# Trends in Bioinformatics: Bi-Clustering with Biological Context Information

Willi Gierke

Supervisor: Cindy Perscheid

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

# Agenda

1. Motivation
2. Task
3. Algorithm
4. Evaluation
5. Discussion

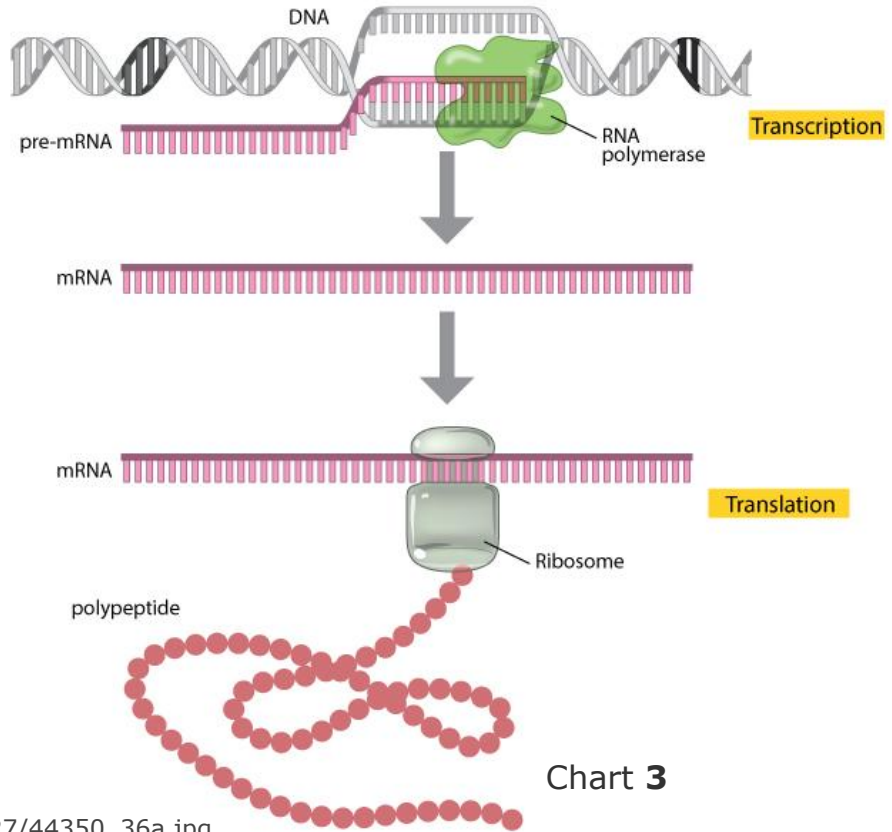
**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018

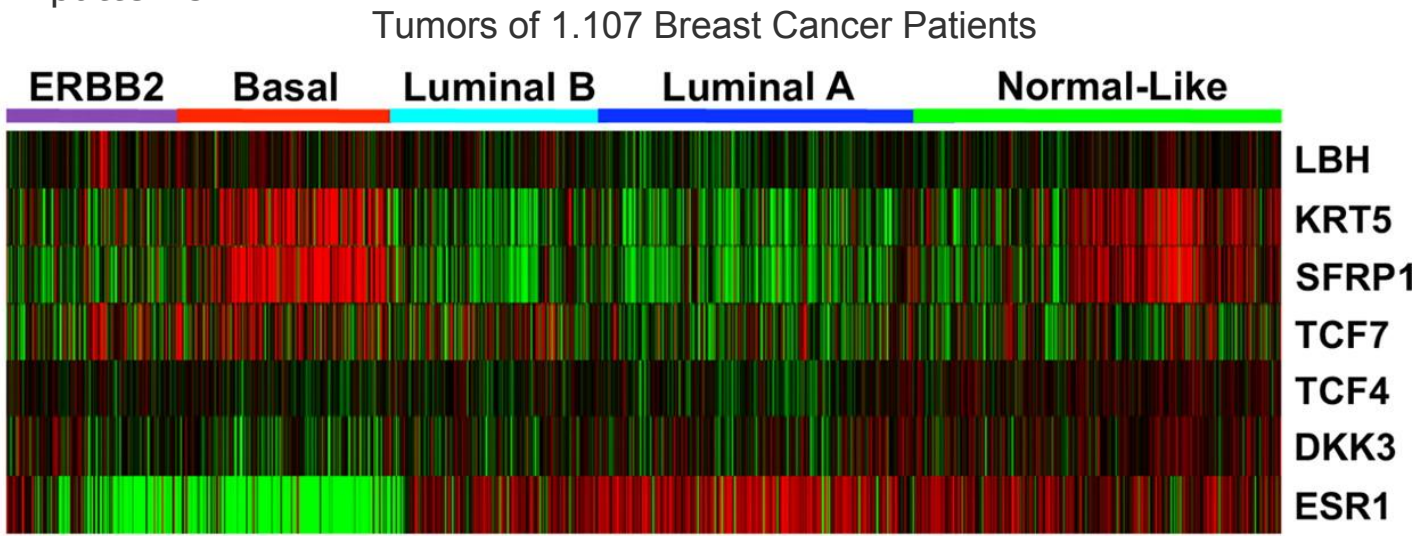
- RNA is translated into e.g. proteins
- Proteins influence functioning and phenotype of the cell

→ Infer from gene expression health condition of the cells




http://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/105292327/44350_36a.jpg

- Analyze sick patients to find common disease-specific expression patterns




**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



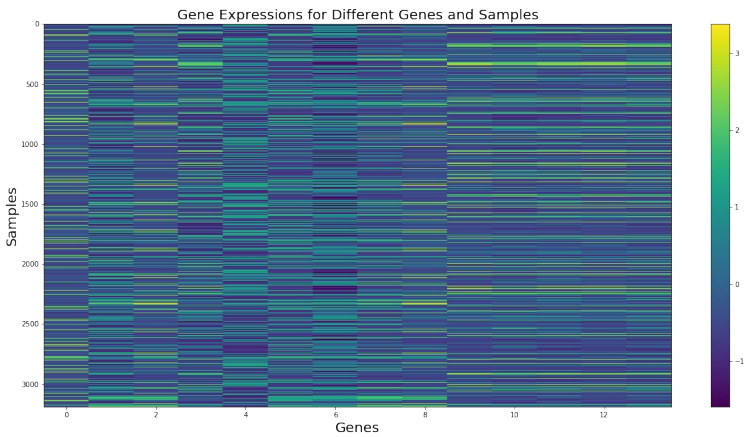
http://mcb.asm.org/content/30/17/4267/F6.large.jpg

# Motivation
# Biclustering

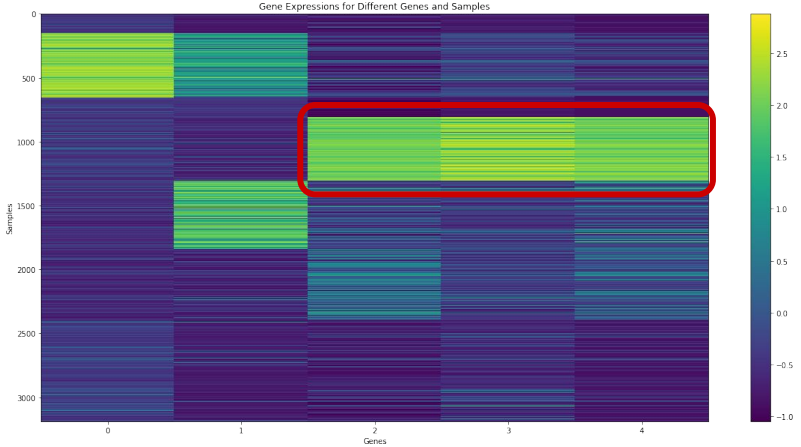
- Activation patterns of genes might only occur under specific conditions

- Necessary to find local patterns in gene expression data

Raw Expression Matrix




Expression Matrix with Biclusters

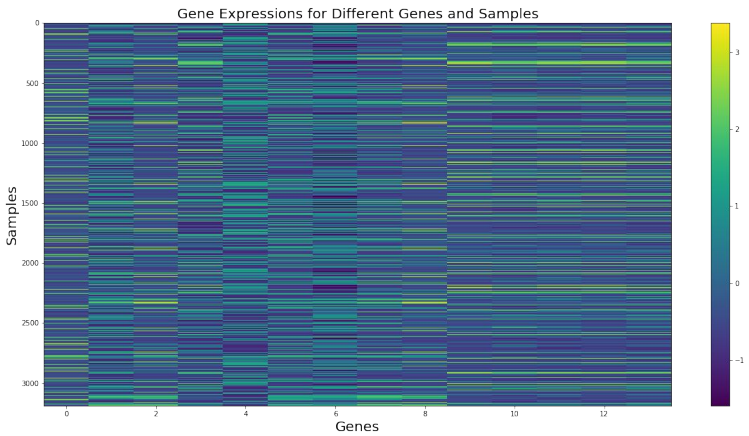


# Motivation
# Biclustering

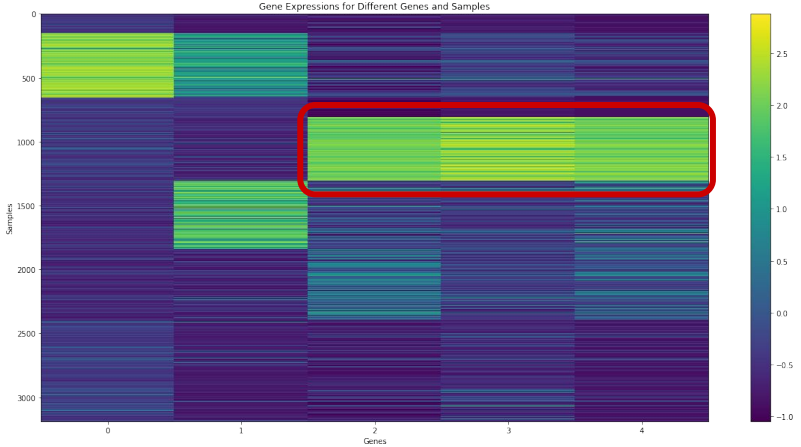
- Single clustering: find groups of genes similar in the complete dataset

- Biclustering: find groups of genes similar only in certain samples

Raw Expression Matrix




Expression Matrix with Biclusters



- Feature engineering: add features to support algorithm decisions
- Knowing some genes co-occur in e.g. pathways could accelerate finding biclusters / improve the quality of the results
- Incentivize to find patterns of interest
- Support finding hidden processes barely expressed by the data

# Motivation
# Biological Context Information

- e.g. Kernel Trick: add distance of instance from center of point mass as feature

  → Classes are linearly separable in feature space

  → Simplifies algorithm decision

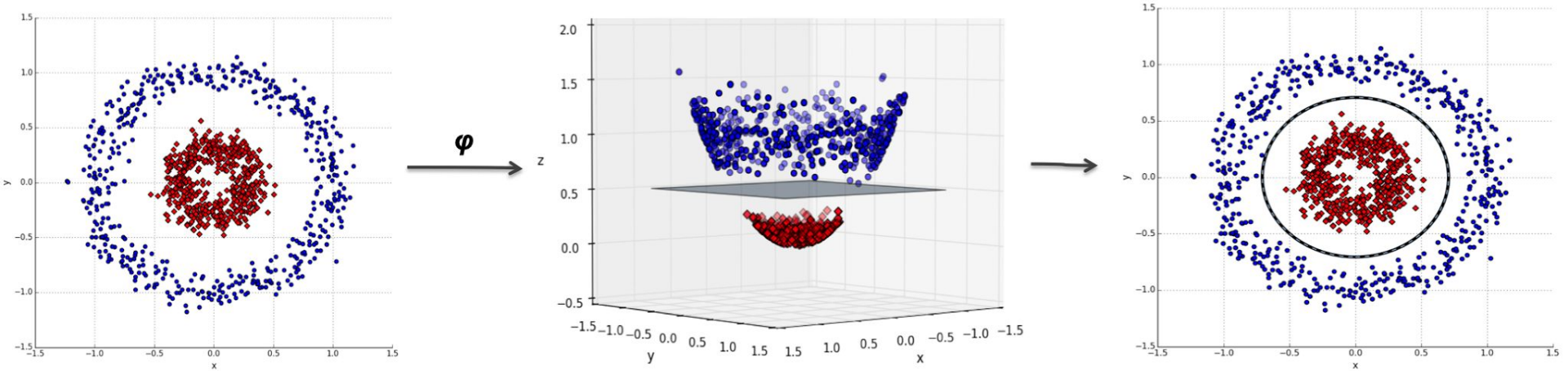



Chart **8**

http://beta.cambridgespark.com/courses/jpm/figures/mod5_kernel_trick.png

# Task
# Data Exploration

- Used a dataset by The Cancer Genome Atlas[1] project
- ~3000 tissue samples
- ~56k expressed genes
- 8 cancer types

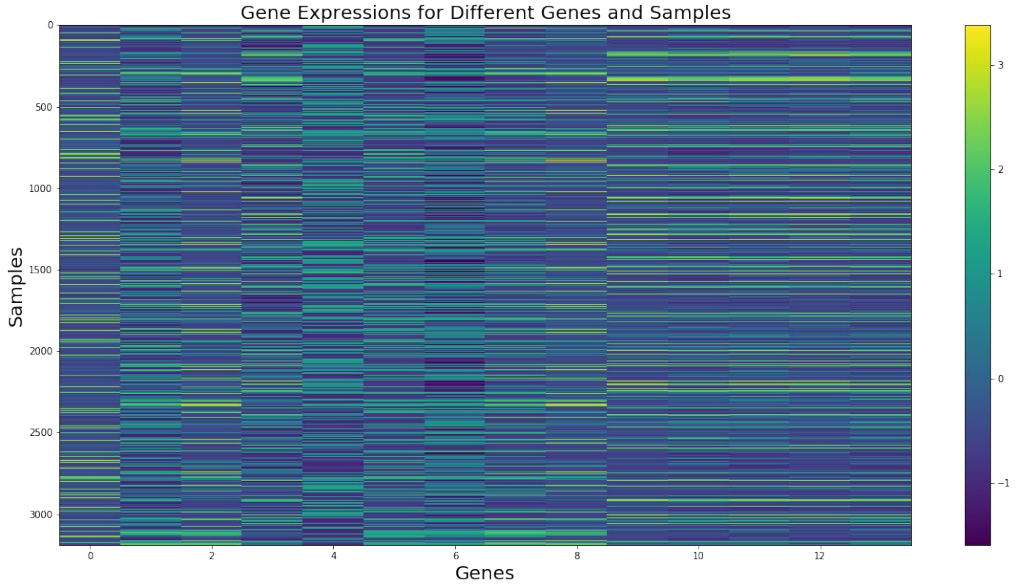

→ Are there groups of genes that behave similarly across different cancer types?
Can we find them easier using context information?



1) https://cancergenome.nih.gov/

- Idea: relevant gene expressions should highly differ between samples

  → Removed genes with low variances → only keep 14 genes



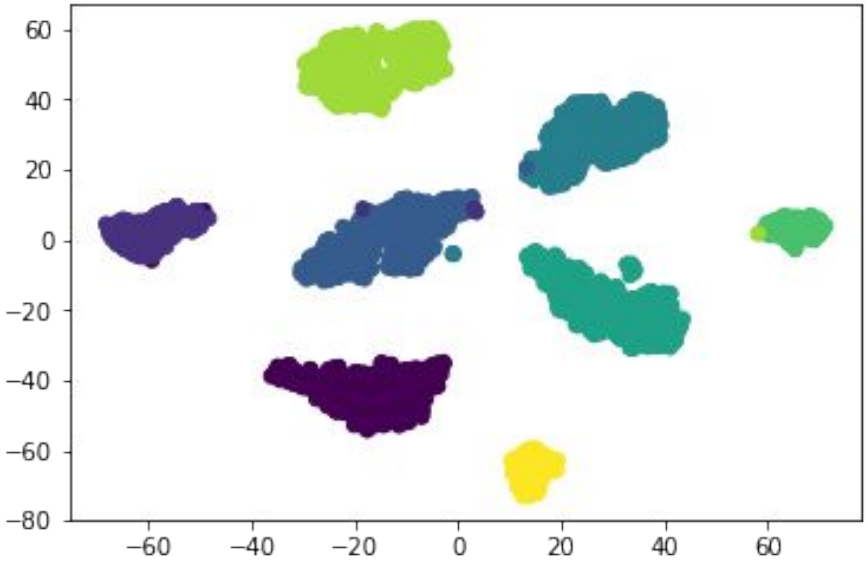

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018

# Task
# Feature Visualization

- Used t-SNE to visualize structure in high-dimensional space

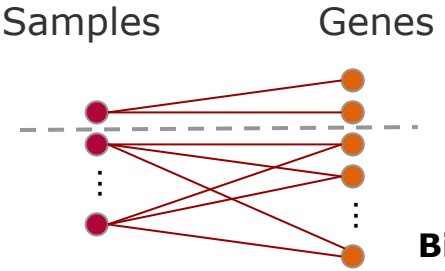


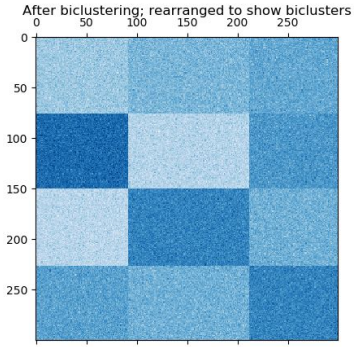

→ Clustering algorithm should definitely find groups

# Task
# Biclustering

- Spectral Coclustering
  - Bipartite graph between samples and genes
  - Edges: entry of the matrix
  - Find subgraphs using normalized cut

→ Assumes chessboard pattern
cannot focus on desired context




**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

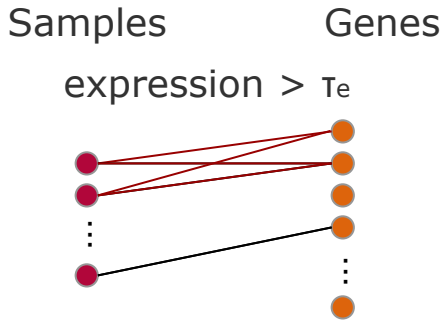
Willi Gierke
09.02.2018



http://scikit-learn.org/stable/_images/sphx_glr_plot_spectral_biclustering_0031.png

# Task
# Biclustering

- BiMax[1]: works on binary data
  - Divide and conquer algorithm
    ≡ find cliques in bipartite graph[2]
  - Edges: entry of the matrix over threshold
  - Find maximal cliques




**Bi-Clustering with Biological Context Information**

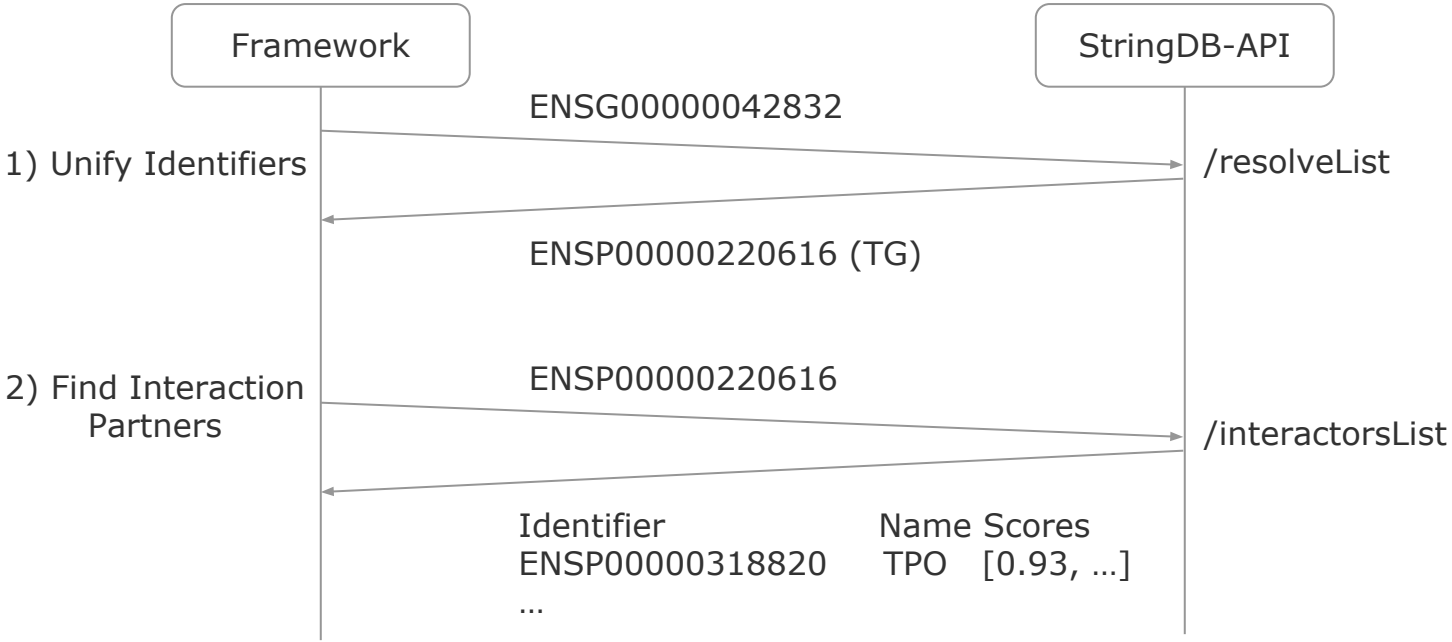
Trends in Bioinformatics

Willi Gierke
09.02.2018



1) A systematic comparison and evaluation of biclustering methods for gene expression data Prelić et al. 2006
2) Exact biclustering algorithm for the analysis of large gene expression data sets Voggenreiter et al. 2012

# Task
# Collecting Context Information




3) Add Context Information To Algorithm

**Bi-Clustering with Biological Context Information**

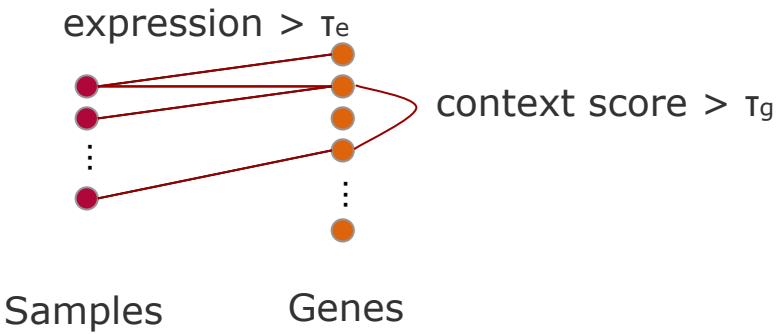
Trends in Bioinformatics

Willi Gierke
09.02.2018

# Algorithm
# Intuition

- Bipartite graph between set of samples and set of genes
- Connection if gene expression exceeds threshold $\tau_e$
- Gene connection if context score exceeds threshold $\tau_g$

→ Find biggest bi-cliques using Bron-Kerbosch algorithm




context score: based on StringDB

Are genes in the same pathway?
Do they occur in the same publication? ⋮

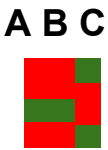
# Algorithm
# Choosing the Thresholds

- Expression threshold: choose with respect to average expression of that gene
- Gene connection threshold: “confidence that genes are connected”
  - Highly depends on score and underlying dataset

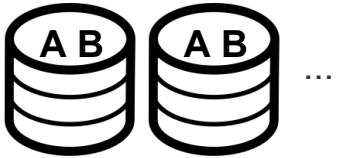
Textmining


PubMed

Coexpression




Database




KEGG GOC

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



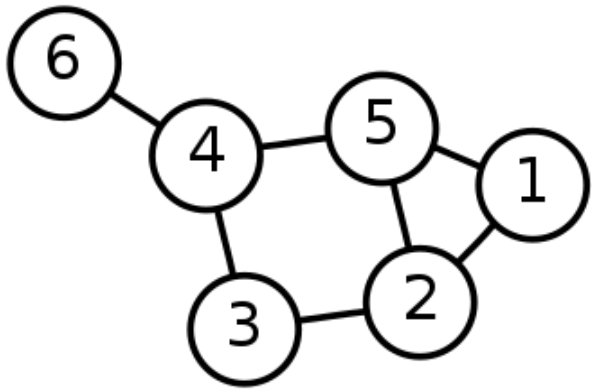
http://www.coreclimax.eu/sites/coreclimax.itc.nl/files/icons/Papers.png
https://d30y9cdsu7xlg0.cloudfront.net/png/15531-200.png

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
                MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
                P ← P \ {v}
                X ← X ∪ {v}
        end for
    end if
```




**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



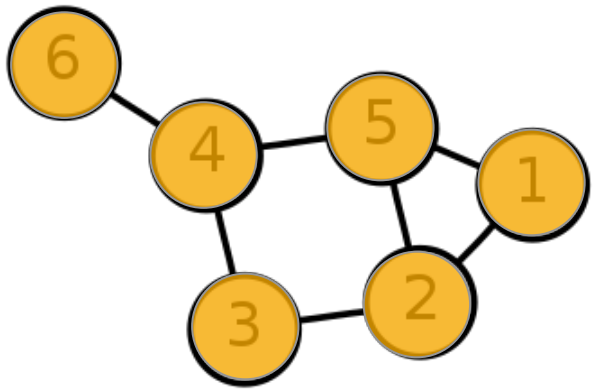
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




R P X
{} {1, 2, 3, 4, 5, 6} {}

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



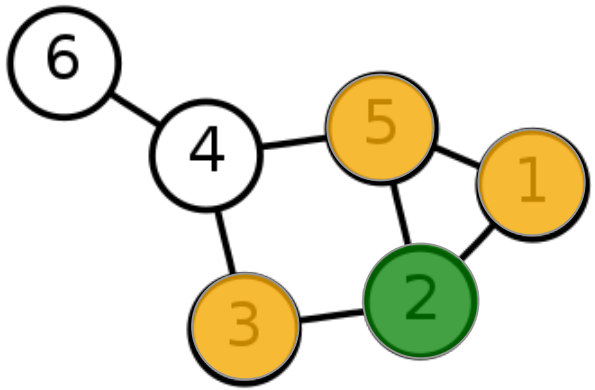
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




| v's | R | P | X |
|---|---|---|---|
| 2 | {} | {1, **2**, 3, 4, 5, 6} | {} |
| | {2} | {1, 3, 5} | {} |
| 4 | | | |
| 6 | | | |

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



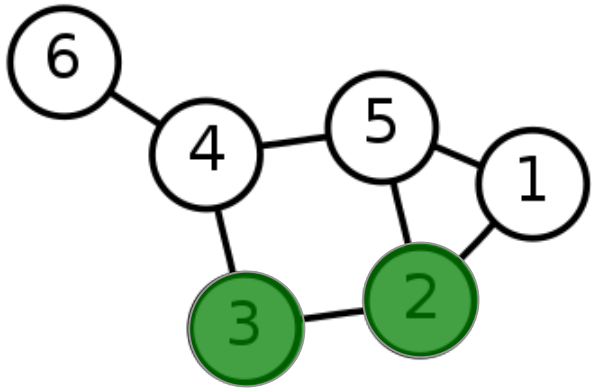
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




| v's | R | P | X |
|---|---|---|---|
| 2 | {} | {1, **2**, 3, 4, 5, 6} | {} |
| v's{2} | | {1, 3, **5**} | {} |
| 3 | {2, 3} | {} | {} → Clique: {2, 3} |
| 5 | | | |
| 4 | | | |
| 6 | | | |

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



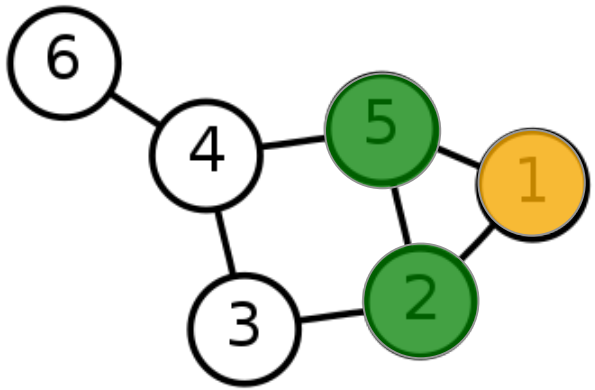
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




| v's | R | P | X |
|---|---|---|---|
| 2 | {} | {1, **2**, 3, 4, 5, 6} | {} |
| v's{2} | | {1, 3, **5**} | {} |
| 3 | | {2, 3} | {} → Clique: {2, 3} |
| 5 | | {2, 5} | {**1**} |
| 4 | | | |
| 6 | | | |

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



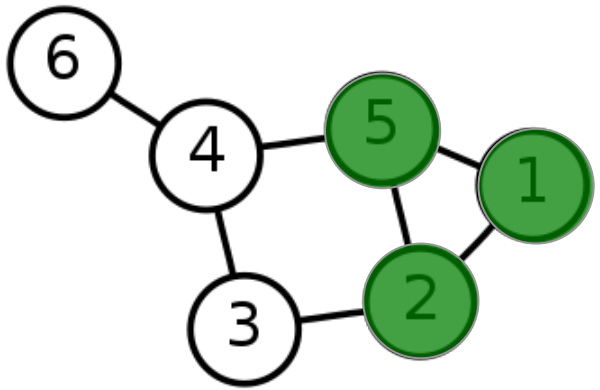
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




| v's | R | P | X | |
|---|---|---|---|---|
| 2 | {} | {1, **2**, 3, 4, 5, 6} | {} | |
| v's | {2} | {1, 3, **5**} | {} | |
| 3 | {2, 3} | {} | {} | → Clique: {2, 3} |
| 5 | v's {2, 5} | {**1**} | {} | |
| 1 | {2, 5, 1} | {} | {} | →Clique:{2,5,1} |
| 4 | | | | |
| 6 | | | | |

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke 09.02.2018



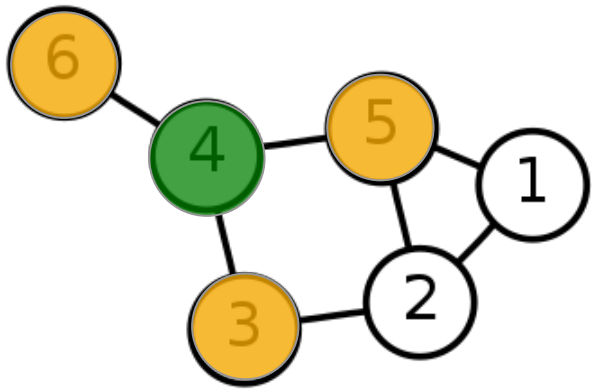
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
            MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
            P ← P \ {v}
            X ← X ∪ {v}
        end for
    end if
```




| v's | R | P | X |
|---|---|---|---|
| 2 | {} | {1, **2**, 3, 4, 5, 6} | {} |
| v's | {2} | {1, 3, **5**} | {} |
| 3 | {2, 3} | {} | {} → Clique: {2, 3} |
| 5 | v's {2, 5} | {**1**} | {} |
| 1 | {2, 5, 1} | {} | {} →Clique:{2,5,1} |
| 4 | {4} | {3, 5, 6} | {} |
| 6 | | ⋮ | |

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke 09.02.2018



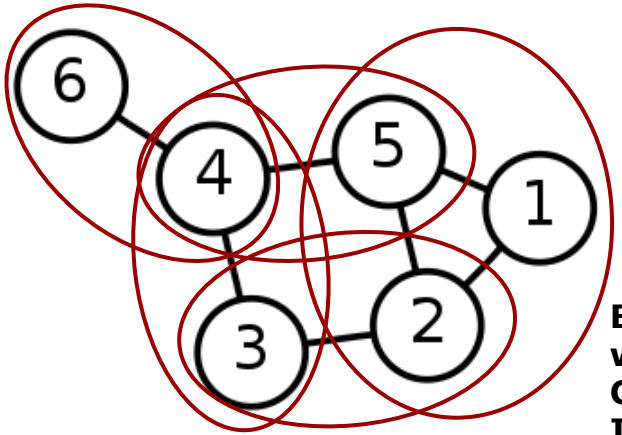
https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Algorithm
# Finding Maximal Cliques with Bron-Kerbosch

**Algorithm 2** Maximal Similarity Cliques

**Input:** $G = (V, E)$
**Output:** *Cliques*

```
MSC (R, P, X):
    if P == ∅ and X == ∅ then
        Cliques ← R
        return
    else
        pivot ← P ∪ X
        for every vertex v in P\N(pivot) do
                MSC (R ∪ {v}, P ∩ N(v), R ∩ N(v))
                P ← P \ {v}
                X ← X ∪ {v}
        end for
    end if
```




5 Maximal Cliques:
1 2 5
2 3
3 4
4 5
4 6
Can be found in $O(3^{n/3})$

**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



https://en.wikipedia.org/wiki/Bron%E2%80%93Kerbosch_algorithm
http://media.springernature.com/full/springer-static/image/art%3A10.1186%2Fs40537-016-0042-7/MediaObjects/40537_2016_42_Figb_HTML.gif

# Evaluation

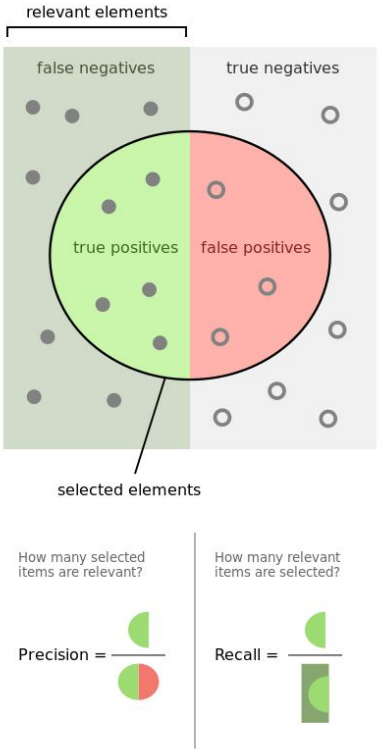
- Is bicluster G plausible?

  → Proportion of pairs of genes in G for which

  - There exists a connecting path (precision)
  - There exists no connecting path (recall)

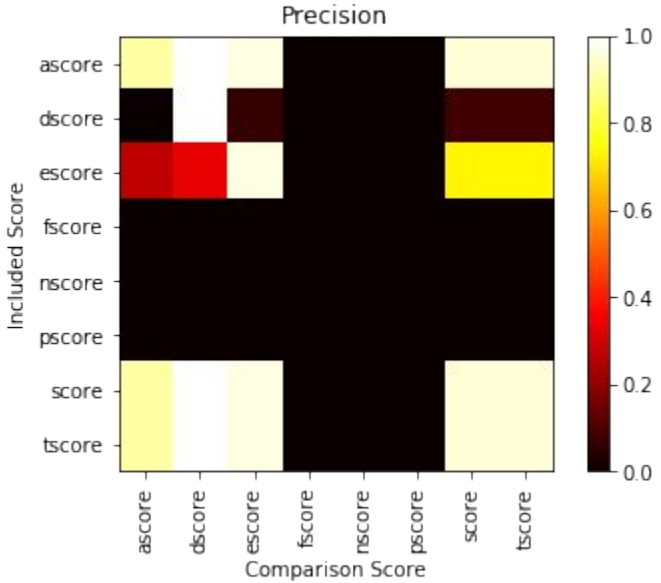
- StringDB offers various scores and combines them

  → Cluster based on one score, evaluate against another




Based on: A systematic comparison and evaluation of biclustering methods for gene expression data by Prelić et al.
https://upload.wikimedia.org/wikipedia/commons/thumb/2/26/Precisionrecall.svg/440px-Precisionrecall.svg.png

- Coexpression, textmining and combined score very useful
- Experimental score of middle quality
- Functional, neighborhood and phyletic profile score bad
  - Not helpful
  - No data




**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018

# Evaluation
# Biological Context

- Cluster purity not desired since biological context is interesting
- Based on ascore, compared against ascore: Keratin 5, Keratin 17 belong to same cluster, even if they are not associated based on the score
- "The gene expression cluster defining basal epithelial cells included keratin 5, keratin 17, integrin-β4, and laminin …"

→ use PathwayLinker[1] to evaluate whether genes are connected via pathways

Cancer as a Paradigm for Translational and Clinical Biomedical Research

César Serrano[1], George D. Demetri[2], in Clinical and Translational Science (Second Edition), 2017

Translation of Cancer Gene Expression Profiling to the Clinic

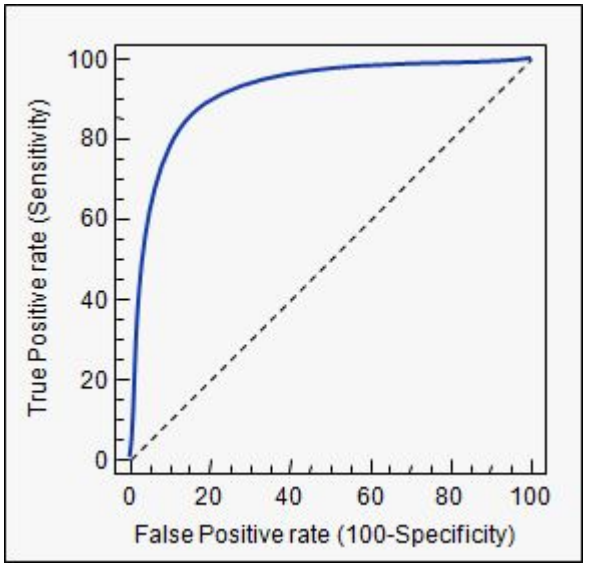
The hypothesis that phenotypic diversity of human cancer might be accompanied by a corresponding diversity in gene expression patterns eventually took shape in 2000. In their seminal study, Perou et al. (2000) proved that systematic investigation of gene expression patterns captured with complementary DNA (cDNA) microarrays led to an improved molecular taxonomy of human breast cancer (Perou et al., 2000). Briefly, cDNA microarray technology consists of labeling RNA samples obtained from patients and control subjects with distinguishable fluorescent dyes and hybridized to gene-specific probes composed of single strands of cDNA (Fodor et al., 1993). Relative levels of gene expression are estimated by measuring the fluorescence intensity of each probe. A hierarchical clustering method is used to group experimental samples on the basis of similarity in their patterns of expression. This technology was first used in a set of 65 surgical specimens of human breast tumors from 42 different individuals. In this study, two broad subgroups of breast cancer could be defined based on the lineage of the two types of cells present in the human mammary gland: basal cells and luminal cells. The gene expression cluster defining basal epithelial cells included keratin 5, keratin 17, integrin-β4, and laminin, whereas ER and ER-associated transcription factors clustered in a



http://pathwaylinker.org

# Evaluation
## Behavior of Varying Thresholds

- Vary thresholds $\tau_e$ and $\tau_g$ e.g. using a ROC curve?
- disadvantage might be:
  - highly depending on thresholds
  - highly depending on used datasets for biological context






https://i.stack.imgur.com/5x3Xj.png

# Lessons Learned

- Handling genes is cumbersome due to numerous identifiers
- Various databases are intransparent
- Results of publications barely reproducible

# Conclusion

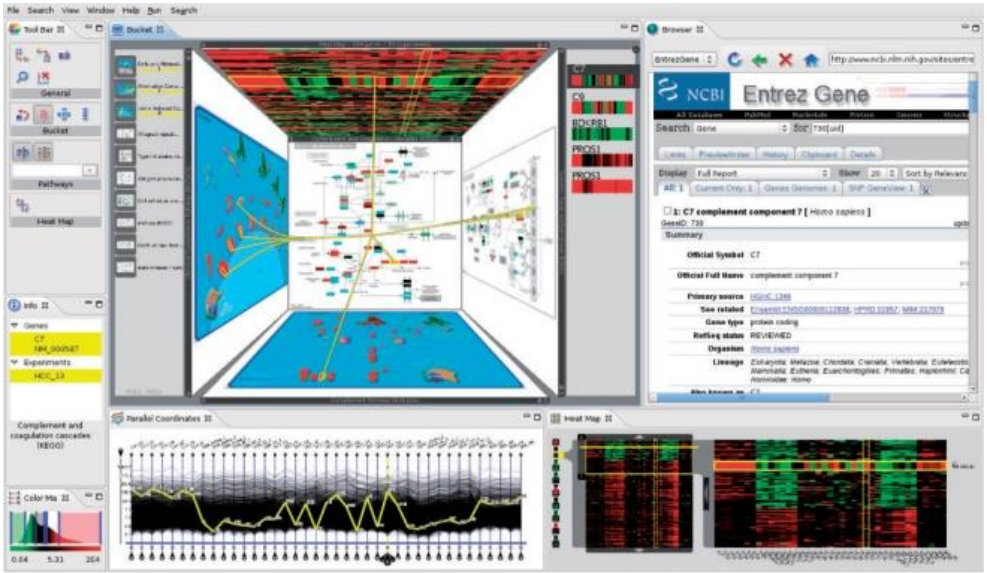
- Context-awareness helps to
  - Find patterns
  - Focus on defined biological processes

- Biclustering mostly NP-complete
  - Can not consider all >19.000 genes without pruning them
  - Lossy heuristics necessary
    - BiMax: binary edges
    - Spectral Coclustering: no overlap

# Discussion

- Other Evaluation Approaches?
- Algorithm Extensions?
- Algorithm Alternatives?
- Different Context Information?

- Gene Expression Data + Pathways → Pathway Network[1]
- Caleydo: connecting pathways and gene expression, Streit et al. 2009




**Bi-Clustering with Biological Context Information**

Trends in Bioinformatics

Willi Gierke
09.02.2018



1) Pathway network inference from gene expression data, Ponzoni et al. 2013