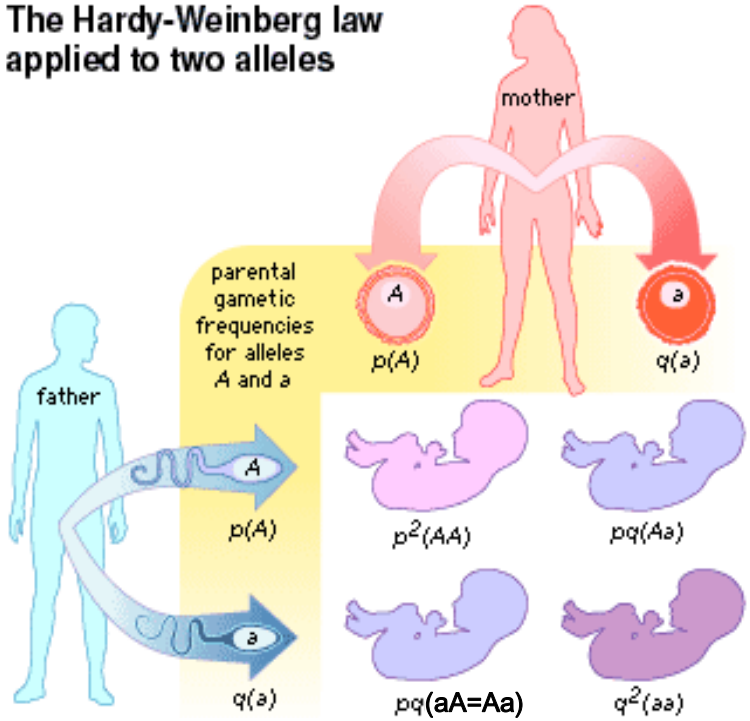
chromosomes of a human (male): 44 pairs and 2 sex

Single Nucleotide Polymorphism





Single Nucleotide Polymorphism



Hardy–Weinberg principle: both allele and genotype frequencies in a population remain constant from generation to generation

Example: mice

AA, Aa → white coat

aa → gray coat

A's: 80%
a's: 20%
} allele frequency

Random mating:
64% **AA** homozygous (0.8*0.8 = 0.64)
32% **Aa** heterozygotes (0.8*0.2*2 = 0.32)
4% **aa** homozygous (0.2*0.2 = 0.04)

96% white coats
4% gray coats
} genotype frequency

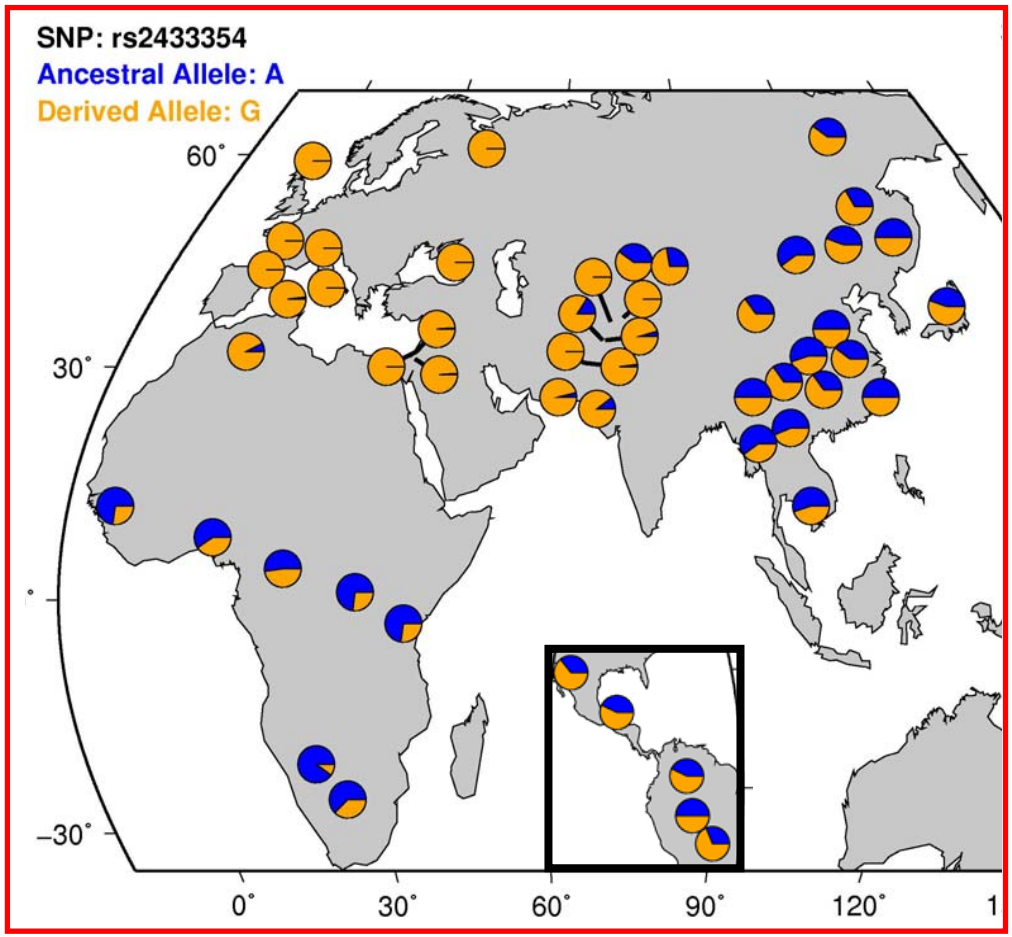
gray coated mice will disappear?

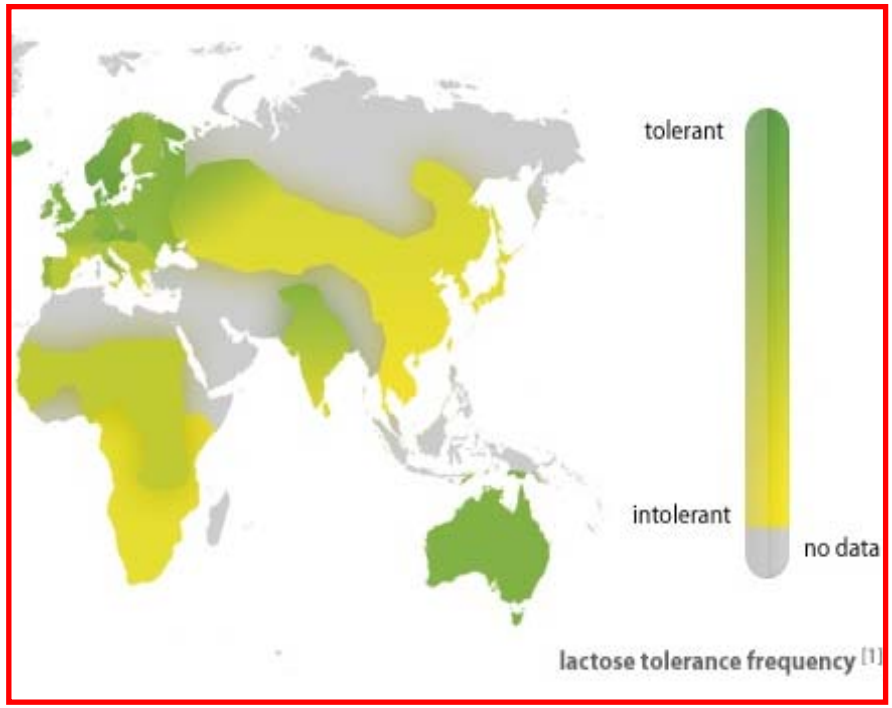
No!

A's: 80% (0.64 + .5*0.32)
a's: 20% (0.04 + .5*0.32)
} allele frequency

Single Nucleotide Polymorphism

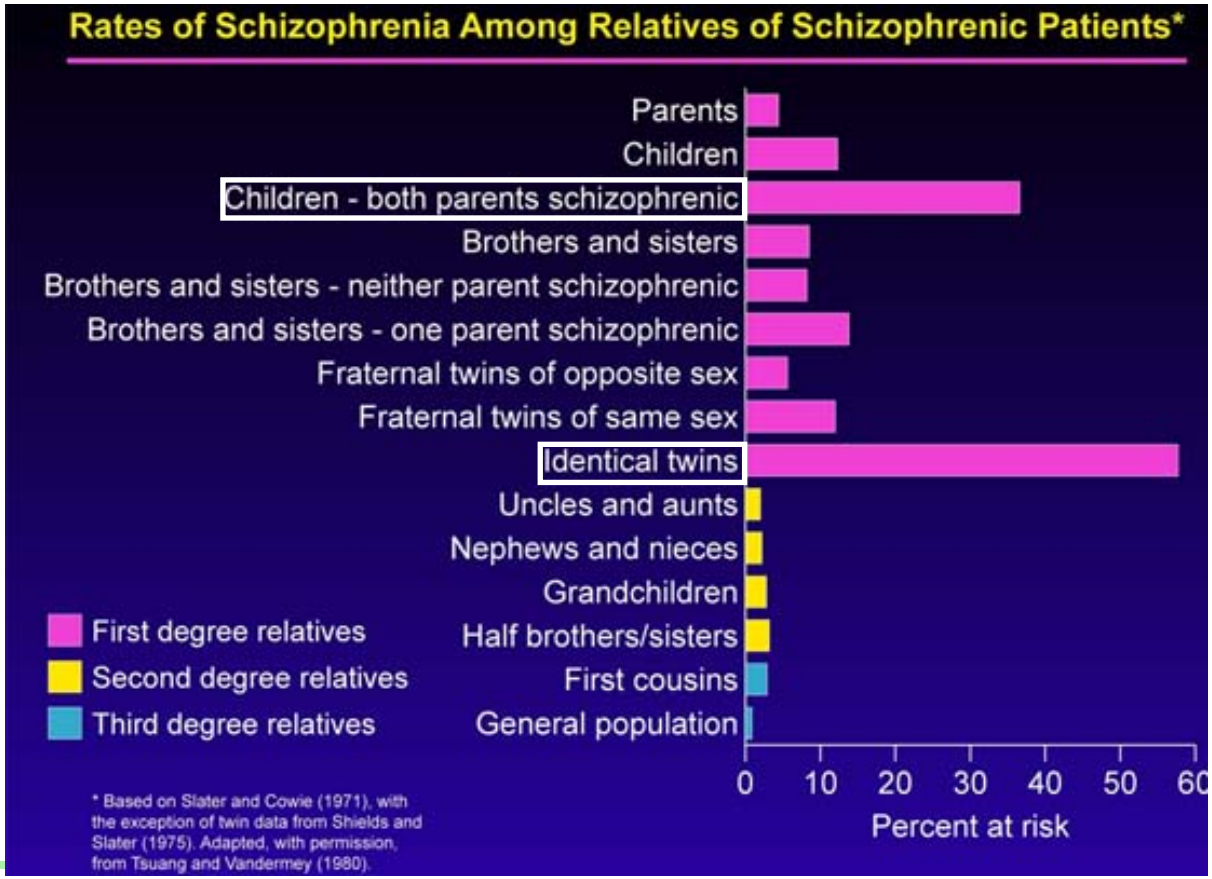
Lactose SNP



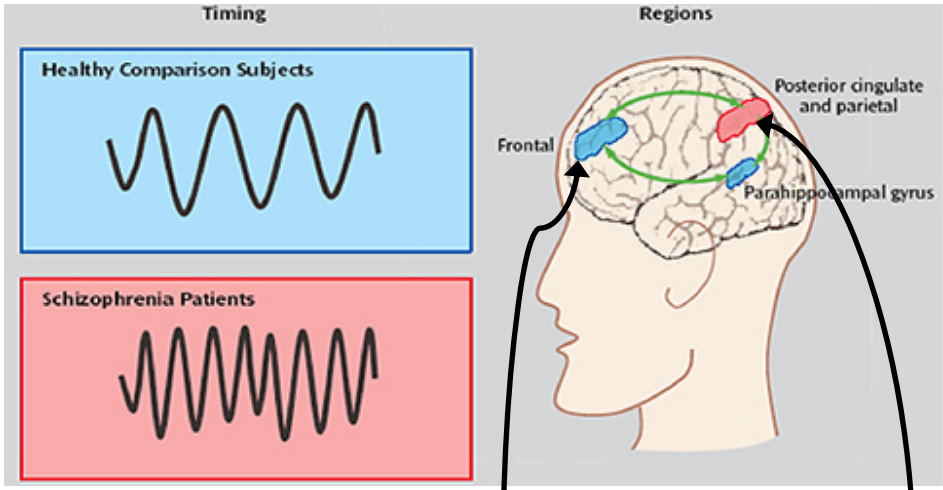


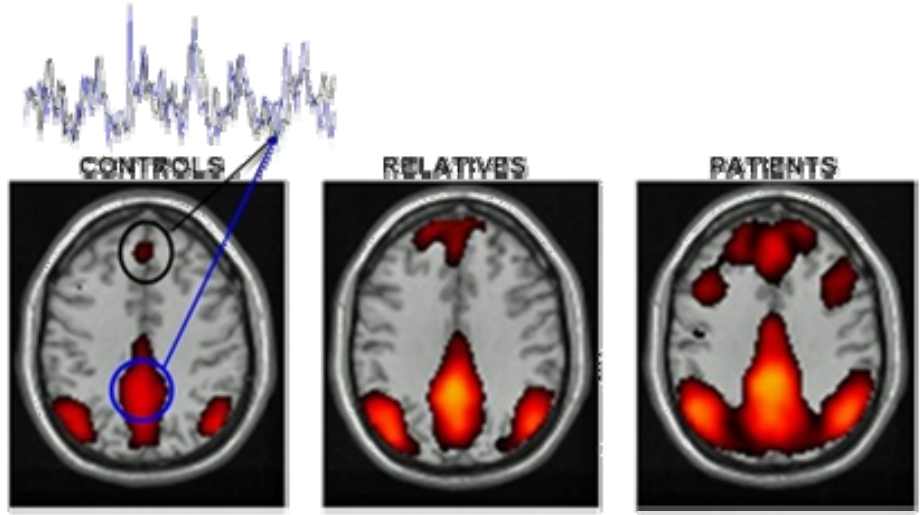
Single Nucleotide Polymorphism: Schizophrenia

Schizophrenia is associated with genetic factors



Single Nucleotide Polymorphism: Schizophrenia



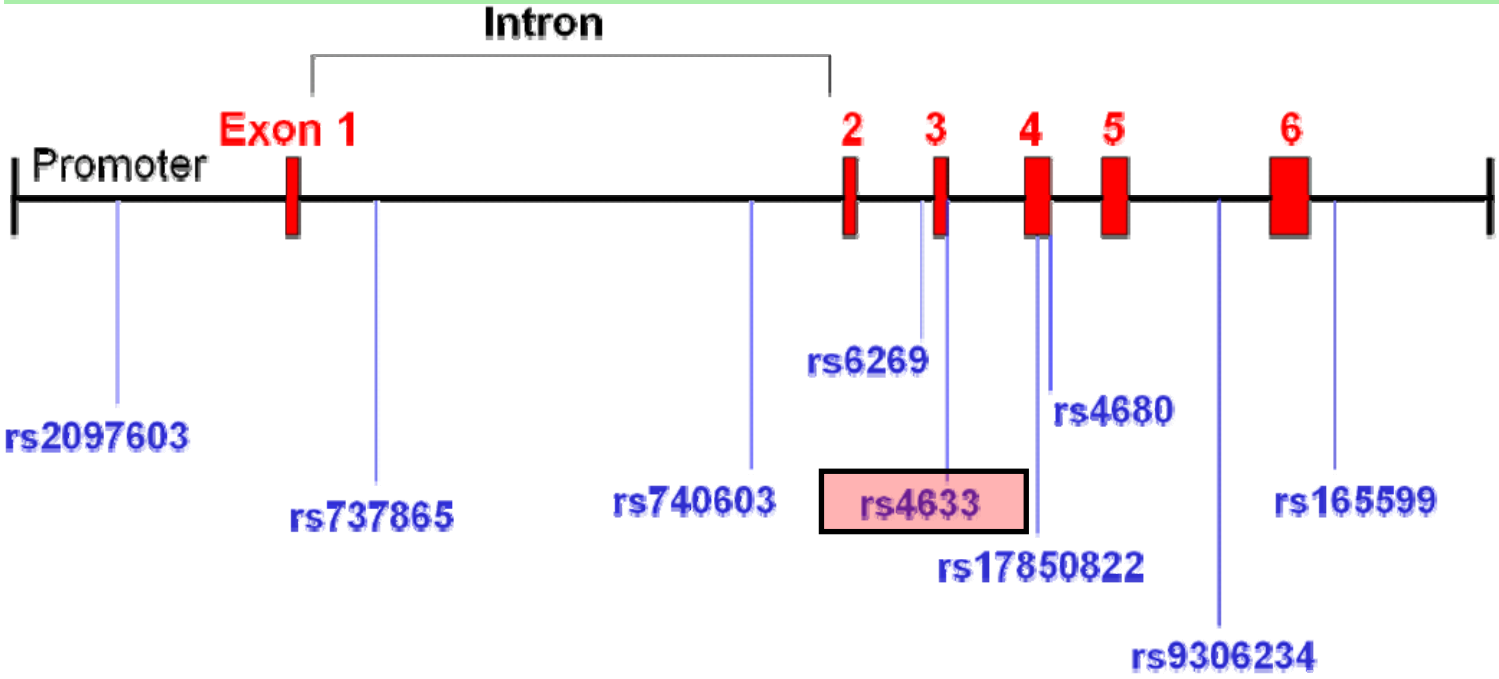


- medial prefrontal cortex and posterior cingulate cortex: associated with self-reflection

- increased connectivity between these regions:
  lost in their own world →trouble performing tasks →schizophrenic

- "People normally suppress this default system when they perform challenging tasks, but we found that patients with schizophrenia don't do this"
  (John D. Gabrieli)

Single Nucleotide Polymorphism: Schizophrenia



- catechol-O-methyltransferase (COMT) gene
- Red: exons (coding sequence)
- Blue: SNPs
- promoter region: expression regulating sequence

rs4633-schizophrenia confirmed by:

- Wang et al., Biological Psychiatry, Jan. 2009
- Gupta et al., Pharmacogenomics, Feb. 2009

Microarray: Preprocessing

Johnson & Johnson

JANSSEN PHARMACEUTICA

Beerse, Belgium

Microarray: Preprocessing

universiteit hasselt

University
Hasselt
Statistics

KATHOLIEKE UNIVERSITEIT LEUVEN

Univ.
Leuven
Statistics

Microarray: Preprocessing

Gene Network Science

http://www.gnsbiotech.com/static_content/johannes-kepler.html

“GNS is collaborating with researchers at Johannes Kepler University, Linz, by way of genomics data processing algorithms developed by the researchers and licensed by GNS.”

Microarray: Preprocessing

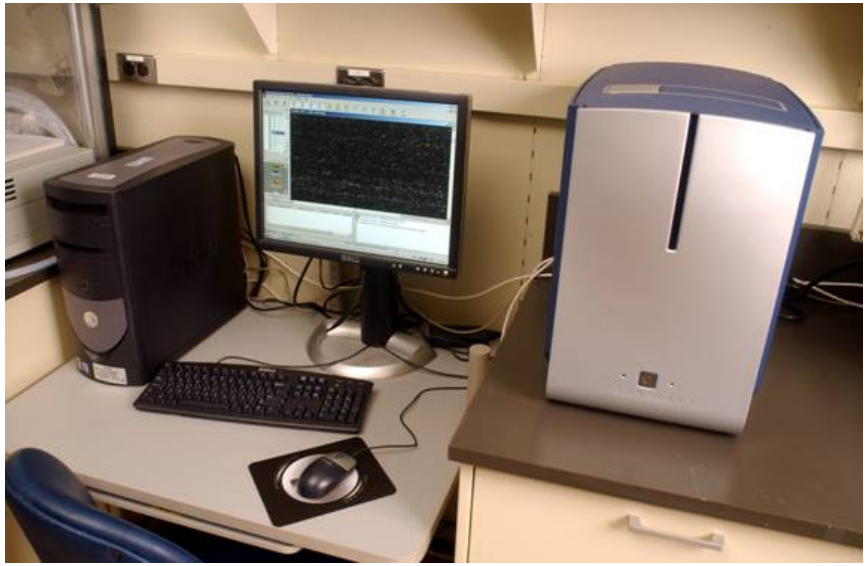
Affymetrix

Fluidics station

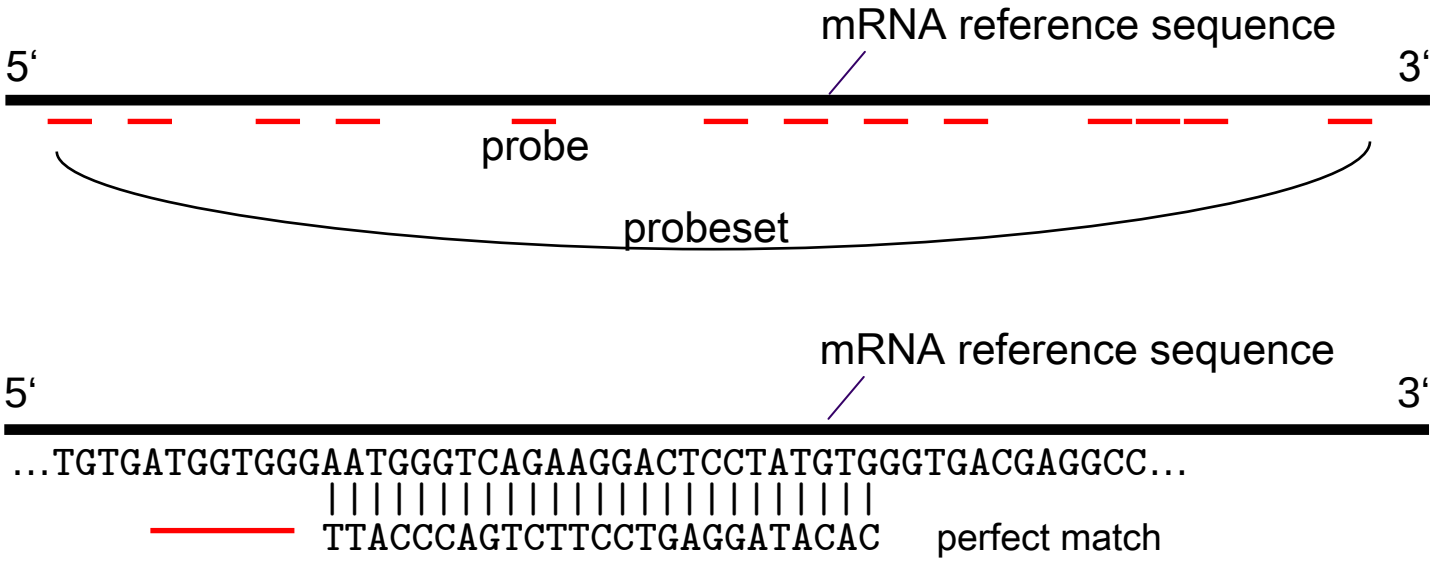
Wash / Stain

Affymetrix

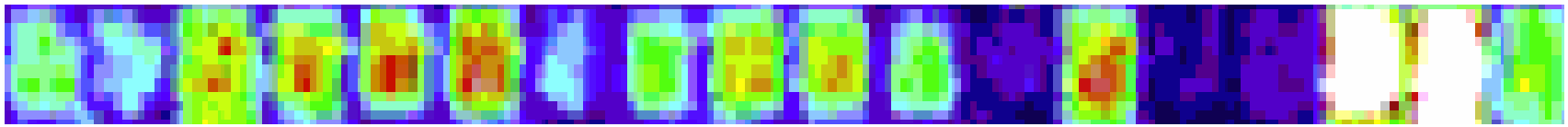
Scanner



Microarray: Preprocessing

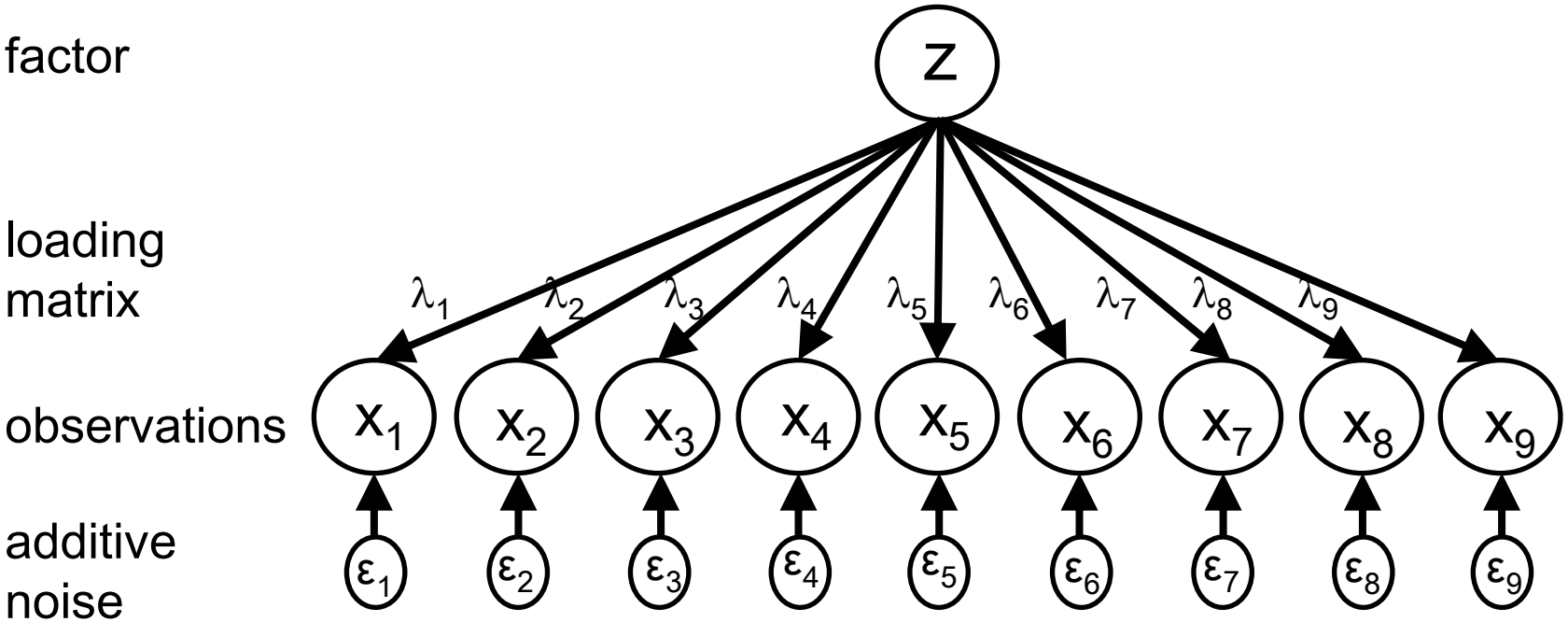


Fluorescence intensity image



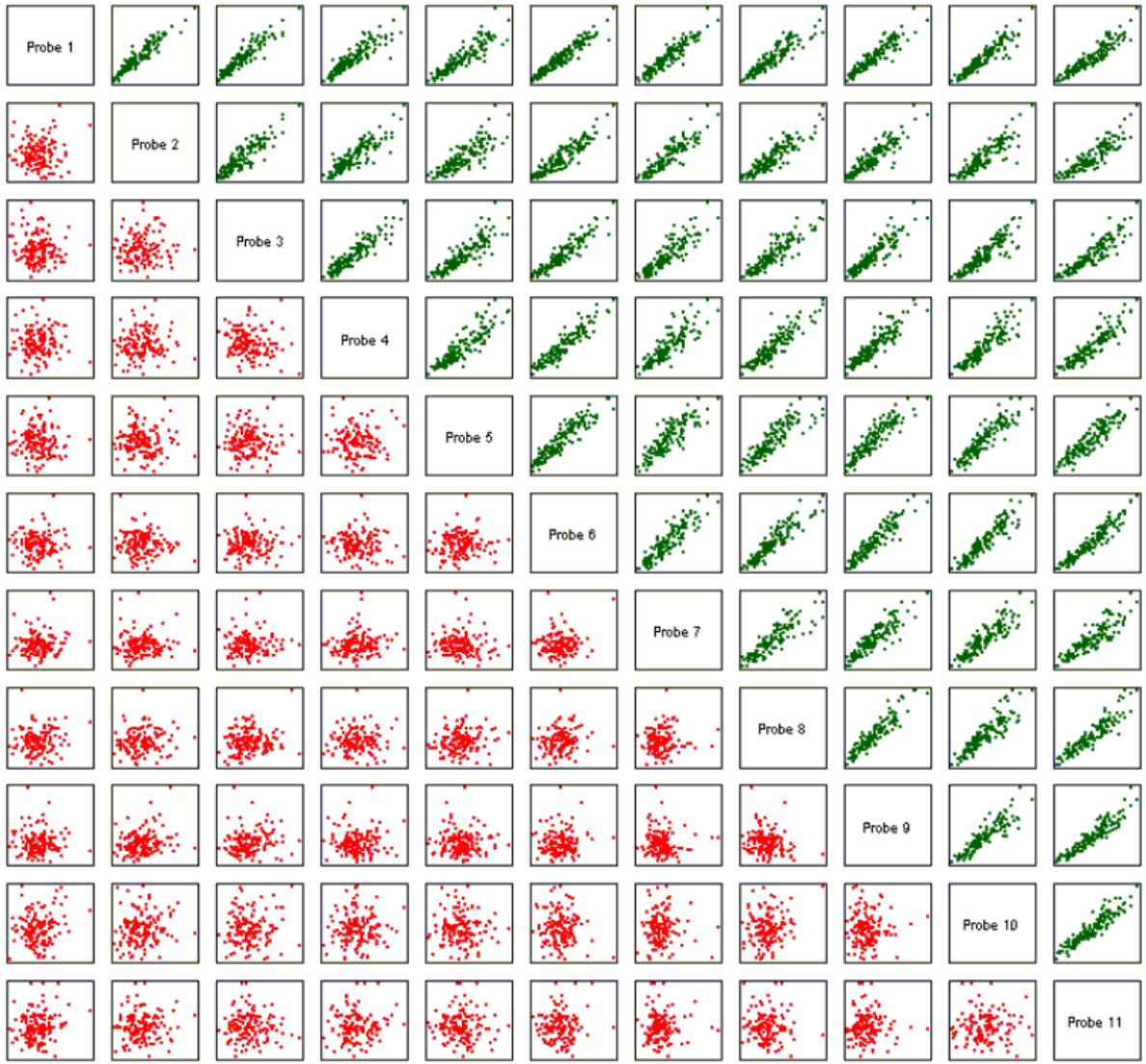
Microarray: Preprocessing

Our new approach based on factor analysis: FARMS
(Factor Analysis for Robust Microarray Summarization)



z = variation in mRNA concentration
λ_i = sensitivity of log-PM_i
ε_i = measurement noise for log-PM_i
x_i = observed log-PM_i

Microarray: Preprocessing



Microarray: Preprocessing

$$x = \lambda z + \varepsilon$$

Generative model:

- z: factor $\mathcal{N}(0,1)$
- ε: noise $\mathcal{N}(0, \Psi)$
- λ: loading vector
- x: data $\mathcal{N}(0, \lambda\lambda' + \Psi)$
- Ψ, $\lambda \rightarrow$ EM-algorithm

Ψ is diagonal
Covariance x: $\lambda\lambda' + \Psi$

Correlations between probes can only be explained by hidden factor

Feature selection: minimal

$$\mathrm{var}(z \mid \boldsymbol{x}) \;=\; \left(1 \;+\; \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}\right)^{-1}$$

Maximum a posterior:

- Data: $\{\boldsymbol{x}\} \;=\; \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$
- Posterior:
  $p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \;\propto\; p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\; p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$
- Likelihood:
  $$p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{0}\;,\; \boldsymbol{\lambda}\boldsymbol{\lambda}^T \;+\; \boldsymbol{\Psi}\right)(\boldsymbol{x}_i)$$
- Prior:
  $p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$

Microarray: Preprocessing

E-step

$$\mathrm{E}_{z_i|\boldsymbol{x}_i}(z_i) = \mu_{z_i|\boldsymbol{x}_i}$$

$$\mathrm{E}_{z_i|\boldsymbol{x}_i}(z_i^2) = \mu_{z_i|\boldsymbol{x}_i}^2 + \sigma_{z_i|\boldsymbol{x}_i}^2$$

$$\mu_{z_i|\boldsymbol{x}_i} = (\boldsymbol{x}_i)^T \left(\boldsymbol{\Lambda}\,\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Lambda}$$

$$\sigma_{z_i|\boldsymbol{x}_i}^2 = 1 - \boldsymbol{\Lambda}^T \left(\boldsymbol{\Lambda}\,\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Lambda} = \left(1 + \boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}\right)^{-1}$$

M-step

$$\lambda_j^{\mathrm{Gauss}} = \left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\,\mathrm{E}_{z_i|\boldsymbol{x}_i}(z_i) + \frac{1}{N}\frac{\mu_\Lambda\,\Psi_{jj}^{\mathrm{new}}}{\sigma_\Lambda^2}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}_{z_i|\boldsymbol{x}_i}(z_i^2) + \frac{1}{N}\frac{\Psi_{jj}^{\mathrm{new}}}{\sigma_\Lambda^2}\right)^{-1}$$

$$\lambda_j^{\mathrm{new}} = \begin{cases} \lambda_j^{\mathrm{Gauss}} & \text{for} \quad \lambda_j^{\mathrm{Gauss}} > 0 \\ 0 & \text{for} \quad \lambda_j^{\mathrm{Gauss}} \leq 0 \end{cases}$$

$$\Psi_{jj}^{\mathrm{new}} = \left[\mathrm{diagvect}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\right]_j - \lambda_j^{\mathrm{new}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}_{z_i|\boldsymbol{x}_i}(z_i)\,\boldsymbol{x}_i\right]_j + \frac{1}{N}\frac{\Psi_{jj}^{\mathrm{new}}}{\sigma_\Lambda^2}\lambda_j^{\mathrm{new}}\left(\mu_\Lambda - \lambda_j^{\mathrm{new}}\right)$$

Microarray: Preprocessing

International competition: “Affycomp” (http://affycomp.biostat.jhsph.edu/)

81	*dfwfcb* / **zhongxue**	0.00	0.00	0.00	0.63	0.31	0.09	0.32	0.26	0.29	0.12	1.00	***1.00***	***1.00***	1.00
88	*en.FARMS* / **clevert**	0.00	0.00	0.00	0.85	0.63	0.16	0.72	0.62	0.65	0.25	***1.00***	***1.00***	***1.00***	***1.00***
89	*VST-RMA* / **dupan**	0.04	0.07	0.23	0.80	0.49	0.12	0.51	0.46	0.47	0.15	0.54	0.96	0.93	0.64
0	*(perfection)*	*0.00*	*0.00*	*0.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
N	**Method / Submitter**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**

AUC (most relevant criterion)

Participants from Berkely, Affymetrix, EBI, Roche, etc.

Johnson & Johnson tested it on 30 internal data sets

→ Now their default microarray normalization method

Microarray: Preprocessing

Extensions FARMS:

- Few arrays differ →few factors differ → factor is Laplace distributed
  Laplace FARMS
  - integral cannot be analytically computed
  - variational approach (from physics, “calculus of variations” → functionals)
  - mixture of Gaussians (“independent factor analysis”, Attias 99)

Laplace

variational: α_i is variance of the local Gaussian approximation of the factor

E-step

$$\sigma^2_{z_i|\boldsymbol{x}_i} = \left(\alpha_i^{-1} + \boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}\right)^{-1}$$

M-step

$$\alpha_i = \sqrt{\mu^2_{z_i|\boldsymbol{x}_i} + \sigma^2_{z_i|\boldsymbol{x}_i}}$$

Gauss

$$\sigma^2_{z_i|\boldsymbol{x}_i} = \left(1 + \boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}\right)^{-1}$$

Microarray: Preprocessing

Extensions FARMS:

- **Conditional FARMS** (factor depends on external input)

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{B}\,\boldsymbol{u} + \boldsymbol{v} & \boldsymbol{v} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{x} &= \boldsymbol{\Lambda}\,\boldsymbol{z} + \epsilon & \epsilon &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned}$$

$$\boldsymbol{z} \mid \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{B}\,\boldsymbol{u}, \boldsymbol{\Sigma})$$

Update Rules:

$$\boldsymbol{\Lambda} = \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\, \mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i^T\right)\right)\left(\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i\boldsymbol{z}_i^T\right)\right)^{-1}$$

$$\boldsymbol{B} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i\right)\boldsymbol{u}_i^T\right)\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}_i\boldsymbol{u}_i^T\right)^{-1}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i\,\boldsymbol{z}_i^T\right) - \boldsymbol{B}\,\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}_i\,\mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i^T\right)$$

$$\boldsymbol{\Psi} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{\Lambda}\,\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{u}_i}\left(\boldsymbol{z}_i\right)\boldsymbol{x}_i^T$$

Classification

Given

- Tissue samples (Tumor, Blood, …)
- Therapy outcome, clinical features, tumor type or state

Goals:

- Diagnosis (tumor type)
- Prognosis of the therapy outcome (alternative therapy, medication)
- Identification of relevant genes (drug design)

Working hypothesis

Gene expression profile → cell state → tumor state → prediction

Prediction: Support Vector Machine and Feature Selection

Classification

Angio-
genesis

and

Metastasis

Classification

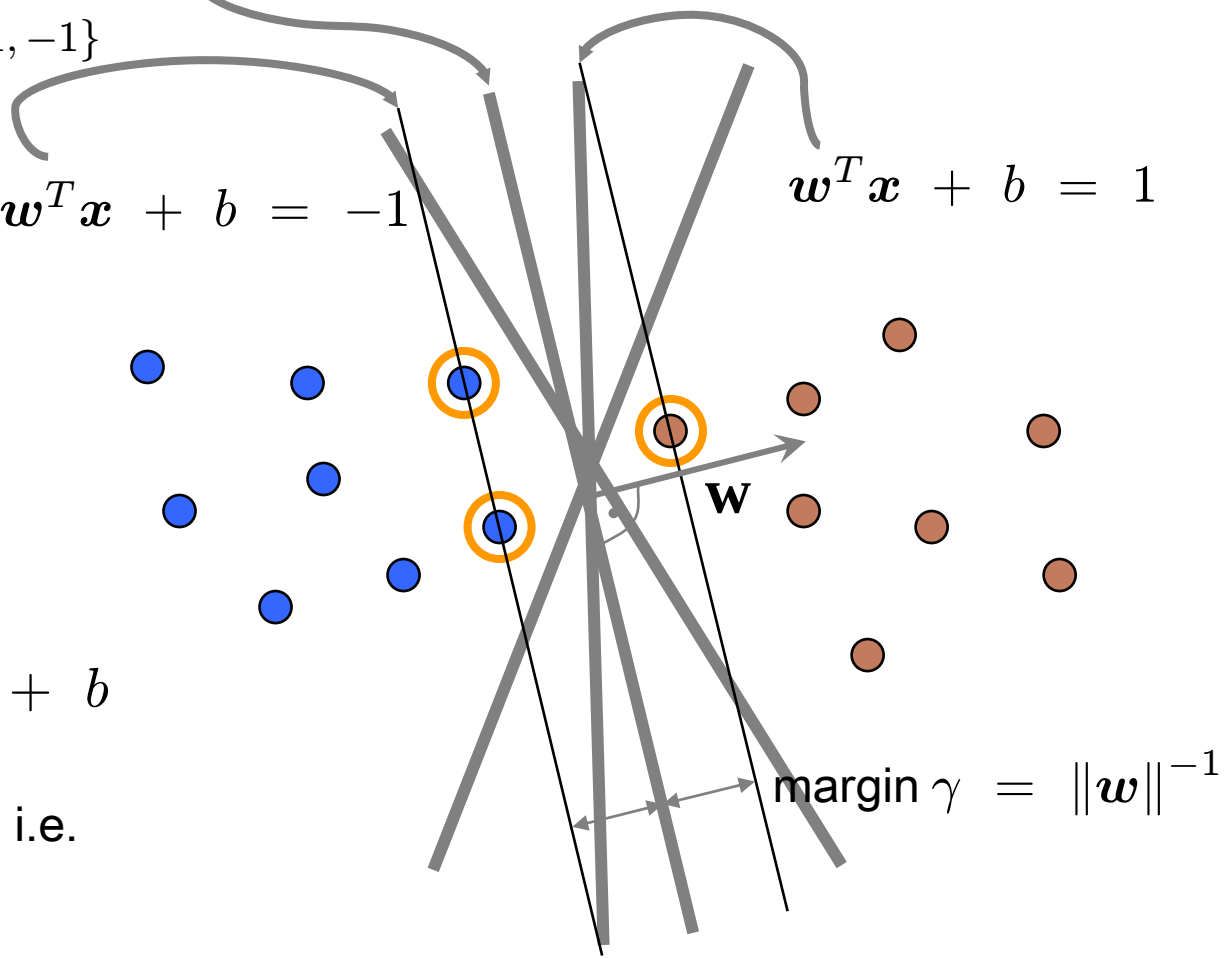
data:

$\{(\boldsymbol{x}^i, y^i)\}\ ,\ 1 \leq\ i\ \leq l$

$\boldsymbol{x}^i\ \in\ \mathbb{R}^n\ ,\ y^i\ \ \in\ \{+1, -1\}$

- class –1
- class +1
- support vectors

search separating hyperplane with maximal margin



linear classifier

$f(\boldsymbol{x})\ =\ \boldsymbol{w}^T\ \boldsymbol{x}\ +\ b$

model selection:
find best classifier, i.e.
optimal $\boldsymbol{w}$ and b

Classification

SVM model selection

primal problem

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\,\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad y^i\,\left(\boldsymbol{w}^T\boldsymbol{x}^i\,+\,b\right)\,\geq\,1$$

dual problem

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j}\alpha_i\,\alpha_j\,y^i\,y^j\,\left(\boldsymbol{x}^j\right)^T\boldsymbol{x}^i\,-\,\sum_{i=1}^{l}\alpha_i$$

$$\text{s.t.} \quad \alpha_i\,\geq\,0$$

$$\sum_{i=1}^{l}\alpha_i\,y^i\,=\,0$$

Lagrangian $\mathcal{L}$ with multipliers α_i

$$\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})\,=\,\frac{1}{2}\,\|\boldsymbol{w}\|^2\,-\,\sum_{i=1}^{l}\alpha_i\,\left(y^i\,\left(\boldsymbol{w}^T\boldsymbol{x}^i\,+\,b\right)\,-\,1\right)$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial\boldsymbol{w}}\,=\,\boldsymbol{w}\,-\,\sum_{i=1}^{l}\alpha_i\,y^i\,\boldsymbol{x}^i\,=\,\mathbf{0}$$

$$\Rightarrow\quad \boxed{\boldsymbol{w}\,=\,\sum_{i=1}^{l}\alpha_i\,y^i\,\boldsymbol{x}^i}$$

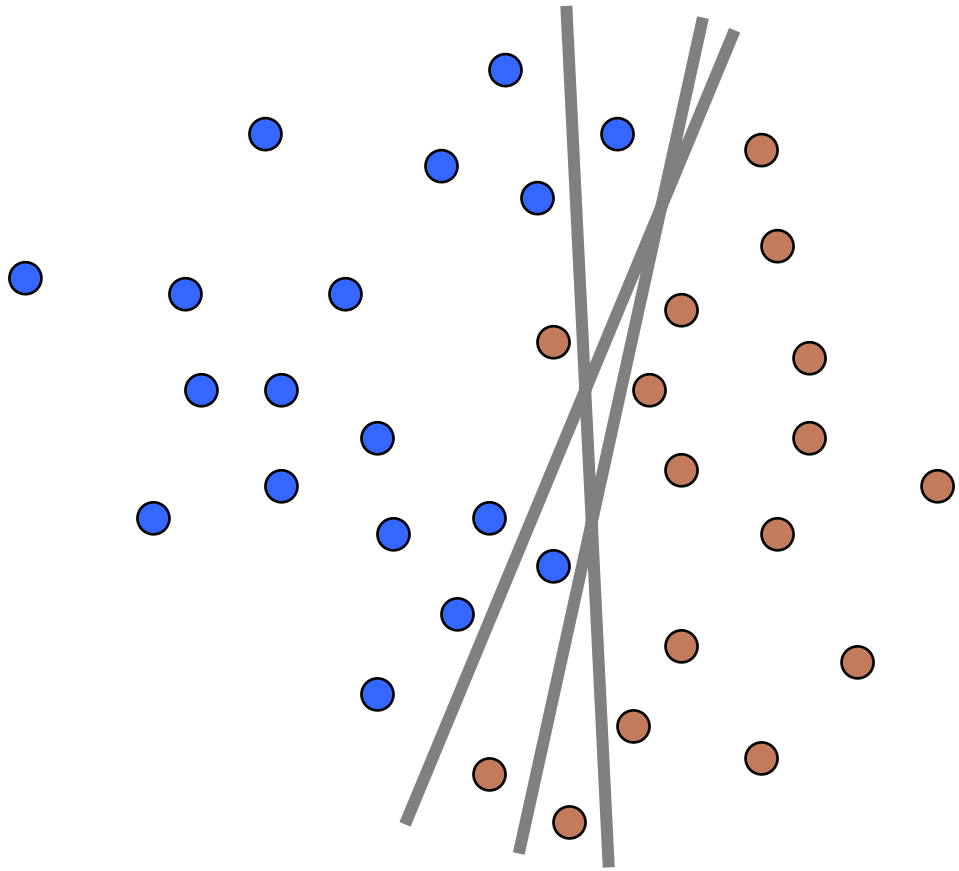
Expansion of $\mathbf{w}$ with respect to the data to classify

$$\frac{\partial\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial b}\,=\,\boxed{\sum_{i=1}^{l}\alpha_i\,y^i\,=\,0}$$

Classification

not linear separable

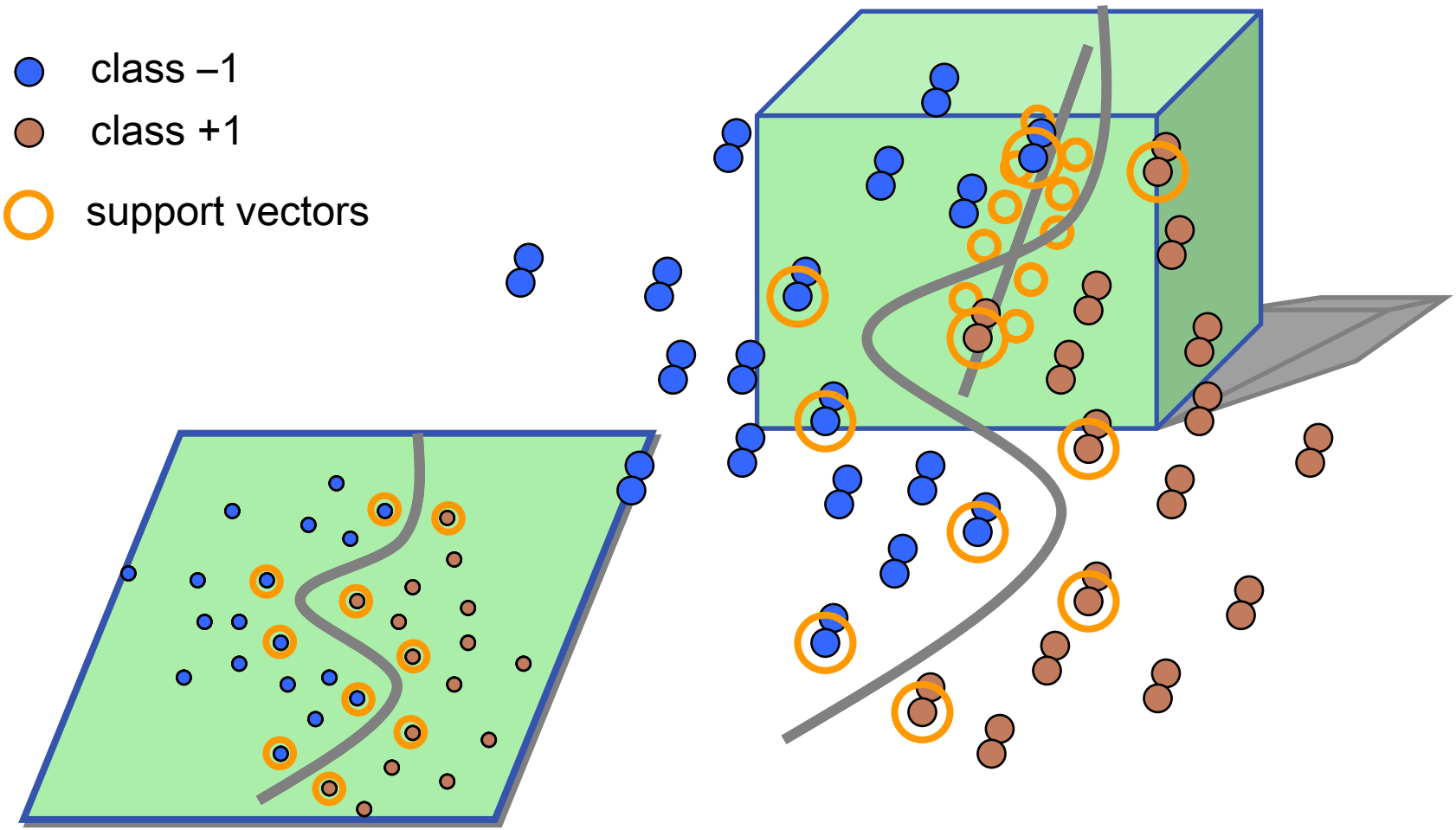
- class –1
- class +1



Classification

non-linear support vector machine

- class –1
- class +1
- support vectors



Classification

Theorem 1 (Mercer)
Let the kernel k be symmetric and from $L_2(X \times X)$ defining a Hilbert-Schmidt operator

$$T_k(f)(\boldsymbol{x}) \;=\; \int_X k(\boldsymbol{x}, \boldsymbol{x}')\; f(\boldsymbol{x}')\; d\boldsymbol{x}' \;.$$

If T_k is positive definite, i.e. for all $f \in L_2(X)$

$$\int_{X \times X} k(\boldsymbol{x}, \boldsymbol{x}')\; f(\boldsymbol{x})\; f(\boldsymbol{x}')\; d\boldsymbol{x}\; d\boldsymbol{x}' \;\geq\; 0 \;,$$

then T_k has eigenvalues $\lambda_j \;\geq\; 0$ with associated eigenfunctions $\psi_j \in L_2(X)$. Further

$$(\lambda_1, \lambda_2, \ldots) \in \ell_1$$

$$k(\boldsymbol{x}, \boldsymbol{x}') \;=\; \sum_j \lambda_j\; \psi_j(\boldsymbol{x})\; \psi_j(\boldsymbol{x}') \;,$$

where ℓ_1 is the space of vectors with finite one-norm and the last sum converges absolutely and uniformly for almost all $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Feature Selection

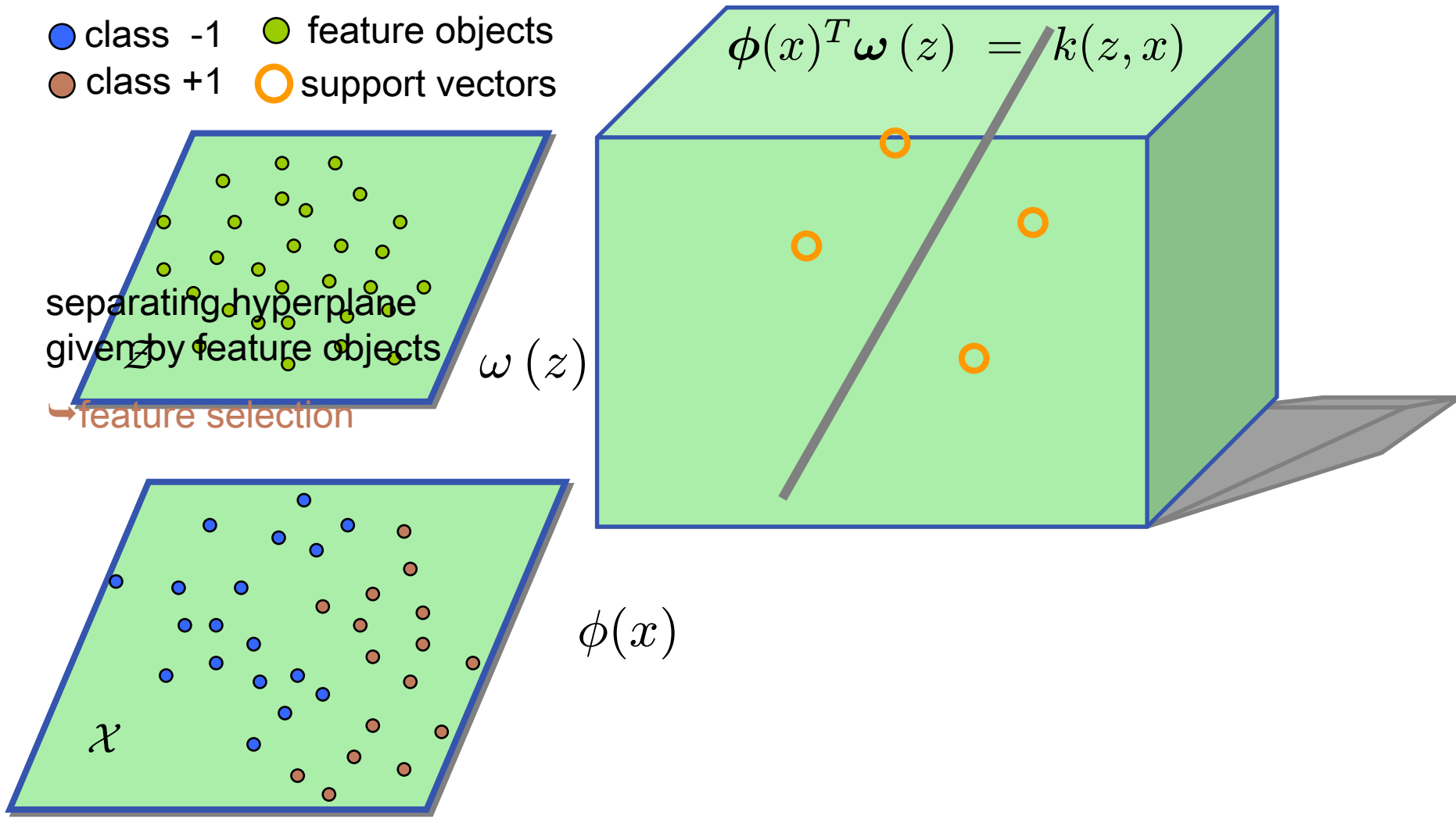
Problem

- Many genes (several 10,000 features) but few samples (about 100 examples)
- SVM theory requires more examples (L) then features (n) for a low error bound on future data (determined by L / n).

Solution

- **Feature selection**: Selection of relevant genes and, therefore, decreasing the input dimension n

Feature Selection



Feature Selection

Theorem 1 (Singular Value Expansion)
Let α be from $H_1 := L^2(\mathcal{Z})$ and let k be a kernel from $H_3 := L^2(\mathcal{X}, \mathcal{Z})$ which defines a Hilbert-Schmidt operator $T_k : H_1 \to H_2$ ($H_2 := L^2(\mathcal{X})$)

$$(T_k\alpha)(x) = f(x) = \int_{\mathcal{Z}} k(x,z)\ \alpha(z)\ dz\ .$$

Then $\|f\|^2_{H_2} = \langle T_k^* T_k \alpha, \alpha\rangle_{H_1}$ (T_k^* adjoint operator of T_k) and

$$k(x,z) = \sum_n s_n\ e_n(z)\ g_n(x)$$

which converges in the L^2-sense. The $s_n \geq 0$ are the singular values of T_k, and $e_n \in H_1$, $g_n \in H_2$ are the corresponding orthonormal functions.
If $\int_{\mathcal{Z}} (k(x,z))^2\ dz \leq K^2$ for all $x \in \mathcal{X}$, then the following sum convergences absolutely and uniformly:

$$f(x) = \sum_n s_n\ \langle \alpha, e_n\rangle_{H_1}\ g_n(x)$$

Feature Selection

$$\min_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \quad \frac{1}{2}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right)^\top \boldsymbol{K}^\top \boldsymbol{K}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) - \boldsymbol{y}^\top \boldsymbol{K}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) + \epsilon\, \mathbf{1}^\top\left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right)$$

$$\text{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^- \leq C\mathbf{1}$$

$\boldsymbol{y}$ is vector of labels
$\boldsymbol{K}$ is kernel matrix
$\mathbf{1}$ is vector of ones
$\boldsymbol{\alpha} = \boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-$
$\boldsymbol{w} = \boldsymbol{Z}\,\boldsymbol{\alpha} = \sum_{j=1}^{l} \alpha_j\, \boldsymbol{z}$

Fast optimization through a new sequential minimal optimization (SMO) technique: only box constraints!

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = \sum_{j=1}^{l} \alpha_j\, \boldsymbol{z}^T\boldsymbol{x} + b = \sum_{j=1}^{l} \alpha_j\, \boldsymbol{K}_{(x)j} + b$$

Microarray: Feature Selection / Classification

Task

Brain tumor (medulloblastoma) patients respond differently to the therapy (chemo, radiation)

➥Negative prognosis: alternative therapy or more intensive control

➥Positive prognosis: reduced medication

60 patients and 7129 genes

S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova and P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, T. R. Golub
Prediction of central nervous system embryonal tumour outcome based on gene expression
Nature 415(687):436-442, 2002

Microarray: Feature Selection / Classification

Classification results

Standard			New method		
Method	F	Error	Method	Features	Error
TrkC (1 gene)	1	20	SVM	40 / 45 / 50	5 / 4 / 5
SVM		15	SVM	40 / 45 / 50	5 / 5 / 5
TrkC & SVM		14	P-SVM	40 / 45 / 50	4 / 4 / 5
KNN	8	13			
KNN & SVM		12			

Standard feature selection with signal-to-noise- and t-statistics

Microarray: Feature Selection / Classification

Task

Breast cancer: 70-80% of the patients do not need a treatment because metastasis does not appear

➥ Prediction of metastasis: therapy selection

➥Alternative treatment or individual medication (toxicity)

78 patients and 25,000 genes

L. J. van't Veer, H, Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend
Gene expression profiling predicts clinical outcome of breast cancer
Nature 415: 530-536, 2002

Microarray: Feature Selection / Classification

Classification results

Standard feature selection					New method				
Method	F	Error	ROC	Test	Method	F	Error	ROC	Test
weighted voting	70	20	0.77	2	SVM	30	12	0.88	2

Standard feature selection with signal-to-noise-statistics

Microarray: Feature Selection / Classification

Signature Diagnostics AG
Prof. Dr. André Rosenthal, CEO

Press release (http://www.signature-diagnostics.de/press10.htm):

> Signature Diagnostics AG announced today that the company has developed a new gene expression classifier, that predicts the risk of recurrence and metastasis (progression of disease) in stage I and II colorectal cancer.

Note: 55% employee patent fee hold by Djork-Arne Clevert, JKU

Diagnosis with Peptide Arrays

CHARITÉ

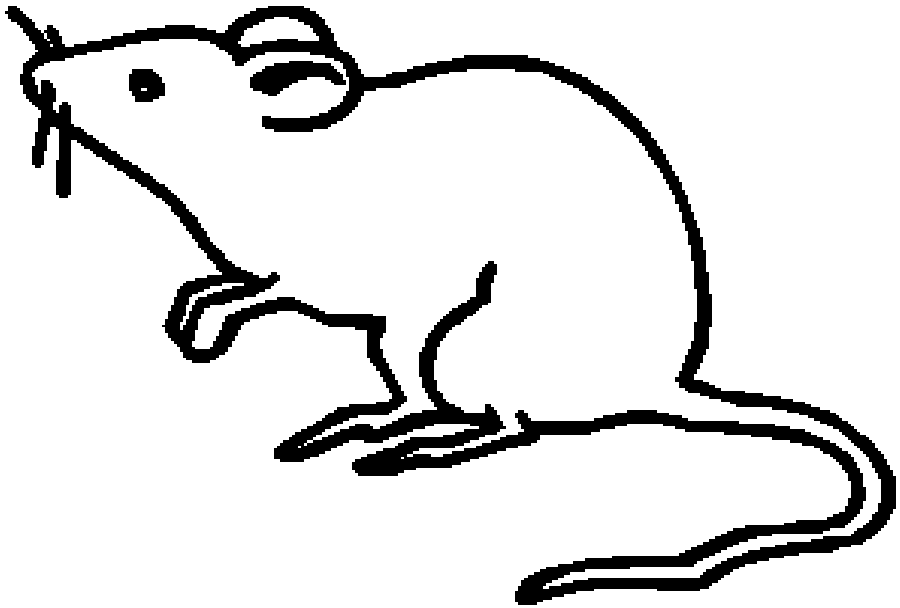
Berlin,
Germany

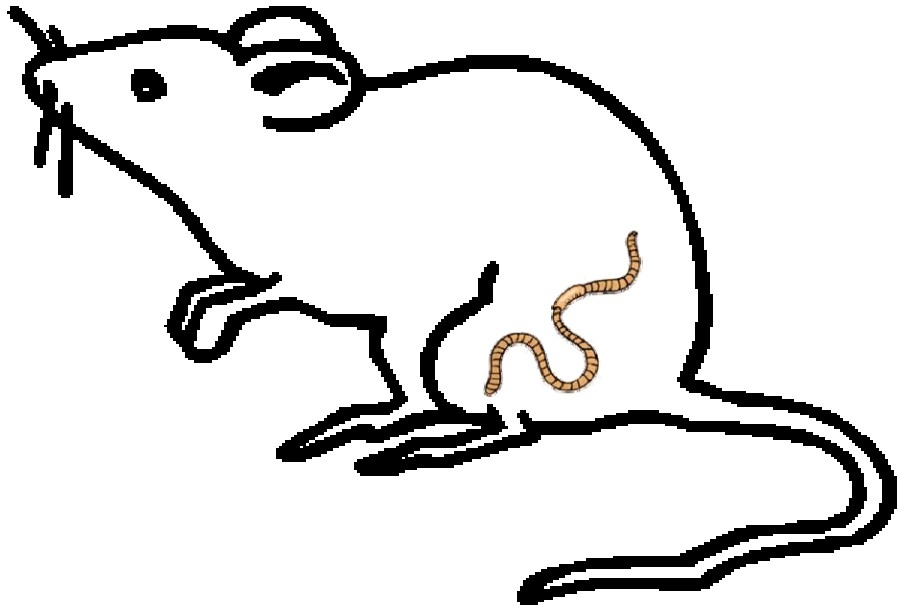
Group:
Institute of theoretical Biology
Systems Immunology
Prof. Michal Or-Guil

Humboldt University

Diagnosis with Peptide Arrays

Mice with and without parasite (*H. polygyrus*, worm)



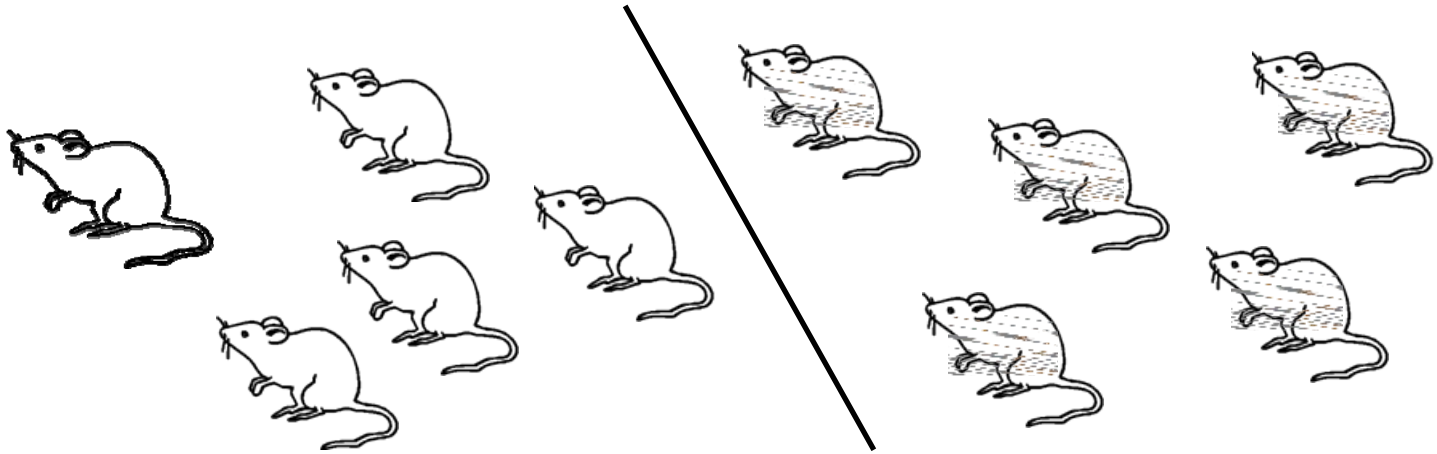


Diagnosis with Peptide Arrays

Given: serum for each mouse containing antibodies

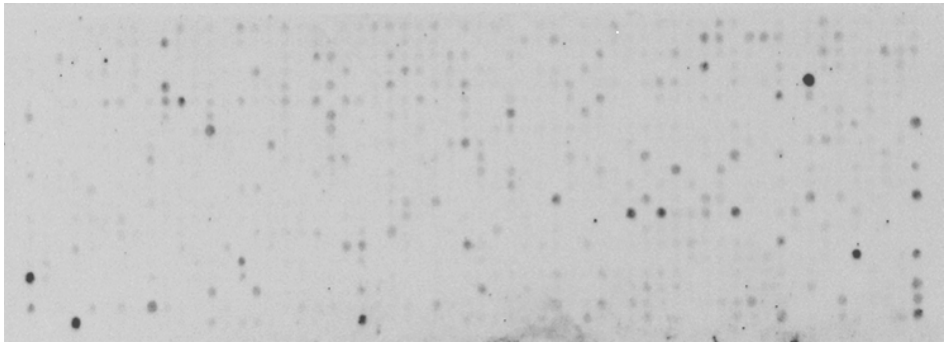
Goal: predict whether the mouse has a parasite

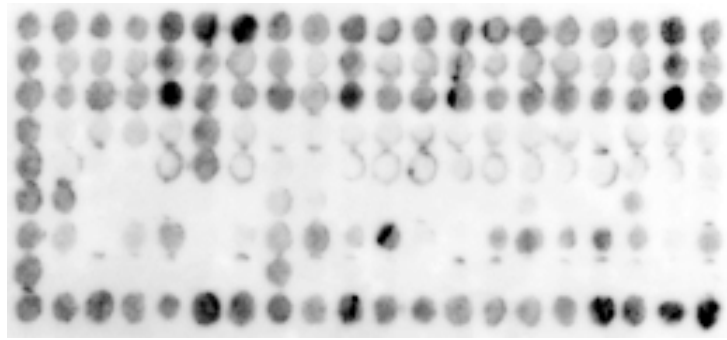
Assumption: Immune system produces antibodies against parasite proteins

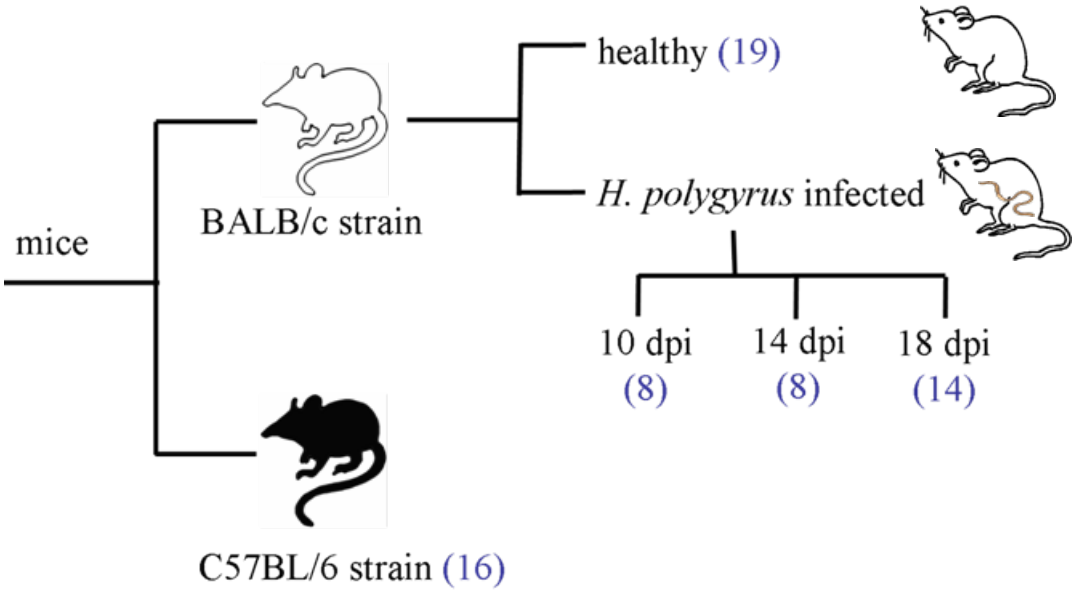


Diagnosis with Peptide Arrays

Antibody detection with random arrays and ML (feature selection and classification)







Diagnosis with Peptide Arrays

Minimal number of peptide features

	IgM	IgG	IgM	IgG	IgM	IgG
Sensitivity	1	1	0.93	0.80	0.93	0.86
Specificity	0.96	0.93	1	0.87	1	1
Diagnostic reliability (%)	**97.7**	90.7	**96.7**	83.4	**95.3**	90.7
Number of features	2	2	6	4	5	4
Significance	<0.001	<0.001	0.004	0.014	<0.001	<0.001

Josef Penninger (Vienna): cancer cells activate the immune system (verified at mice)

→ Diagnosis of cancer by peptide arrays

Copy Number Variation

BIOINF

MERCK SERONO

Geneva, Switzerland

Copy Number Variation

BIOINF

MERCK

Darmstadt,
Germany

Copy Number Variation

complex diseases

e.g.: bipolar disorder

Copy Number Variation

Copy-number variant (CNV):

➥1 kb or larger (50 kb) DNA sequences

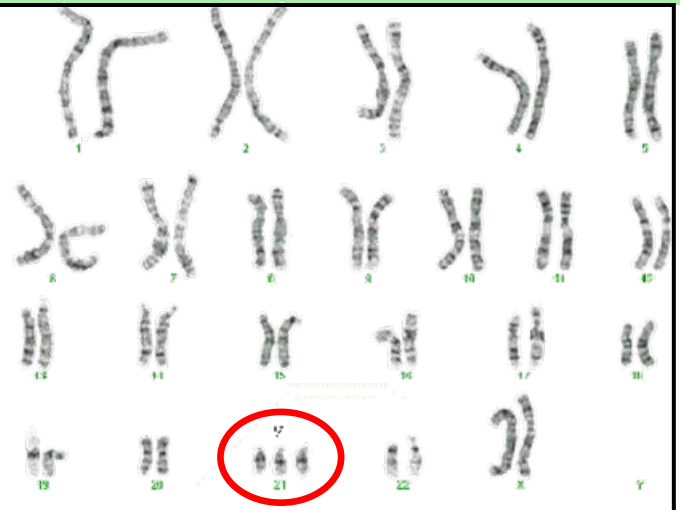
➥variable copy-number compared to a reference

➥insertions, deletions and duplication

➥ 2004 first publications: Sebat et al., Science
Iafrate et al., Nature Genetics

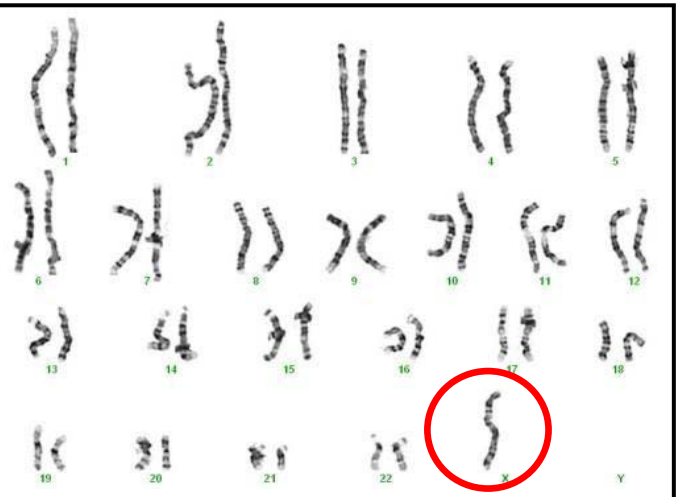
Copy Number Variation

Downs syndrome



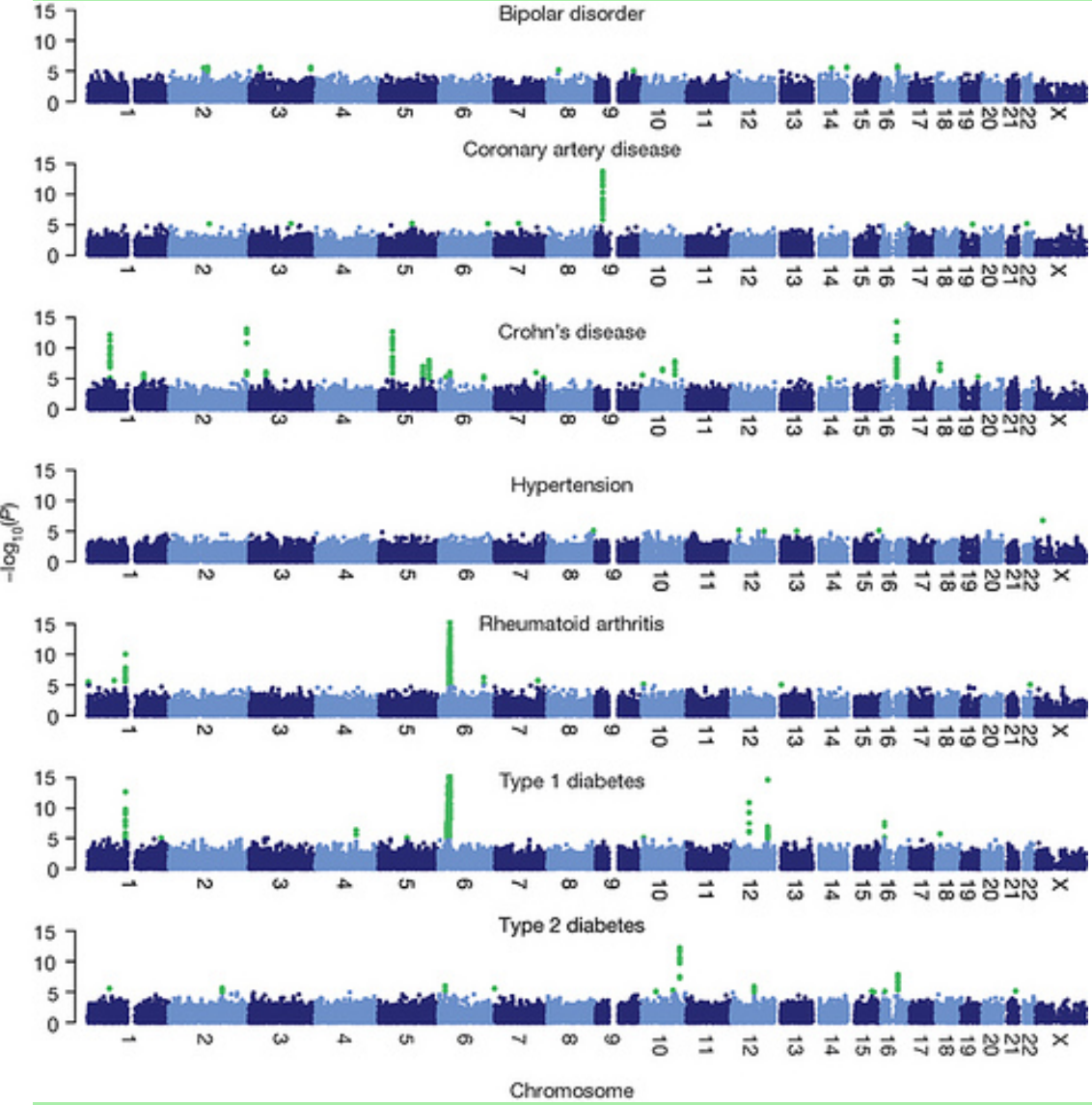
XXY

Monosomy

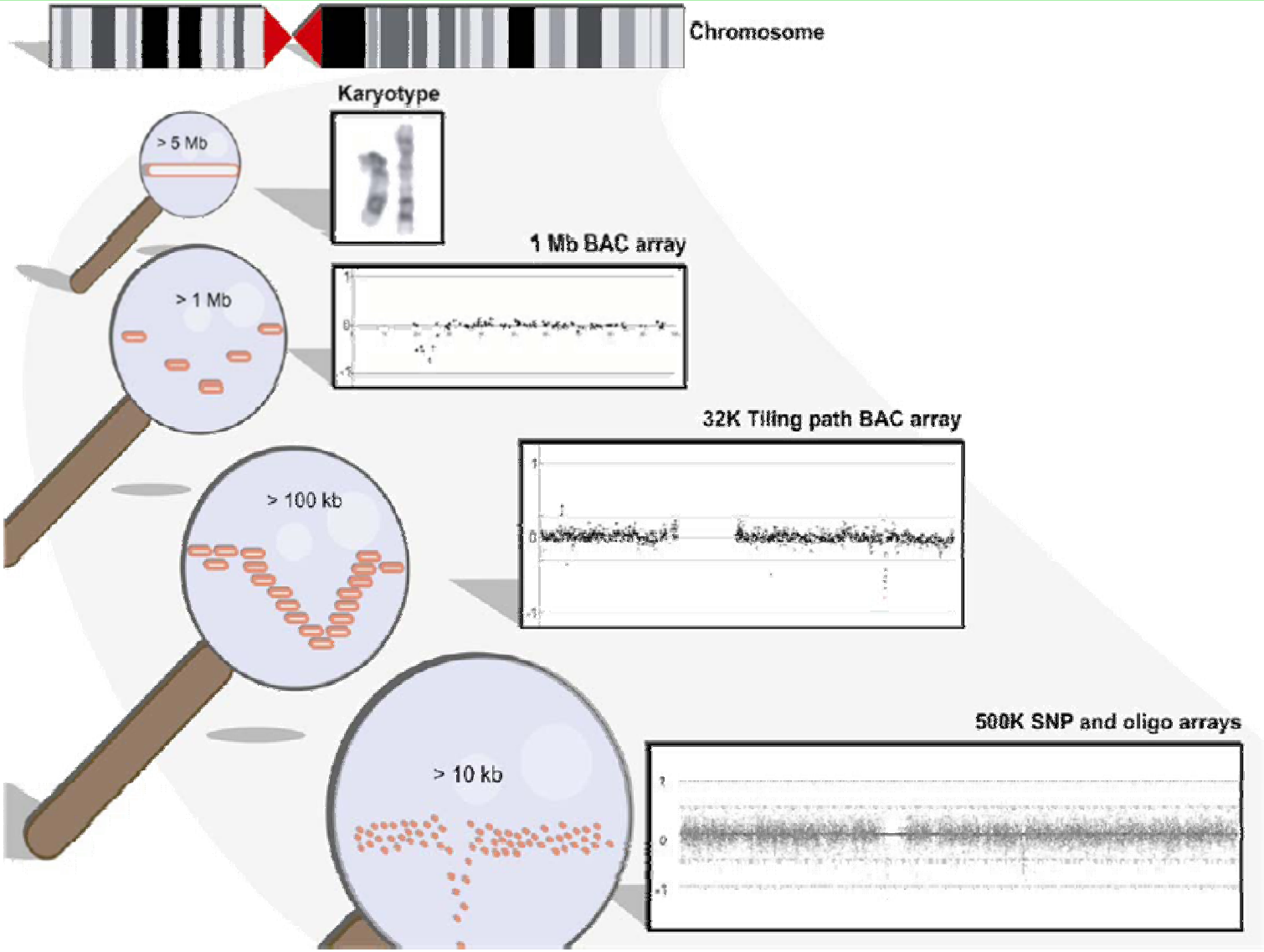


Robertsonian

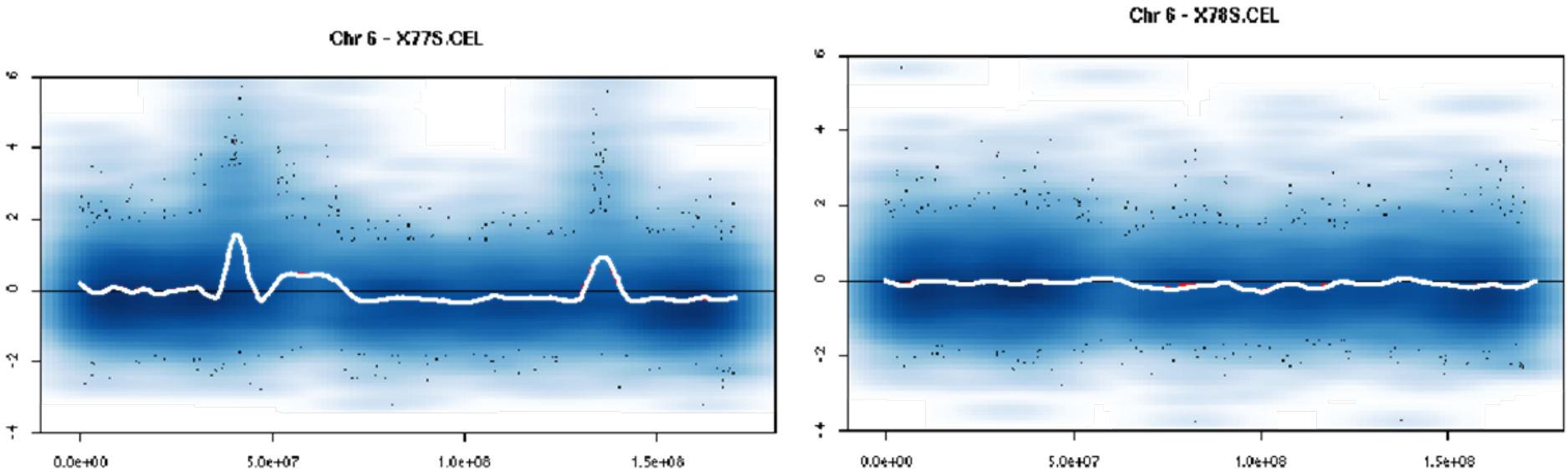
Copy Number Variation



Copy Number Variation



Copy Number Variation

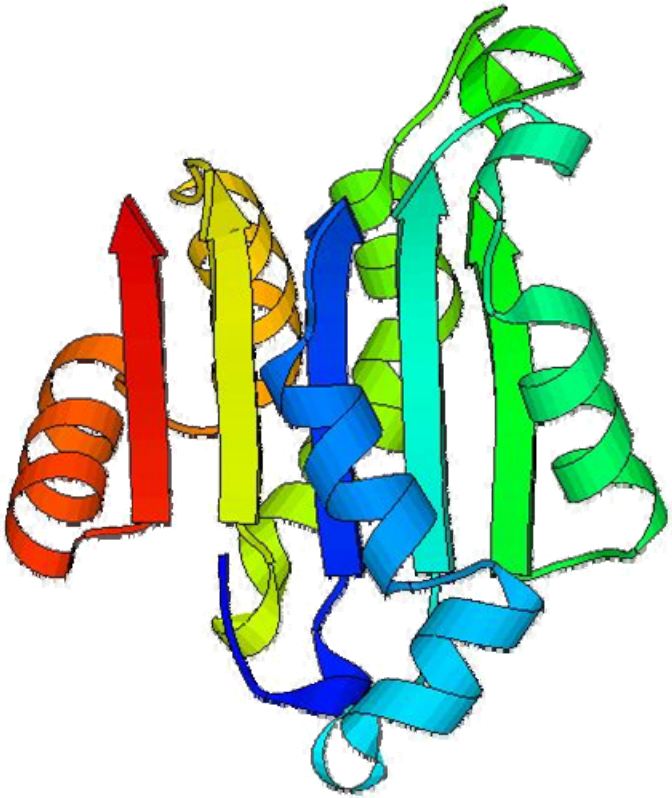


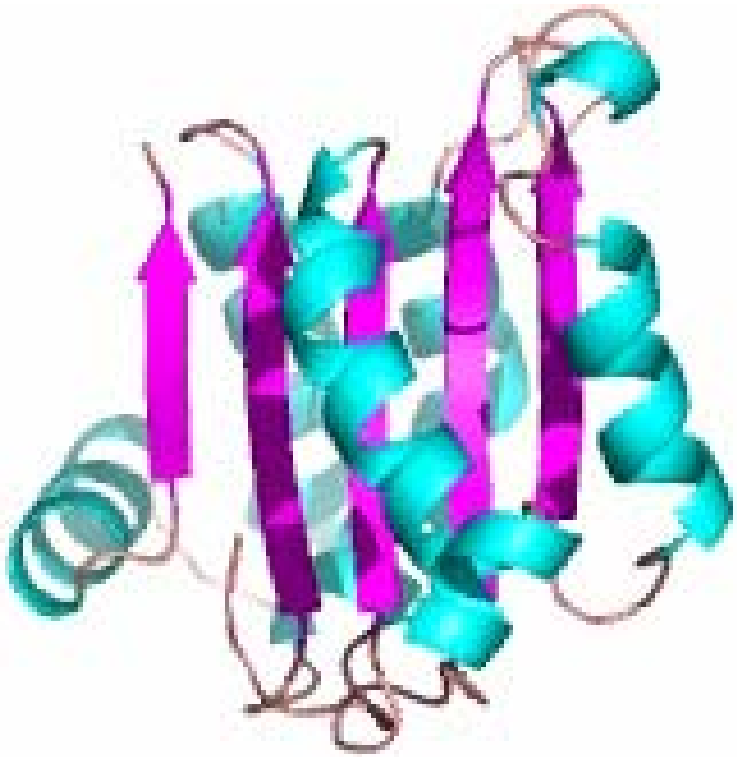
Smooth scatter plots of an Affymetrix SNP 6.0 array for Chr. 6 (using FARMS)

Left: defective amplifications in a sample stemming from lung cancer tissue

Right: the same region corresponding to a healthy cell

Sequence Classification

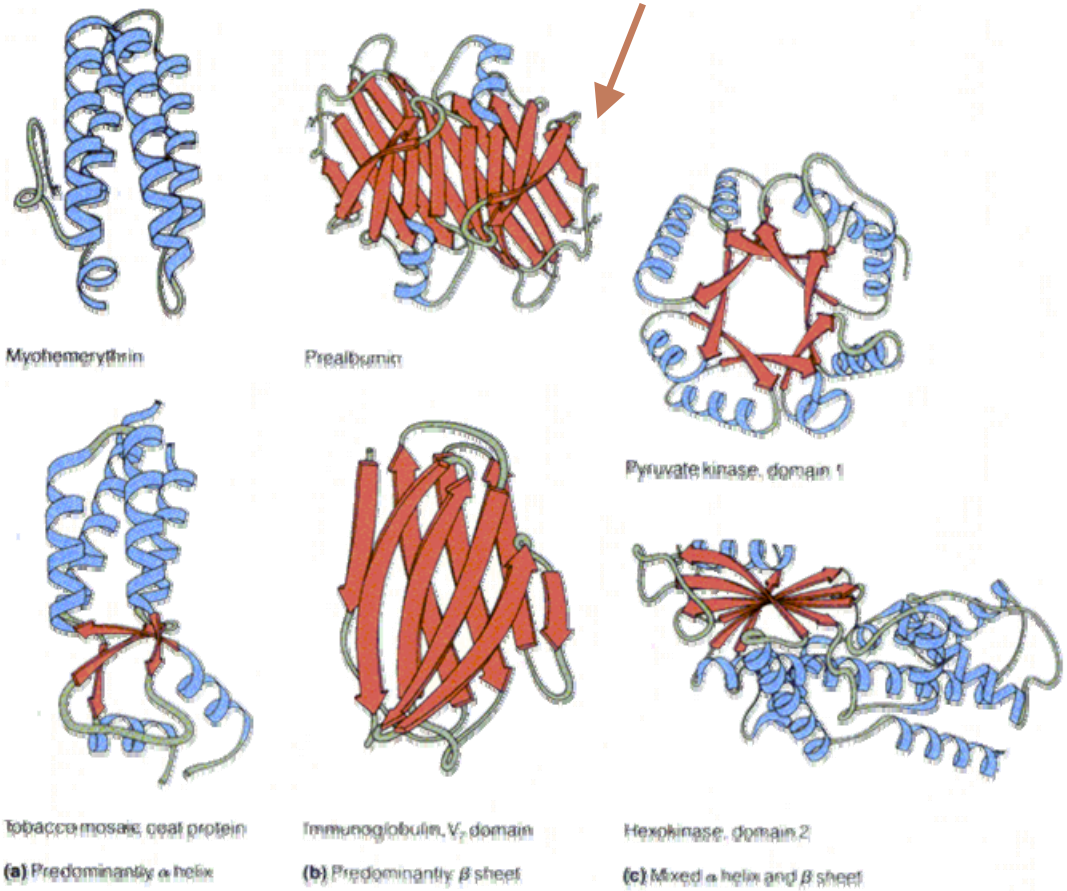




Sequence Classification

The problem

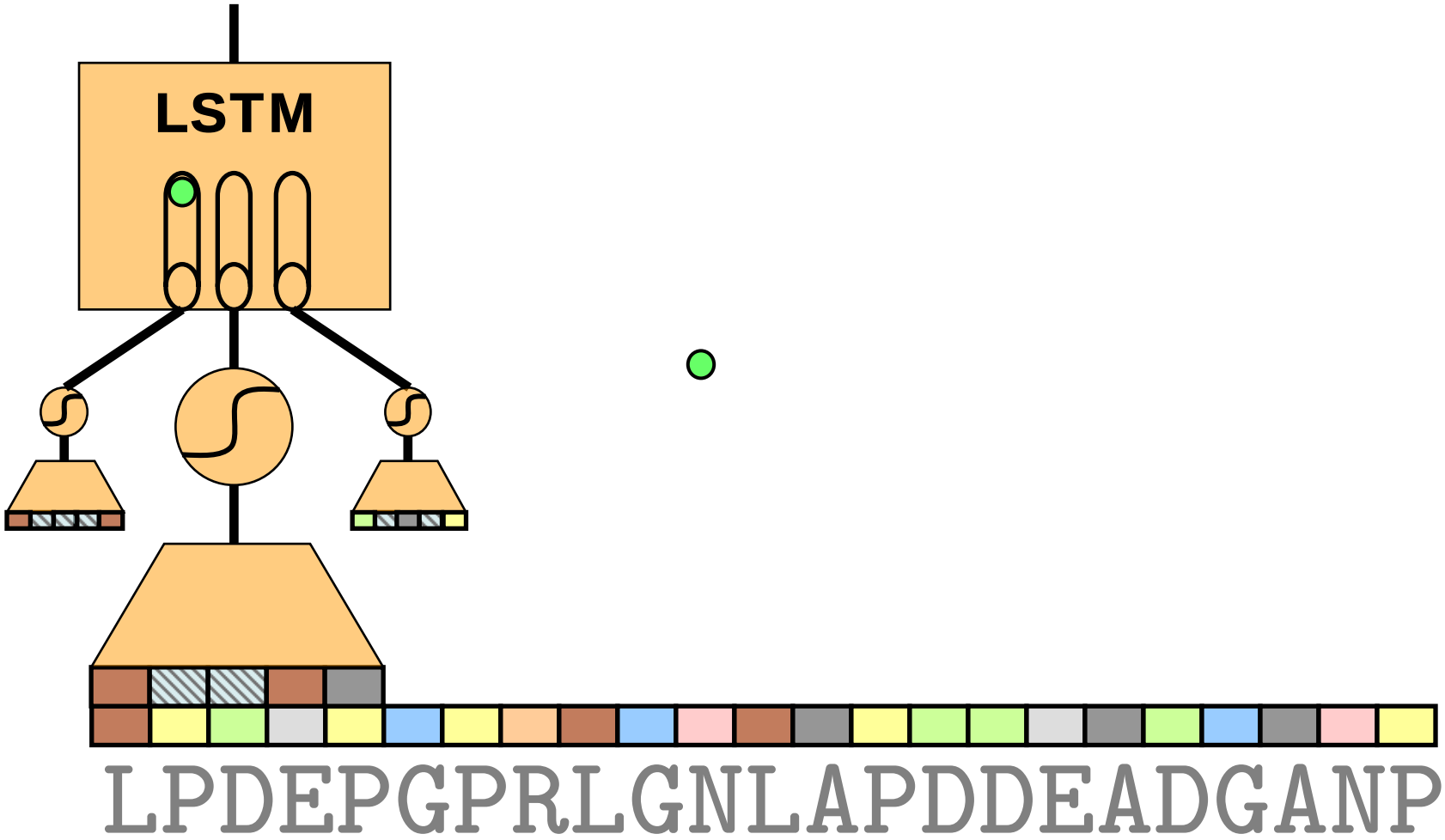
➥ classify a protein given as sequence of amino acids into functional or structural classes



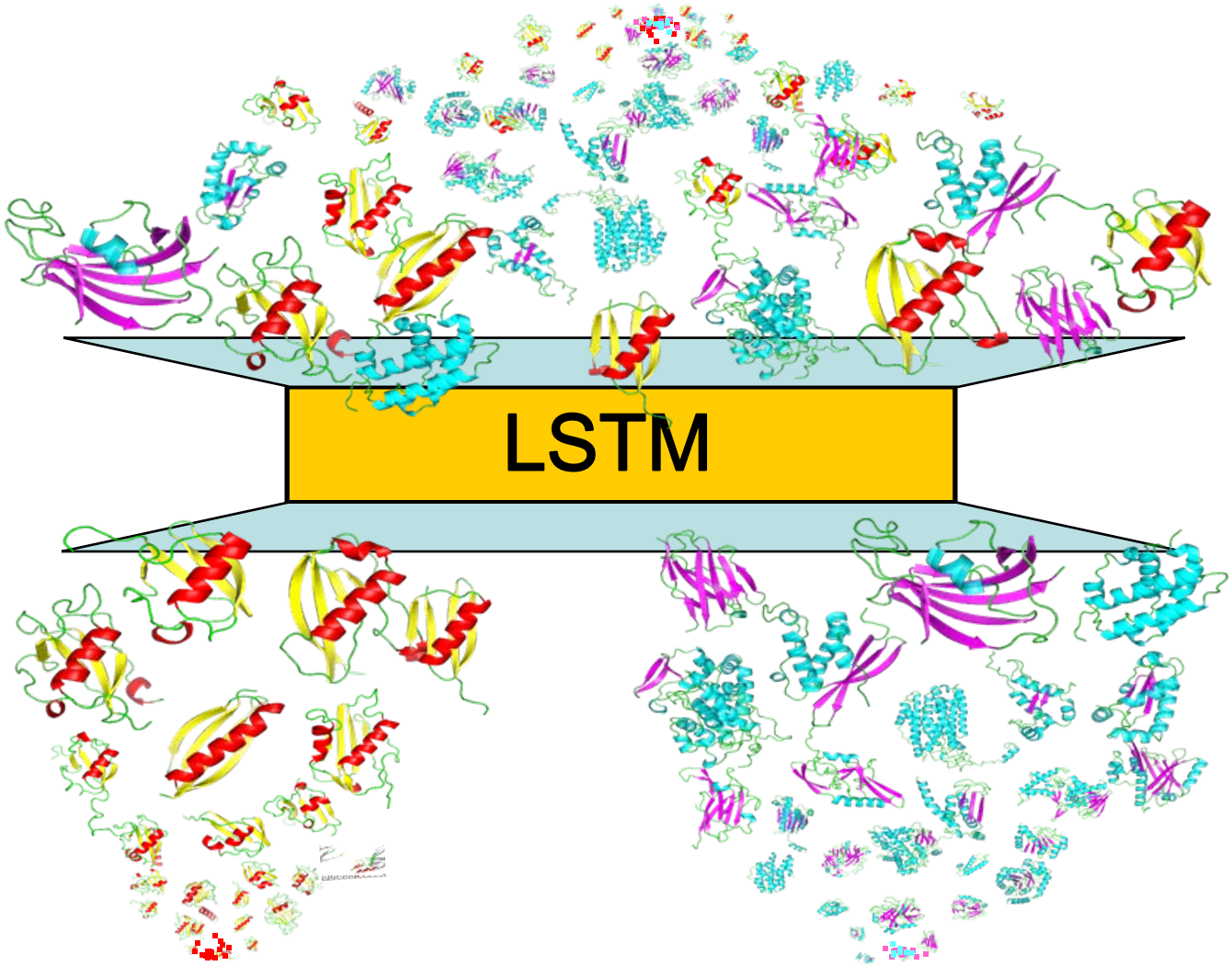

```
APSRKFFVGGNWKMNGRKQSL
GELIGTLNAAKVPADTEVVCA
PPTAYIDFARQKLDPKIAVAA
QNCYKVTNGAFTGEISPGMIK
DCGATWVVLGHSERRHVFGES
DELIGQKVAHALAEGLGVIAC
IGEKLDEREAGITEKVVFEQT
KVIADNVKDWSKVVLAYEPVW
AIGTGKTATPQQAQEVHEKLR
GWLKSNVSDAVAQSTRIIYGG
SVTGATCKELASQPDVDGFLV
GGASLKPEFVDIINAKQ
```

Sequence Classification

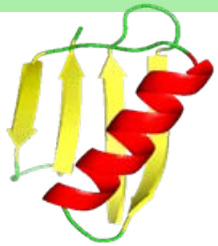
Long Short-Term Memory (LSTM)



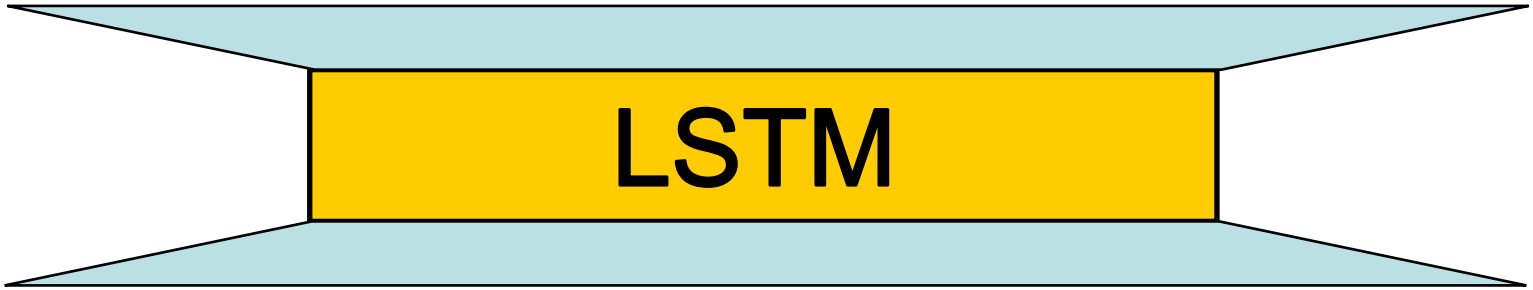
Sequence Classification



Sequence Classification



MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE



MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

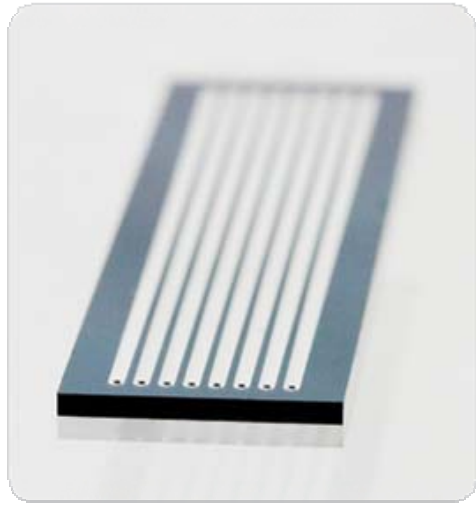
Next Generation Sequencing

- massive parallel sequencing =
  high-throughput sequencing =
  next-generation sequencing

- Roche's 454:
  400,000 reads (sequenced fragments)
  of 250–400bp length in up to 16 lanes

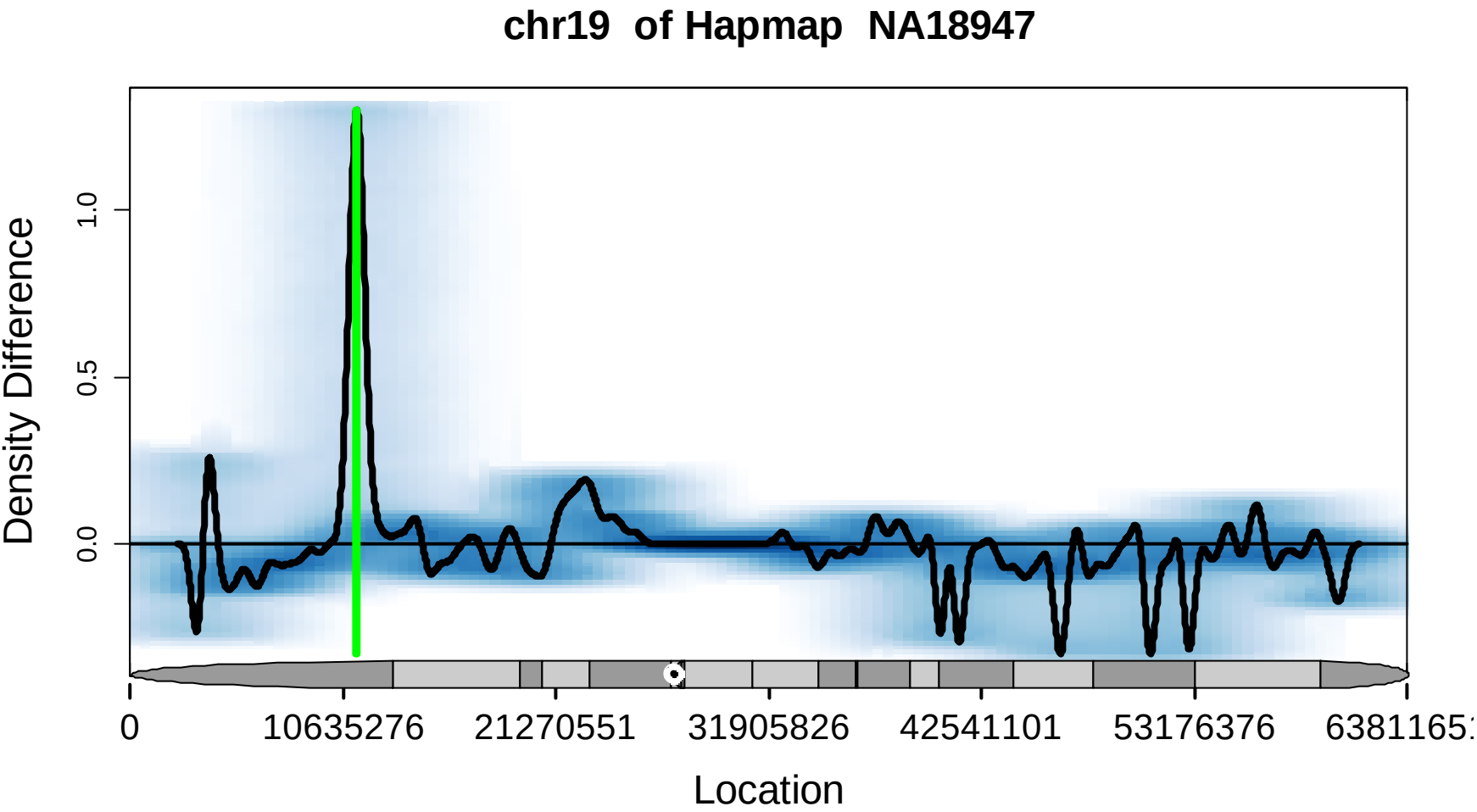
- Illumina's Solexa:
  30–60 million reads in one run in 8 lanes
  (a lane may be a sample) of 25-72bp long
  ⇨ higher coverage

FIGURE 1: ILLUMINA GENOME ANALYZER FLOW CELL



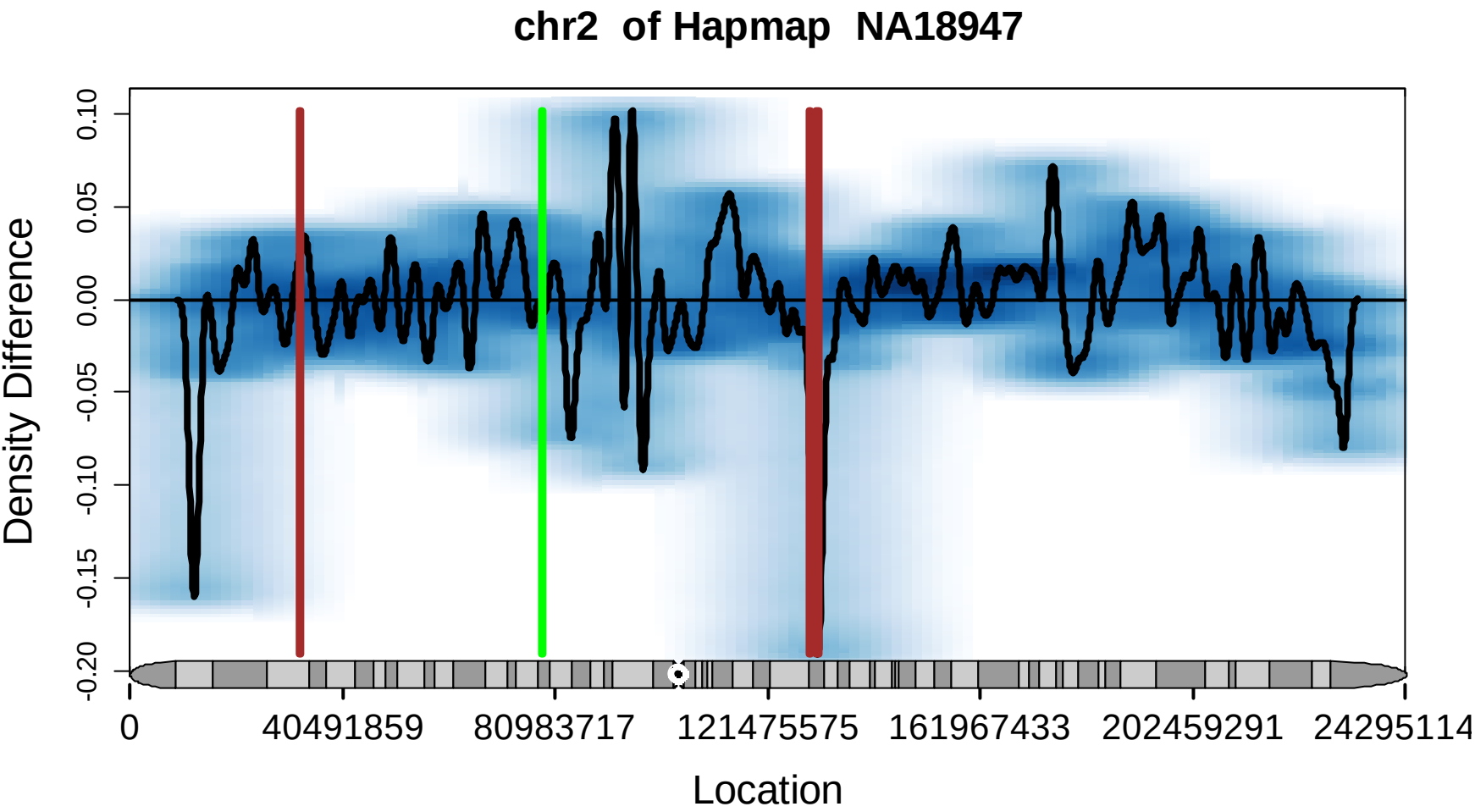
Next Generation Sequencing

An amplification (vertical line) in chromosome 19 detected by BAC arrays



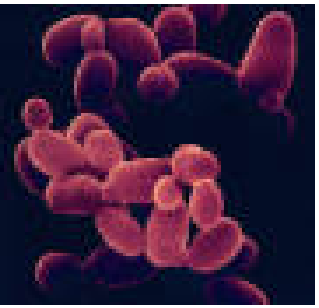
Next Generation Sequencing

Unexplained



Nucleosome Position

Prof. Douglas Murray
Institute for Advanced Biosciences



Nucleosome Position

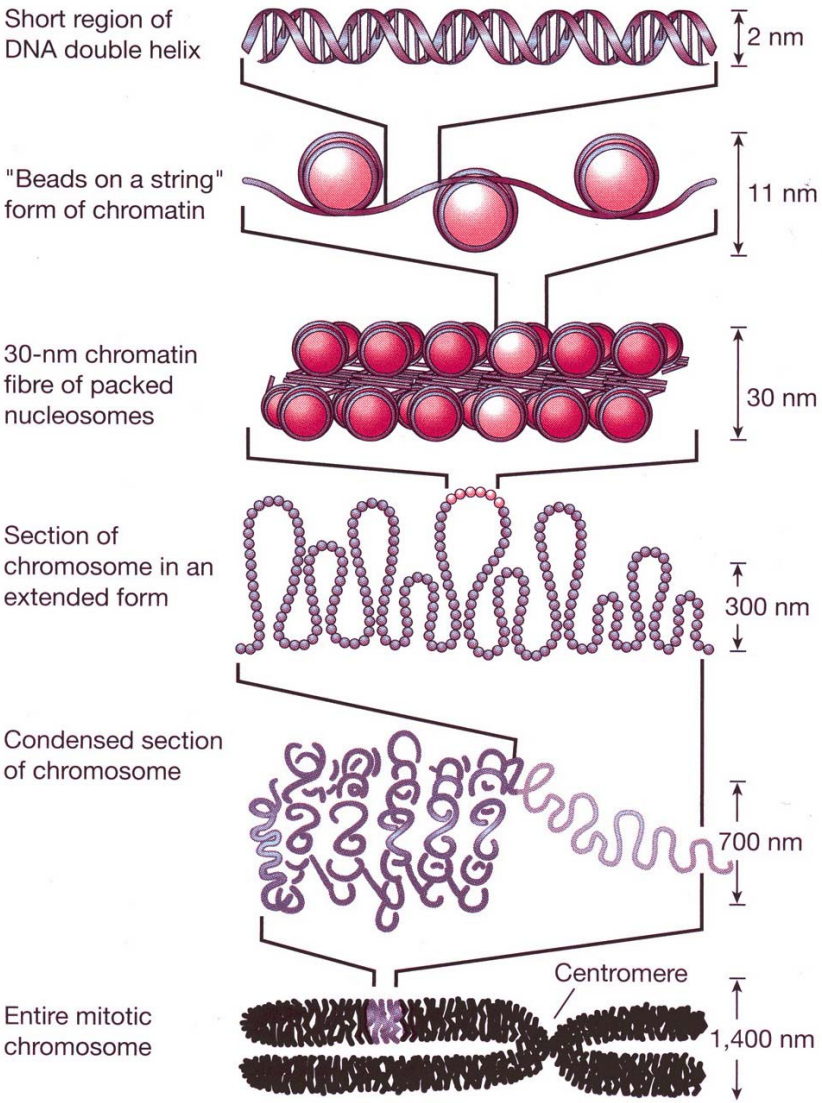
MEDIZINISCHE UNIVERSITÄT
INNSBRUCK

Division of Molecular Biology
Chromatin- and Epigenetics Laboratory
a. Prof. Alexandra Lusser

universität wien

Institute for Theoretical Chemistry
Prof. Dr. Peter Schuster
Rainer Machne

Nucleosome Position



DNA

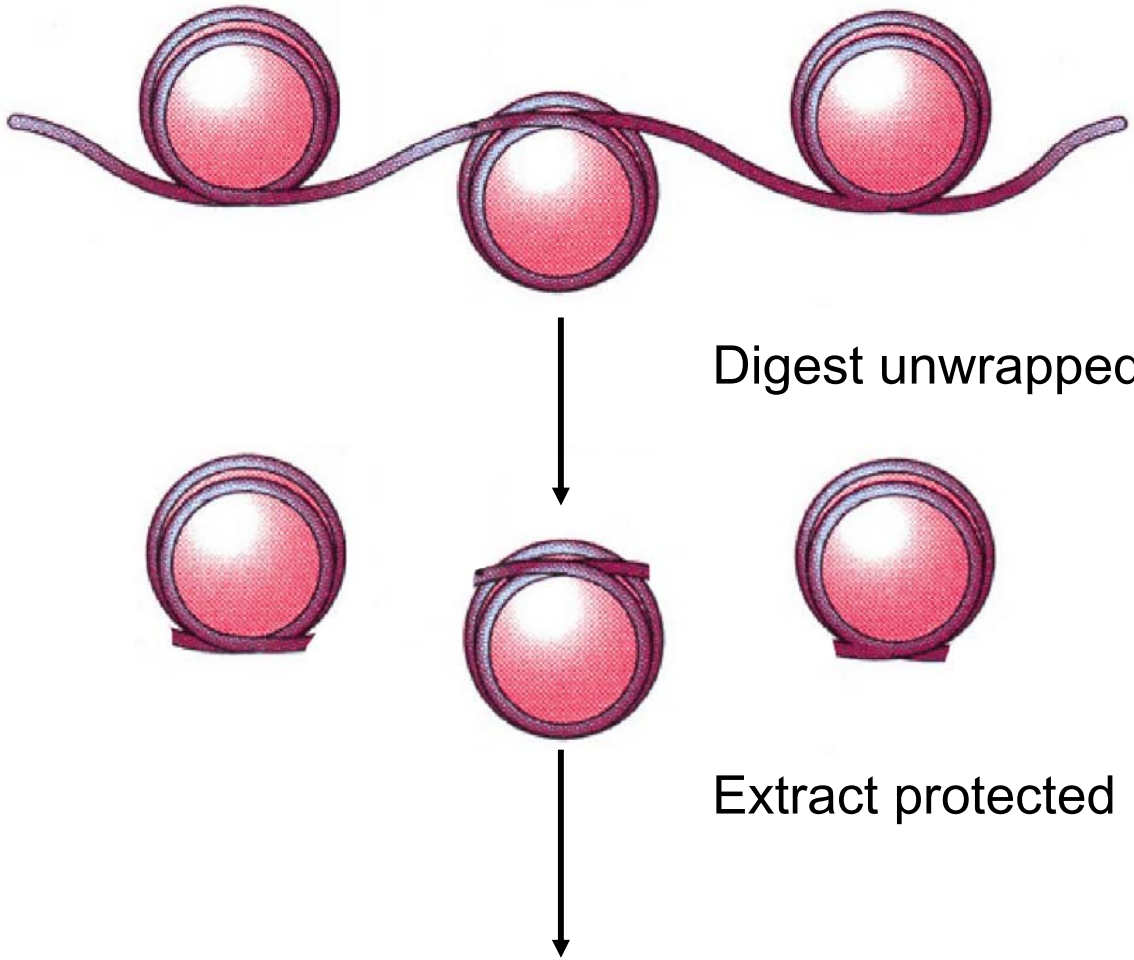
Nucleosomes

Chromosome

Felsenfeld & Groudine, *Nature* (2003)

Nucleosome Position

Isolation of natural nucleosome DNAs



We need help from other disciplines in computational science

Help from Math

Next generation sequencing:

Task1: read mapping hg18 against hg18

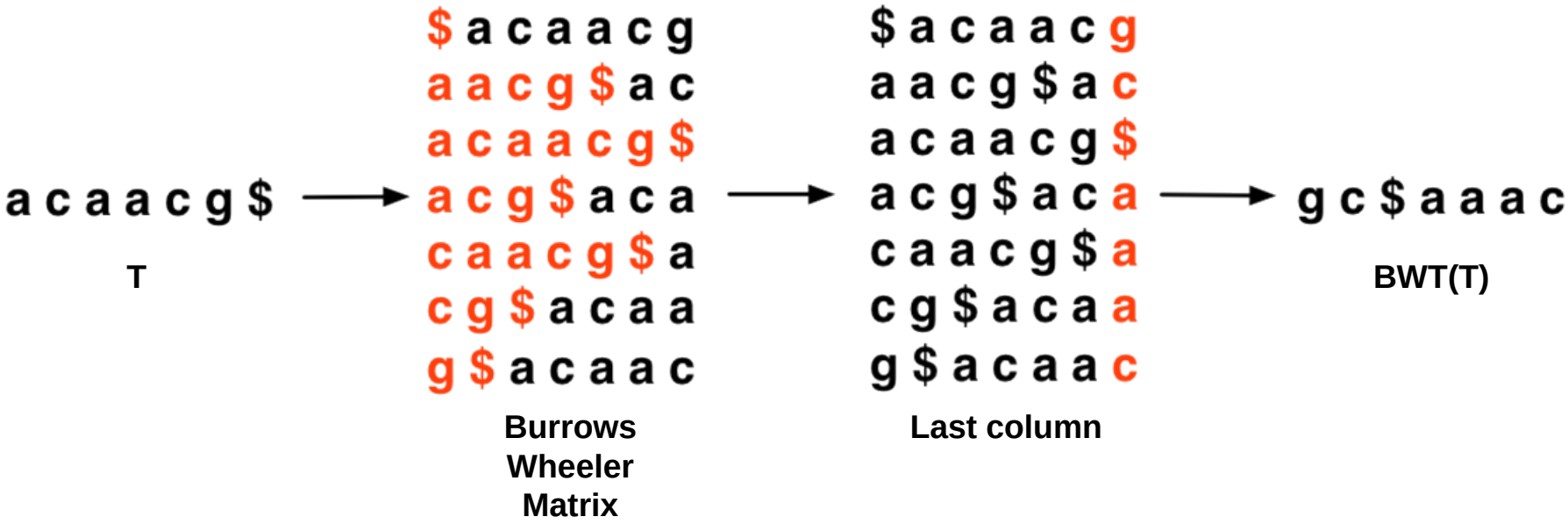
- about 100,000,000,000 reads of length 32 bp
- 64 runs
- one run: memory of 20GB, 50 processes simultaneously, 20h
- 53 days of computation

Task2: read mapping 1000Genomes data

- 10,000,000 to 50,000,000 reads per data set
- 836 data sets
- one data set: memory of 2GB, processes simultaneously, 8h
- 278 days of computation

Help from Math

Burrows-Wheeler-transformation (in bzip2)



Find substrings in strings

can results of polynom factorization help?

HELP

→Symbolic Computation

Help from Math

Estimate posterior for latent variable models:

$$p(\boldsymbol{z} \mid \{\boldsymbol{x}\}) = \frac{p(\{\boldsymbol{x}\} \mid \boldsymbol{z})\ p(\boldsymbol{z})}{\int p(\{\boldsymbol{x}\} \mid \boldsymbol{z})\ p(\boldsymbol{z})\ d\boldsymbol{z}}$$

Evaluate

$$\int p(\{\boldsymbol{x}\} \mid \boldsymbol{z})\ p(\boldsymbol{z})\ d\boldsymbol{z}$$

to compute moments of

$$p(\boldsymbol{z} \mid \{\boldsymbol{x}\})$$

HELP

→Statistics (sampling)
→Computational Mathematics

Help from Math

lower bound on likelihood

$$\boxed{\ln \mathcal{L}(\{\boldsymbol{x}\};\boldsymbol{w})} = \ln p(\{\boldsymbol{x}\};\boldsymbol{w}) = \ln \int_Z p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w})\, d\boldsymbol{z} =$$

$$\ln \int_Z \frac{Q(\boldsymbol{z} \mid \{\boldsymbol{x}\})}{Q(\boldsymbol{z} \mid \{\boldsymbol{x}\})} p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w})\, d\boldsymbol{z} \boxed{\geq}$$ Jensen's inequality

$$\int_Z Q(\boldsymbol{z} \mid \{\boldsymbol{x}\}) \ln \frac{p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w})}{Q(\boldsymbol{z} \mid \{\boldsymbol{x}\})}\, d\boldsymbol{z} =$$

$$\int_Z Q(\boldsymbol{z} \mid \{\boldsymbol{x}\}) \ln p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w})\, d\boldsymbol{z} - \int_Z Q(\boldsymbol{z} \mid \{\boldsymbol{x}\}) \ln Q(\boldsymbol{z} \mid \{\boldsymbol{x}\})\, d\boldsymbol{z}$$

maximize

$$h(\boldsymbol{w}) = \int_Z Q(\boldsymbol{z} \mid \{\boldsymbol{x}\}) \ln p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w})\, d\boldsymbol{z}$$

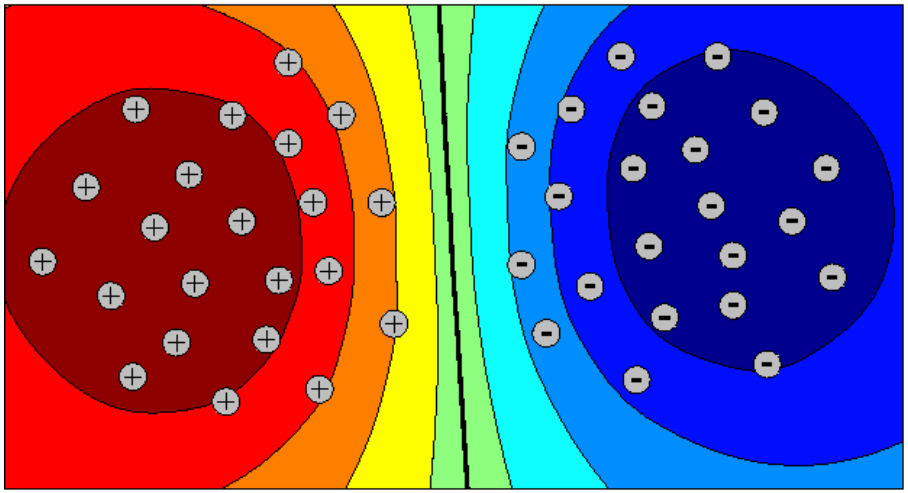
$$\ln p(\{\boldsymbol{x}\},\boldsymbol{z};\boldsymbol{w}) =$$
$$\phi\left(\|\boldsymbol{x} - \rho(\boldsymbol{w},\boldsymbol{z})\|\right) + \ln g(\boldsymbol{w})$$

HELP

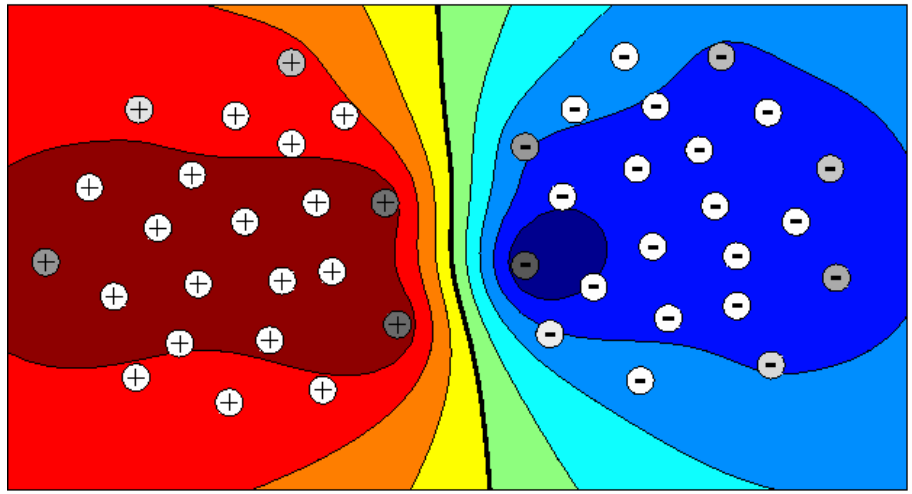
→Computational Mathematics

Help from Math

before charge redistribution



after charge redistribution



Energy: $\mathrm{E} = \frac{1}{2} \sum_{i=1}^{l} \varphi_i \, Q_i = \frac{1}{2} \boldsymbol{Q}^T \boldsymbol{P} \boldsymbol{Q} =$

$$\frac{1}{2} \sum_{i,j=1}^{l} y_i \, y_j \, \alpha_i \, \alpha_j \, P_{ij}$$

dual problem

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{i,j} \alpha_i \, \alpha_j \, y^i \, y^j \, k\left(\boldsymbol{x}^j, \boldsymbol{x}^i\right) - \sum_{i=1}^{l} \alpha_i$$

$$\text{s.t.} \quad \alpha_i \geq 0$$

$$\sum_{i=1}^{l} \alpha_i \, y^i = 0$$

Potential: $\varphi_i = \sum_{j=1}^{l} P_{ij} \, Q_j \quad \left(Q_i = y_i \, \alpha_i \right)$

Potential coefficients: P_{ij} (P_{ii} self – potential)

classification

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} y^i \, \alpha_i \, \boldsymbol{x}^T \, \boldsymbol{x}^i + b$$