Evolution of gene expression - a genome wide comparison between human and mouse

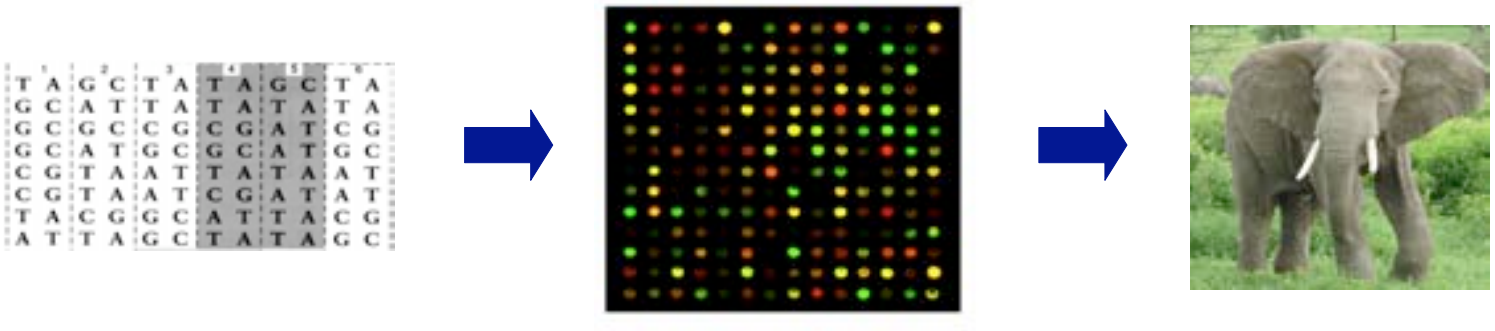
Marta Łuksza
Max Planck Institute for Molecular Genetics, Berlin

Johannes Berg
Physikalisches Institut Albert-Ludwigs-Universität Freiburg

Michael Lässig
Institute for Theoretical Physics, University of Cologne

Motivation

- **Gene expression data** quantify evolution at a **molecular phenotype level**

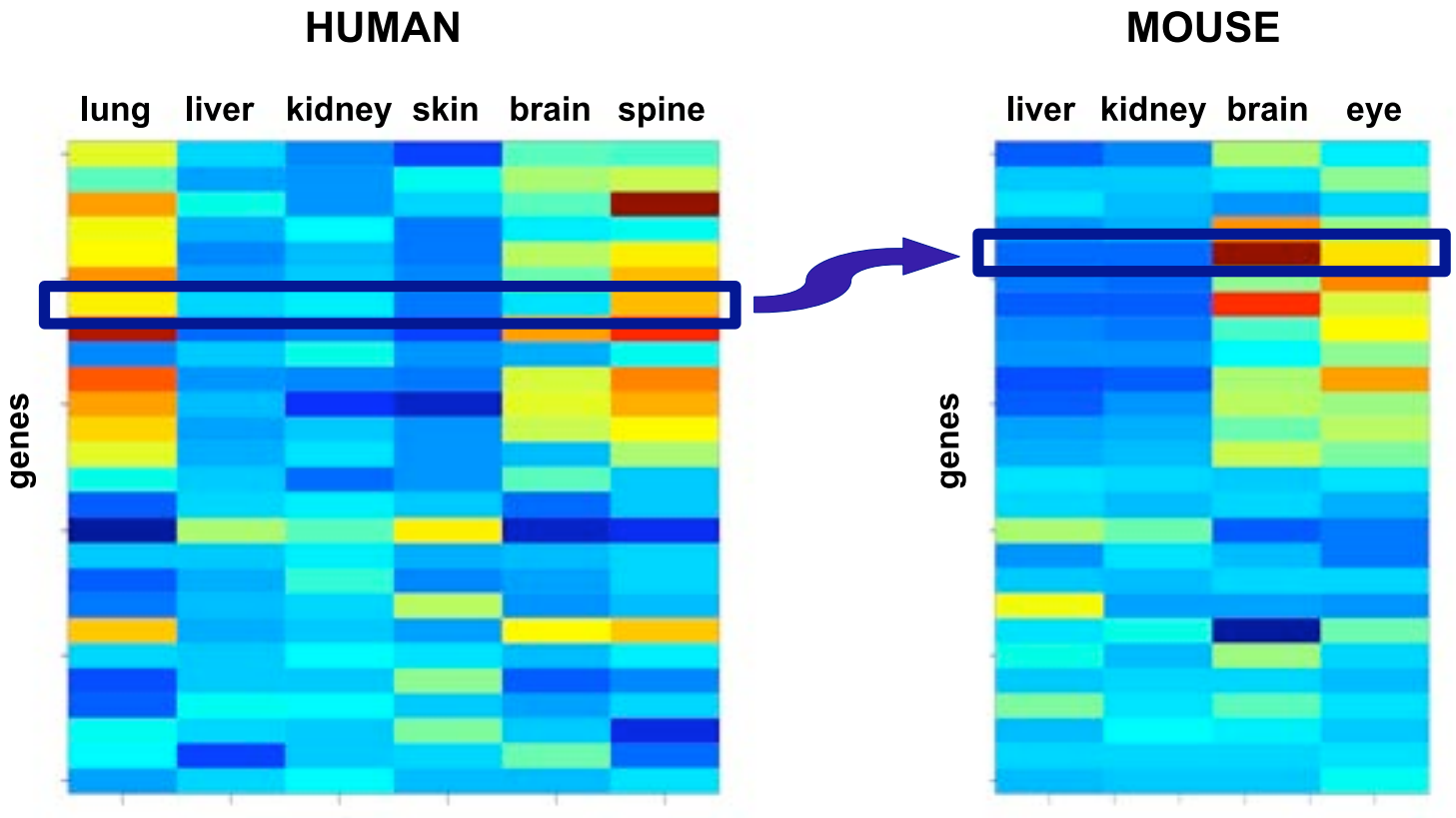


sequence | gene expression | organism

- We need to understand:
  - Statistical basis: evolution of **single genes** and **gene clusters**.
  - Patterns of conservation
  - Patterns of adaptive changes
- **Dataset of this study:** Novartis gene atlas, genome-wide microarray for human and mouse (Su et al. 2004).

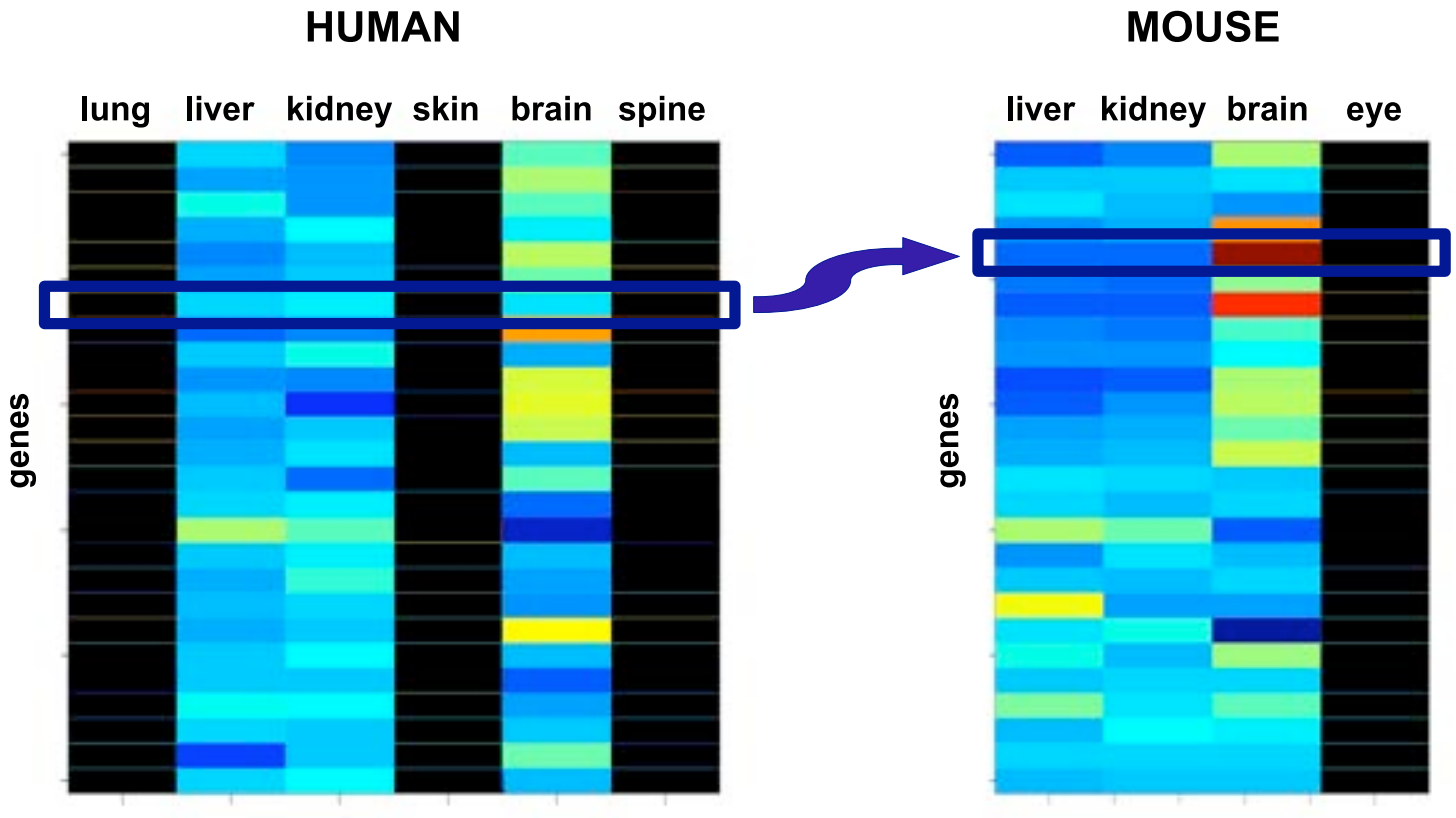
Cross-species comparison of single genes

- Use only the set of **common tissues**



Cross-species comparison of single genes

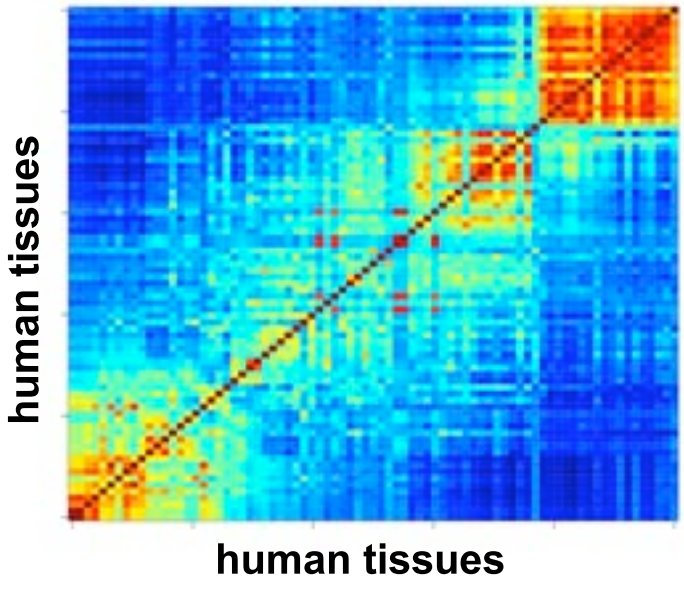
- Use only the set of **common tissues**



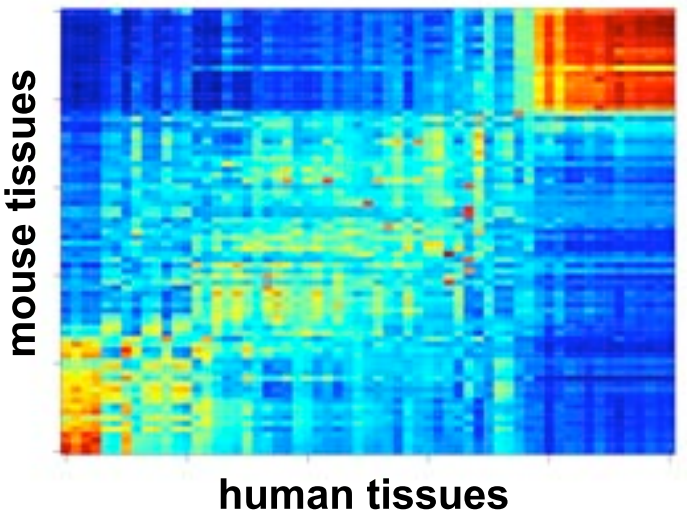
- **But**
  - a lot of information is disregarded
  - mapped tissues can differ between species

Tissue expression dependencies

- **Our method:** quantify **tissue correlations** both **within** and **between** species



within species

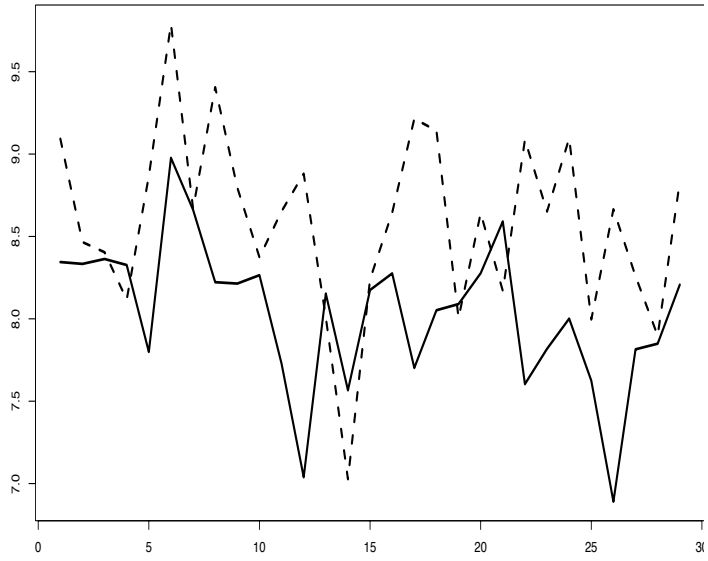


across species

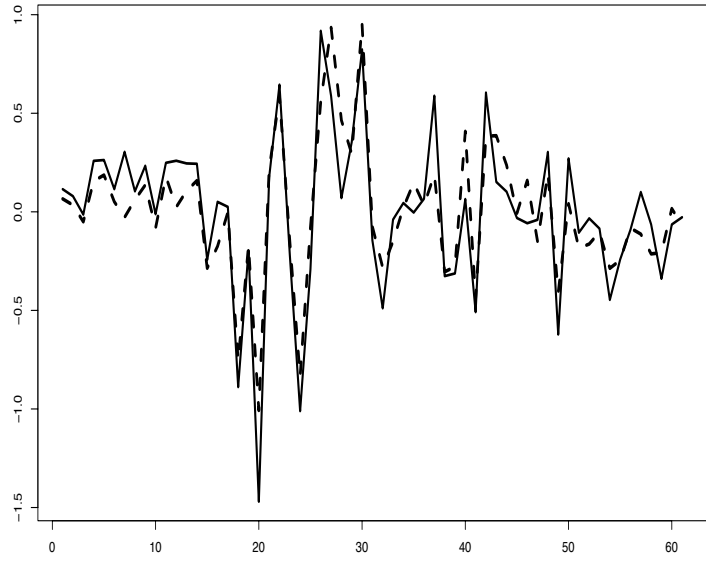
[using orthologous gene pairs]

Cross-species tissue mapping

- Our method accurately maps expression profiles from different species



common tissues only



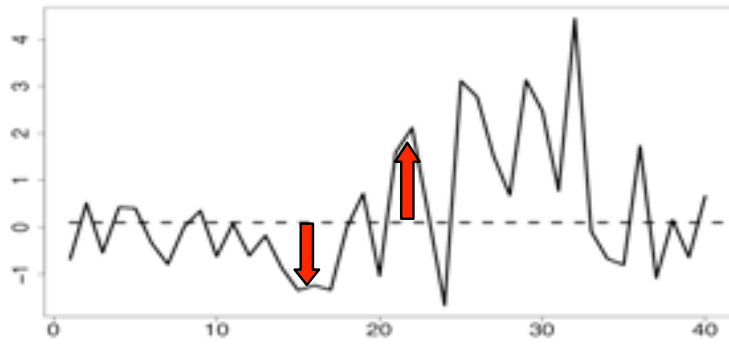
quantitative tissue mapping

[expression profiles of an orthologous gene pair along tissues]

Statistics of expression data

- Compare **deviations** from the mean

$$\mathbf{x_i} = \xi_\mathbf{i} - \langle \xi_\mathbf{i} \rangle, \quad \hat{\mathbf{x}}_\mathbf{i} = \hat{\xi}_\mathbf{i} - \langle \hat{\xi}_\mathbf{i} \rangle$$



- Normalized vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$ follow Gaussian distributions

$$P_1(\mathbf{x}) \sim \exp(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{g}^{-1}\mathbf{x}), \quad P_2(\hat{\mathbf{x}}) \sim \exp(-\frac{1}{2}\hat{\mathbf{x}}^\mathsf{T}\hat{\mathbf{g}}^{-1}\hat{\mathbf{x}})$$

- **Related** genes show expression similarity as described by covariance matrix **G**

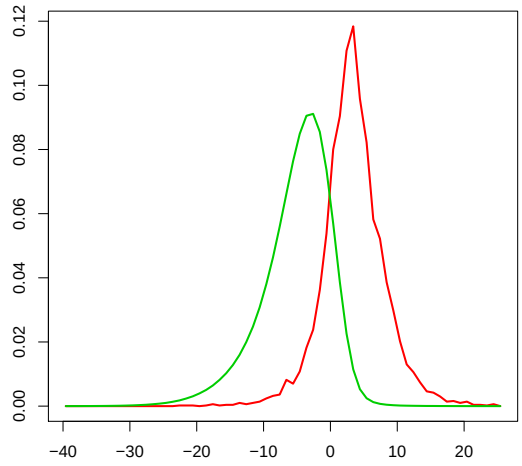
$$Q(\mathbf{x}, \hat{\mathbf{x}}) \sim P_1(\mathbf{x})P_2(\hat{\mathbf{x}}) \exp(-\frac{1}{2}[\mathbf{x}\ \hat{\mathbf{x}}]^\mathsf{T}(\mathbf{G}^{-1})[\mathbf{x}\ \hat{\mathbf{x}}])$$

$[\mathbf{x}\ \hat{\mathbf{x}}]$ is a concatenation of vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$

G is a concatenation of within- and across-species covariance matrices

Statistics of expression data: scoring expression similarity

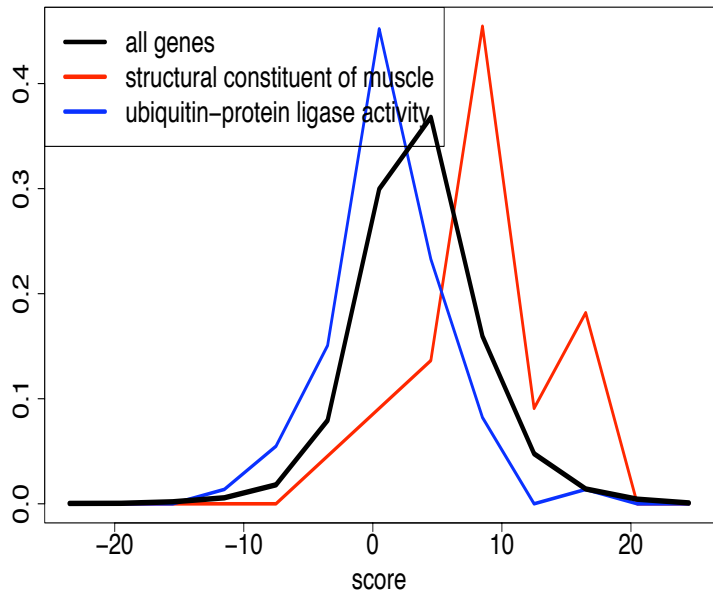
- $P_1(\mathbf{x})P_2(\hat{\mathbf{x}})$ – null model
- $Q(\mathbf{x}, \hat{\mathbf{x}})$ – functional model
- **Log-likelihood score:** $S(\mathbf{x}, \hat{\mathbf{x}}) = \log \frac{Q(\mathbf{x},\hat{\mathbf{x}})}{P_1(\mathbf{x})P_2(\hat{\mathbf{x}})}$



- 67% of orthologs have significant score (p-value < 0.05)
  - red curve - distribution of scores of orthologous pairs
  - green curve - distribution of scores of random gene pairs

Is expression conservation/divergence related to gene function?

- Example:



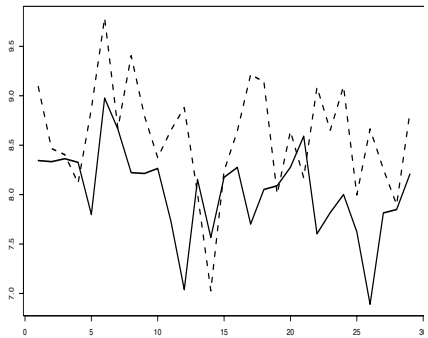
- Kolmogorov-Smirnov test to compare distribution of scores specific to functional classes (as given by Gene Ontology)

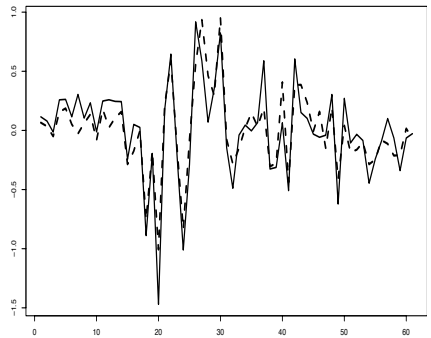
Functional conservation and divergence

	p-value
calcium ion binding	3e-06
plasma membrane	5e-06
cytoplasm	7e-06
structural constituent of muscle	1e-05
immune response	2.1e-05
NADH dehydrogenase activity	2.6e-05
structural constituent of cytoskeleton	4.2e-05
extracellular region	5.3e-05
striated muscle contraction	6.4e-05
mitochondrial respiratory chain complex I	8.4e-05
serine-type endopeptidase inhibitor activity	0.000123
ubiquitin cycle	3e-06
ubiquitin-protein ligase activity	3.7e-05
cation binding	0.002606
rhodopsin-like receptor activity	0.006037
DNA-directed RNA polymerase activity	0.006571
interleukin receptor activity	0.007411
carbohydrate metabolic process	0.007443

Statistics of expression data: summary

- **Quantitative mapping of tissues**





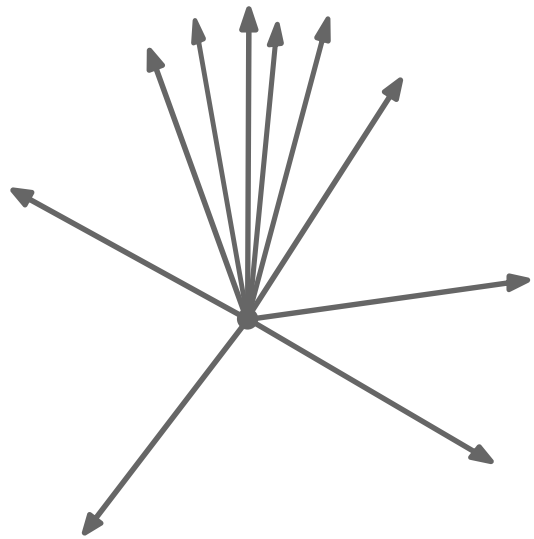
before after

- **Similarity measure** for expression of genes defined by the tissue covariance matrices
  - within species: $\mathbf{x_1} \cdot \mathbf{x_2} = \mathbf{x_1 g^{-1} x_2}$
  - across species: $\mathbf{x} \cdot \hat{\mathbf{x}}$

- **Statistical scoring of similarities**: log-likelihood score $S(\mathbf{x}, \hat{\mathbf{x}})$

Coexpression of genes

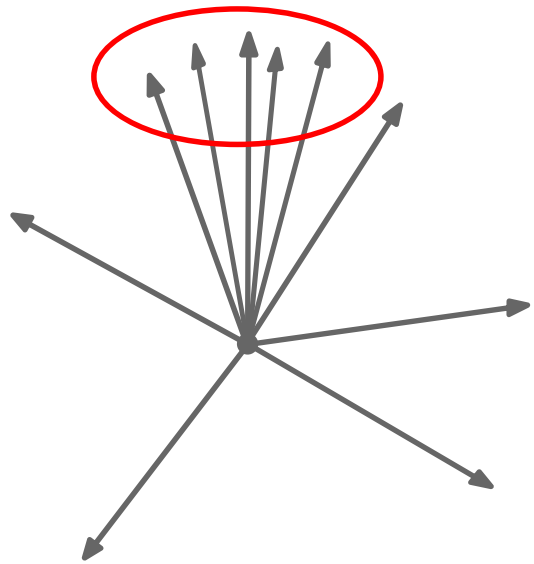
- Vectors in **high-dimensional** space



- Define a **cluster score**
- Find **significance** of the cluster

Coexpression of genes

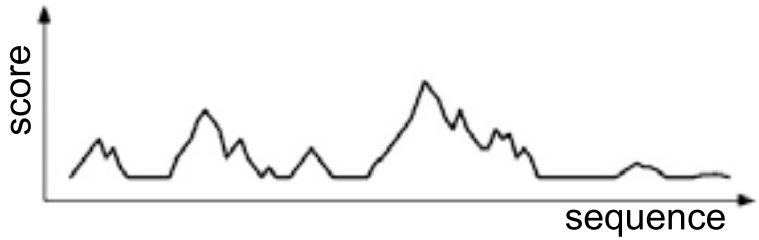
- Vectors in **high-dimensional** space



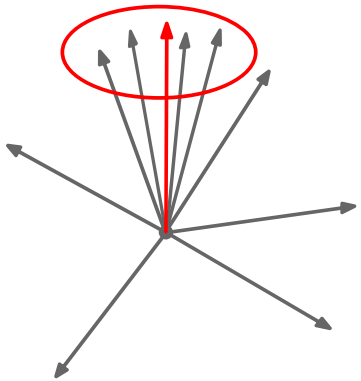
- Define a **cluster score**
- Find **significance** of the cluster

Coexpression of genes: probabilistic cluster analysis

- Local alignment score statistics
  - **many score islands**
  - *p*-value given by the Gumbel distribution
    (Karlin and Altschul (1990))



- Cluster score statistics analogy
  - Sequence letters $\longrightarrow$ vectors
  - reference sequence $\longrightarrow$ the direction vector repeated N times
  - Alphabet size $\longrightarrow$ vector space dimensionality
- **But:** there is **no ordering** of vectors



Cluster member scoring

- **Null model** (multivariate Gaussian)

$$P(\mathbf{x}) \sim \exp\left\{-\frac{\mathbf{x}\cdot\mathbf{x}}{2}\right\}$$

- **Cluster model** (for given direction vector $\mathbf{z}$)

$$Q_{cl}(\mathbf{x}|\mathbf{z}) = (Z_\eta)^{-1} P(\mathbf{x}) \exp\{\eta(\mathbf{x}\cdot\mathbf{z})\}$$

- **Log-likelihood score**

$$s_{cl}(\mathbf{x}|\mathbf{z}) = \log\frac{Q(\mathbf{x}|\mathbf{z})}{P(\mathbf{x})} = \eta(\mathbf{x}\cdot\mathbf{z}) - \mu$$

  where $\mu = \log Z_\eta$ is the offset given by normalization.

- Note: $s_{cl}(\mathbf{x}|\mathbf{z})$ is a similarity measure of $\mathbf{x}$ and $\mathbf{z}$

Fixed direction cluster

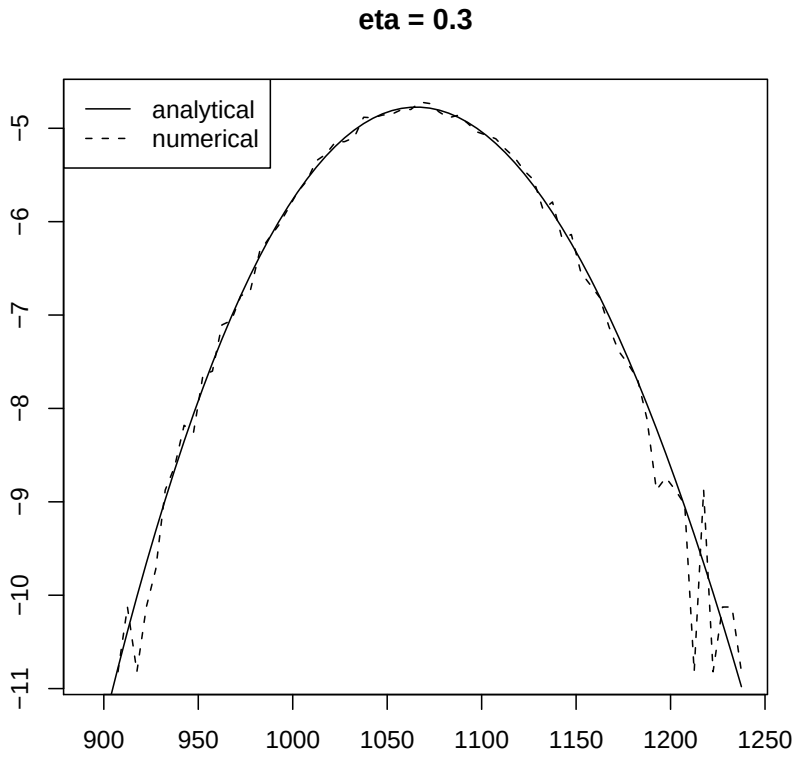
- Given vector $\mathbf{z}$
  – is it a centre of a cluster?
- N vectors in M-dimensional space, $\mathbf{x_1}, \ldots, \mathbf{x_N}$
- **Total cluster score** given by the sum of scores of **positively scoring** vectors

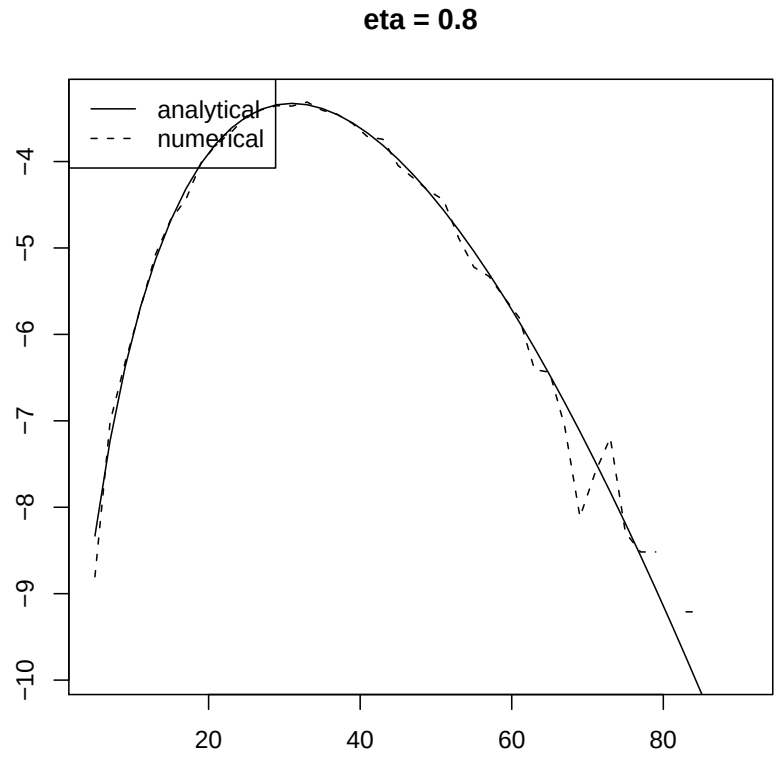
$$S_{cl}(\{\mathbf{x_i}\}|\mathbf{z}) = \sum_{i=1}^{N} \max\left(s_{cl}(\mathbf{x_i}|\mathbf{z}), 0\right)$$

- Compute the **distribution of cluster scores** under the **null model**.

Fixed direction cluster

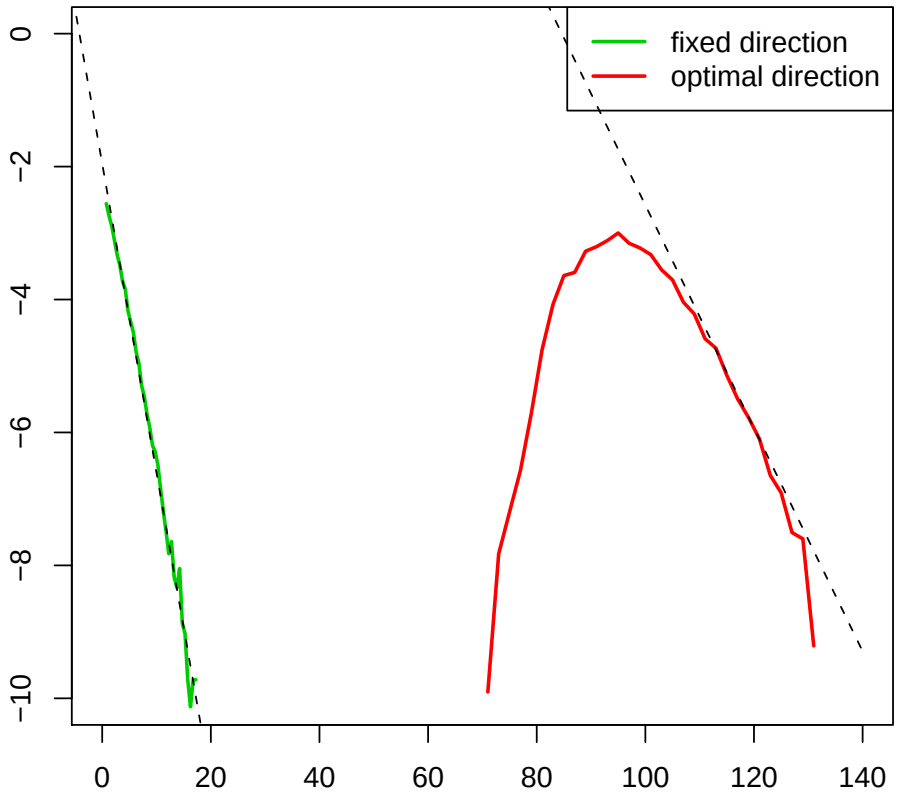
- Numerical experiments
  - Draw N vectors $\mathbf{x_1}, \ldots, \mathbf{x_N}$ from the null distribution P
  - Compute the cluster score with $\mathbf{z}$ set to the north pole
  - Repeat many times to get the numerical distribution of cluster scores
- Analytical and numerical distributions of cluster scores





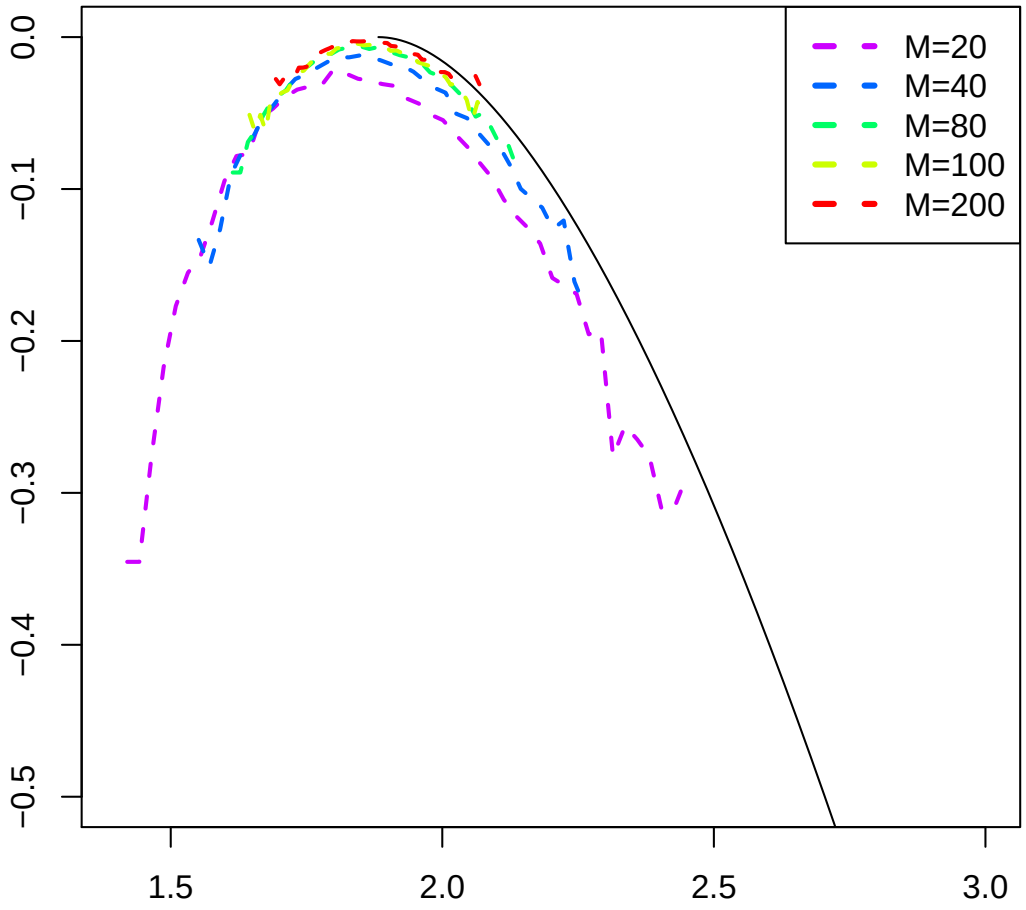
Maximum score cluster

- What happens if we look for the **optimal** direction **z**?
- **Maximum cluster score** distribution and the **fixed direction cluster score** distribution have **different slopes**



Maximum score cluster

- Solution is **asymptotic in** M – the number of dimensions
- Comparison with the numerical experiments

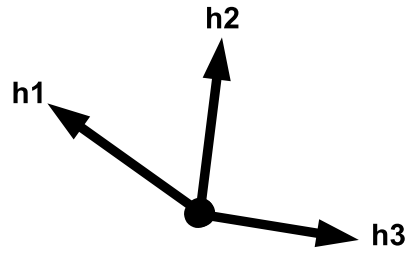


Conserved coexpression clusters

Single species cluster

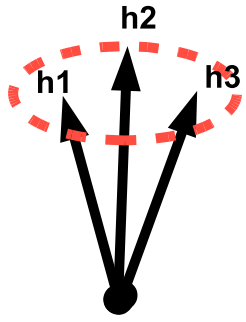
Null model:

independent genes



Functional model:

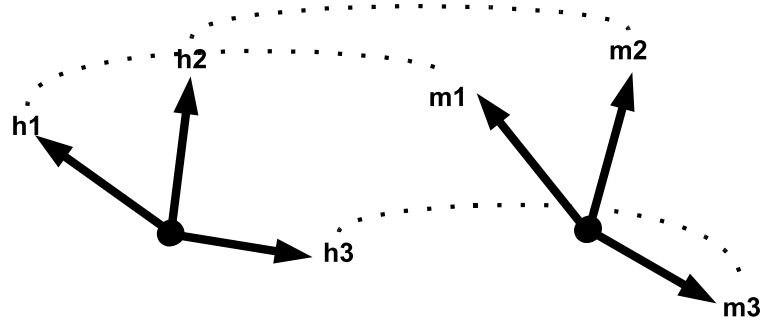
enhanced similarity to cluster centre



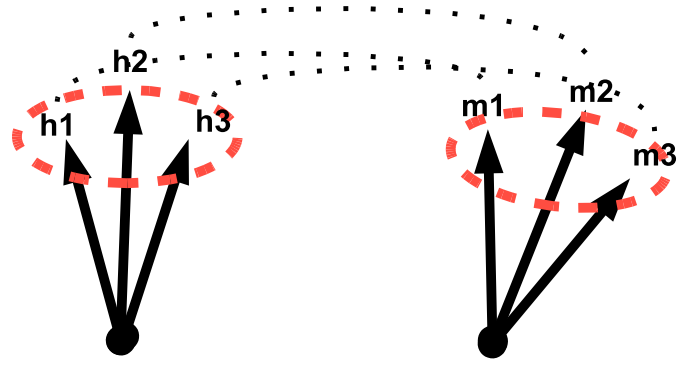
$$s_{Cl}(\mathbf{x}|\mathbf{z}) = \eta\,(\mathbf{x}\cdot\mathbf{z}) - \mu$$

Cross-species cluster

independent gene pairs



enhanced similarity to cluster centres



$$s_{Cl}([\mathbf{x}\ \hat{\mathbf{x}}]|[\mathbf{z}\ \hat{\mathbf{z}}]) = \eta\,(\mathbf{x}\cdot\mathbf{z}) + \hat{\eta}\,(\hat{\mathbf{x}}\cdot\hat{\mathbf{z}}) - \mu$$

Clustering problem

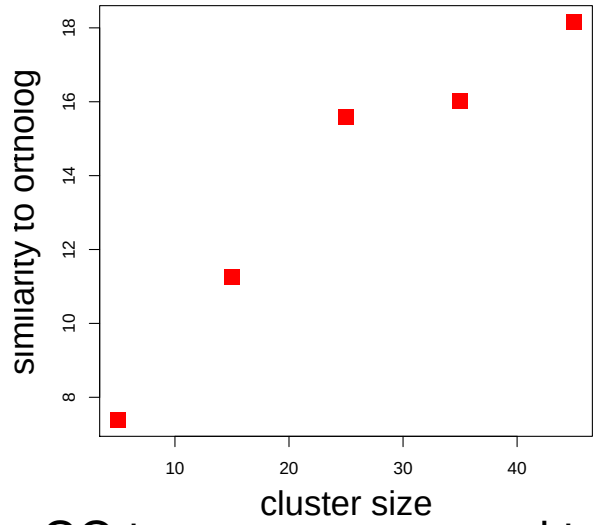
- **Cluster assignment** function, every element is mapped to one of the clusters, $m : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$ (at most N clusters)
- **Many cluster centers** (pairs of orthologs) $[\mathbf{x_{m(1)}}\ \hat{\mathbf{x}}_{\mathbf{m(1)}}], \ldots, [\mathbf{x_{m(G)}}\ \hat{\mathbf{x}}_{\mathbf{m(G)}}]$
- Find cluster centers and cluster assignment that **maximizes the total score** over all clusters:

$$\sum_{k \in \{1,\ldots,N\}, k \notin \mathrm{Image}(m)} s_{Cl}\left([\mathbf{x_k}\ \hat{\mathbf{x}}_{\mathbf{k}}] \middle| [\mathbf{x_{m(k)}}\ \hat{\mathbf{x}}_{\mathbf{m(k)}}]\right)$$

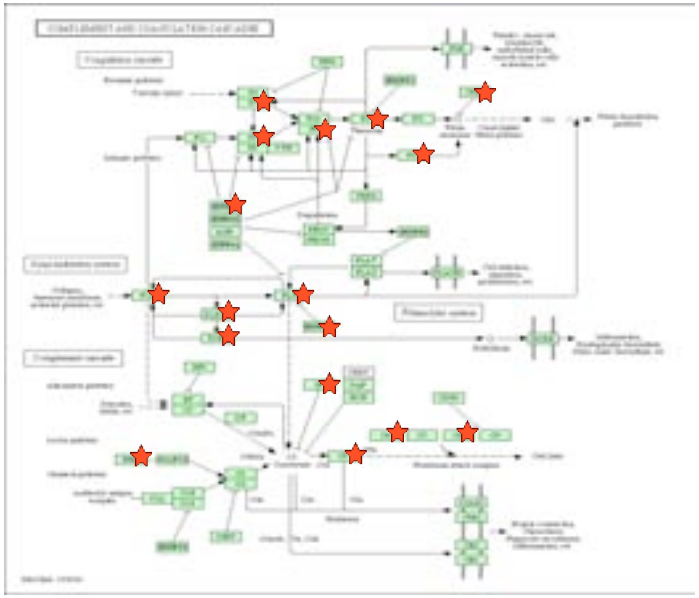
- solved by the Affinity Propagation clustering algorithm (Frey and Dueck, Science 2007)
- compute significance of the the resulting clusters

Conserved coexpression clusters

- Genes with **many interactions** have **more conserved expression** patterns (averaged over significant human clusters)



- Clusters are enriched in GO terms or correspond to KEGG pathways



Summary and outlook

- **Method**
  - Statistical theory for single gene expression comparison, both within- and across-species
  - Coexpression significance
  - Algorithm for detection of conserved co-expression
- **Conserved gene clusters** are functional modules
- **Adaptive changes** of gene expression patterns?
  - Are there joint clusters with significant pattern changes?