Biclustering in drug design

Djork-Arné Clevert, Günter Klambauer, Andreas Mayr, Andreas Mitterecker, Ulrich Bodenhofer, Martin Heusel, and Sepp Hochreiter

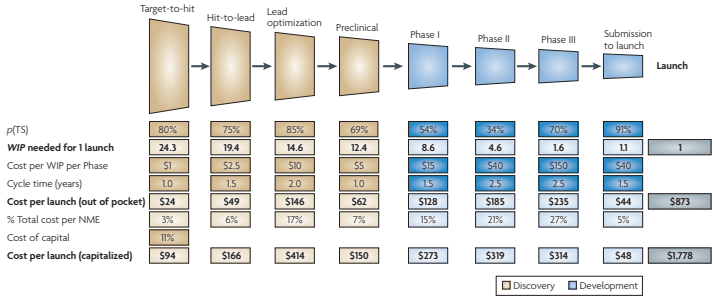
Institute of Bioinformatics, Johannes Kepler University Linz

Potsdam, September 25th, 2012

Pharmaceutical "ice age"

- key patent expirations
- cost-constrained healthcare system
- prescription of generic drugs
- more regulatory requirements
- decreasing number of new drugs approved by the US Food and Drug Administration (FDA) or the European Medicines Agency (EMEA)
- rapidly rising R&D cost

R&D costs for one new molecular entity



Paul et al., Nature Reviews Drug Discovery, 2010, 9(3):203-214

Motivation

- drug candidate failing in Phase I yields a "out of pocket" cost of $428M ("capitalized" cost $610M)
- approx. 80% promising drug candidates fail before end of Phase I (e.g. undetected toxicity)

- **Aim:** Increase the productivity of the R&D process and avoid expensive late-stage clinical failures
  - de-risk drug candidates during the early preclinical stages
  - reduce the time gap between the selection of drug candidate and the identification of potential side effects
- **Data:** Transcriptional effects of the drug candidate on a cell line (gene expression), phenotypic data (biological assays), chemical structure and properties (chemotypes)
- **Problem:** How to extract the relevant information?

Biclustering applications in QSTAR

- **Definition:** Biclustering simultaneously organizes a data matrix into subsets of rows and columns in which the entities of each row subset are similar to each other on the column subset and vice versa.
- Gene expression
  - columns compounds, rows genes ⇒ e.g. compounds that trigger the same pathway
- Bioassays
  - columns compounds, rows bioassay activity ⇒ e.g. compounds that are active on similar targets
- Structural fingerprints
  - columns compounds, rows fingerprints ⇒ e.g. compounds that share a chemical substructure

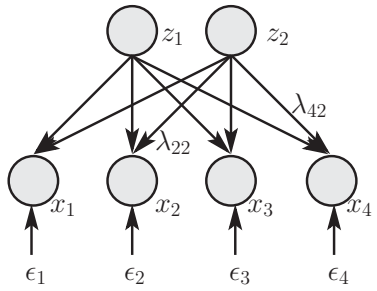
FABIA: The model I

factor z

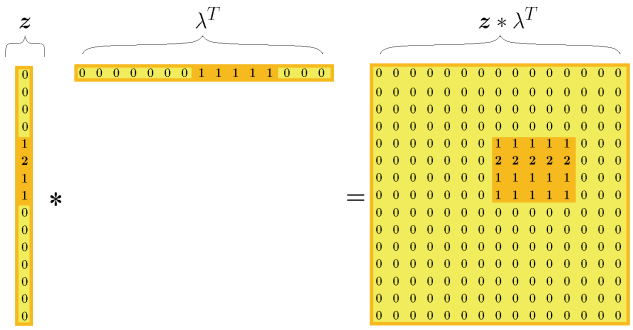
loading matrix Λ

observations x

noise ϵ



Biclustering: The idea



FABIA: The model II

$$x = \boldsymbol{\lambda} z \; + \; \boldsymbol{\epsilon} \; = \sum_{i=1}^{p} \lambda_i \; z_i^T \; + \; \boldsymbol{\epsilon}$$

- x are the observations
- $\boldsymbol{\lambda}$ is the matrix of factor loadings
- $z = (z_1, \ldots, z_p)^T$ is the factor matrix
- p number of biclusters
- $\lambda_i \in \mathbb{R}^n$ is the sparse prototype vector of the i-th bicluster
- $z_i \in \mathbb{R}^l$ is the sparse vector of factors of the i-th bicluster
- $\epsilon \in \mathbb{R}^{n \times l}$ is additive noise
- the noise is independent of z
- ϵ is $\mathcal{N}(0, \boldsymbol{\Psi})$-distributed (the covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ is diagonal)

FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

Factor prior

- Sparseness on the factor
- Laplace prior
- $p(z) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{i=1}^{p} e^{-\sqrt{2}|z_i|}$

Problem

Laplace prior on factors leads to intractable likelihood:

$$p(x \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) \;=\; \int p(x \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Solution: Prior on factors is replaced by maximum of a Gaussian function family $\Rightarrow$ variational approach

$$p(z) \;\approx \arg\max_{\xi} \; p(z|\xi)$$

Motivation Biclustering Conclusions

FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

Factor prior

- Sparseness on the factor
- Laplace prior
- $p(z) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{i=1}^{p} e^{-\sqrt{2}|z_i|}$

Problem

Laplace prior on factors leads to intractable likelihood:

$$p(x \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) \;=\; \int p(x \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Solution: Prior on factors is replaced by maximum of a Gaussian function family $\Rightarrow$ variational approach

$$p(z) \;\approx \arg\max_{\xi} \; p(z|\xi)$$

Biclustering in drug design 9

Motivation Biclustering Conclusions

FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

Factor prior

- Sparseness on the factor
- Laplace prior
- $p(z) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{i=1}^{p} e^{-\sqrt{2}|z_i|}$

Problem

Laplace prior on factors leads to intractable likelihood:

$$p(x \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) \;=\; \int p(x \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Solution: Prior on factors is replaced by maximum of a Gaussian function family $\Rightarrow$ variational approach

$$p(z) \;\approx \arg\max_{\xi} \; p(z|\xi)$$

Biclustering in drug design 9

FABIA: Variational EM updates

E-step:

$\mathrm{E}(\tilde{z}_j \mid x_j) \;=\; \left(\boldsymbol{\lambda}^T \, \boldsymbol{\Psi}^{-1} \, \boldsymbol{\lambda} \,+\, \Xi_j^{-1}\right)^{-1} \boldsymbol{\lambda}^T \, \boldsymbol{\Psi}^{-1} \, x_j$ and

$\mathrm{E}(\tilde{z}_j \, \tilde{z}_j^T \mid x_j) \;=\; \left(\boldsymbol{\lambda}^T \, \boldsymbol{\Psi}^{-1} \, \boldsymbol{\lambda} \,+\, \Xi_j^{-1}\right)^{-1} \,+\, \mathrm{E}(\tilde{z}_j \mid x_j) \; \mathrm{E}(\tilde{z}_j \mid x_j)^T$ where

$\Xi_j \;=\; \mathrm{diag}\left(\mathrm{diagvect}\left(\sqrt{\mathrm{E}(\tilde{z}_j \, \tilde{z}_j^T \mid x_j)}\right)\right)$ is the update for the variational parameter.

M-step:

$$\boldsymbol{\lambda}^{\mathrm{new}} \;=\; \frac{\frac{1}{l} \sum_{j=1}^{l} x_j \, \mathrm{E}(\tilde{z}_j \mid x_j)^T \;-\; \frac{\alpha}{l} \, \boldsymbol{\Psi} \, \mathrm{sign}(\boldsymbol{\lambda})}{\frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{z}_j \, \tilde{z}_j^T \mid x_j)}$$

$$\boldsymbol{\Psi}^{\mathrm{new}} \;=\; \boldsymbol{\Psi}^{\mathrm{EM}} \,+\, \mathrm{diag}\left(\mathrm{diagvect}\left(\frac{\alpha}{l} \, \boldsymbol{\Psi} \, \mathrm{sign}(\boldsymbol{\lambda}) (\boldsymbol{\lambda}^{\mathrm{new}})^T\right)\right) \;, \quad \text{where}$$

$$\boldsymbol{\Psi}^{\mathrm{EM}} \;=\; \mathrm{diag}\left(\mathrm{diagvect}\left(\frac{1}{l} \sum_{j=1}^{l} x_j x_j^T \;-\; \boldsymbol{\lambda}^{\mathrm{new}} \frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{z}_j \mid x_j) \; x_j^T\right)\right).$$

α controls the degree of sparseness (parameter of the Laplacian prior)

FABIA: Variational EM updates

E-step:

$\mathrm{E}(\tilde{z}_j \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \boldsymbol{\Xi}_j^{-1})^{-1} \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} x_j$ and

$\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \boldsymbol{\Xi}_j^{-1})^{-1} + \mathrm{E}(\tilde{z}_j \mid x_j)\, \mathrm{E}(\tilde{z}_j \mid x_j)^T$ where

$\boldsymbol{\Xi}_j = \mathrm{diag}(\mathrm{diagvect}(\sqrt{\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j)}))$ is the update for the variational parameter.

M-step:

$$\boldsymbol{\lambda}^{\mathrm{new}} = \frac{\frac{1}{l}\sum_{j=1}^{l} x_j\, \mathrm{E}(\tilde{z}_j \mid x_j)^T - \frac{\alpha}{l}\boldsymbol{\Psi}\,\mathrm{sign}(\boldsymbol{\lambda})}{\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j)}$$

$$\boldsymbol{\Psi}^{\mathrm{new}} = \boldsymbol{\Psi}^{\mathrm{EM}} + \mathrm{diag}(\mathrm{diagvect}(\frac{\alpha}{l}\boldsymbol{\Psi}\,\mathrm{sign}(\boldsymbol{\lambda})(\boldsymbol{\lambda}^{\mathrm{new}})^T)), \quad \text{where}$$

$$\boldsymbol{\Psi}^{\mathrm{EM}} = \mathrm{diag}(\mathrm{diagvect}(\frac{1}{l}\sum_{j=1}^{l} x_j x_j^T - \boldsymbol{\lambda}^{\mathrm{new}} \frac{1}{l}\sum_{j=1}^{l}\mathrm{E}(\tilde{z}_j \mid x_j)\, x_j^T)).$$

α controls the degree of sparseness (parameter of the Laplacian prior)

Biclustering of bioassays and compounds

Matrix plot (dim 270,000 x 4,000)

Bioassay data details

- Data source: ChEMBL
- # of assays: ca. 4,000
- # of compounds: ca. 270,000
- Sparseness: ca. 1:2,000

FABIA paper

BIOINFORMATICS **ORIGINAL PAPER** Vol. 26 no. 12 2010, pages 1520–1527
doi:10.1093/bioinformatics/btq227

Gene expression Advance Access publication April 23, 2010

FABIA: factor analysis for bicluster acquisition

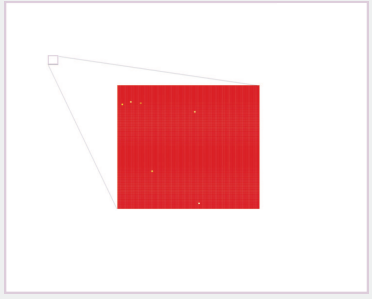
Sepp Hochreiter[1,*], Ulrich Bodenhofer[1], Martin Heusel[1], Andreas Mayr[1], Andreas Mitterecker[1], Adetayo Kasim[2], Tatsiana Khamiakova[2], Suzy Van Sanden[2], Dan Lin[2], Willem Talloen[3], Luc Bijnens[3], Hinrich W. H. Göhlmann[3], Ziv Shkedy[2] and Djork-Arné Clevert[1,4]

[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria, [2]Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, [3]Johnson & Johnson Pharmaceutical Research & Development, Division of Janssen Pharmaceutica, Beerse, Belgium and [4]Department of Nephrology and Internal Intensive Care, Charité, Berlin, Germany

- Developed for -omics data, where FABIA outperformed all prevalent biclustering methods
- Bioassay data and fingerprint data are **sparse** and **big** ⇒ computational expensive factorization ⇒ improvements required
  - sparse matrix algebra where only existing entries are coded
  - an efficient computation and a very low memory footprint

Biclustering of bioassays and compounds

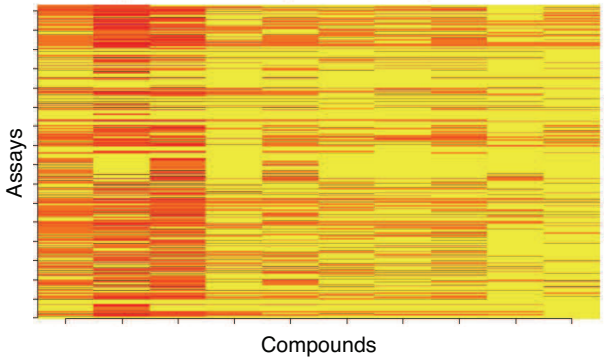
Matrix plot - close up (dim 100 x 100)



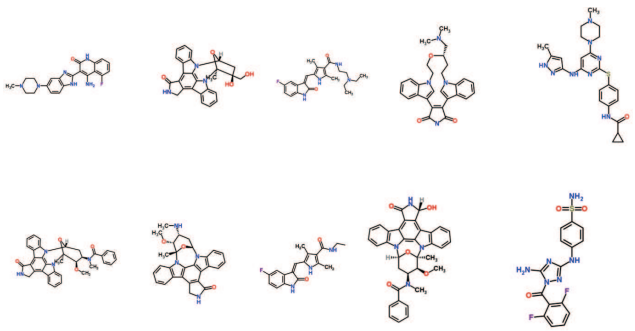
Bioassay data details

- Data source: ChEMBL
- # of assays: ca. 4,000
- # of compounds: ca. 270,000
- Sparseness: ca. 1:2,000

Biclustering of bioassays and compounds



Compounds of the bioassay bicluster



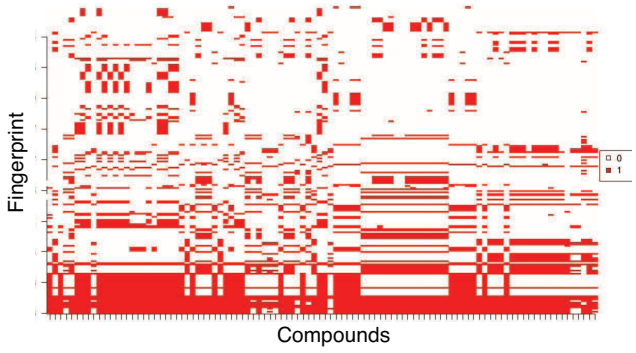
Biclustering of fingerprints and compounds

Matrix plot (dim 16e+6 x 1e+6)

Bioassay data details

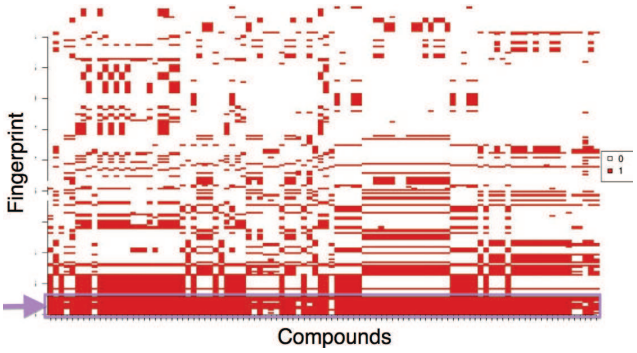
- Data source: ChEMBL
- # of fingerprints: ca. 16,000,000
- # of compounds: ca. 1,000,000
- Sparseness: ca. 1:150,000

Bicluster of fingerprints and compounds I

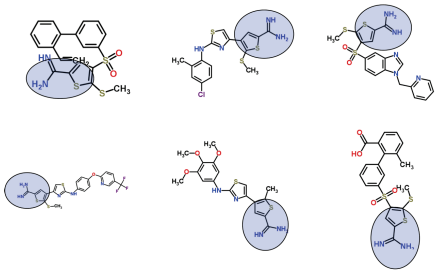


Computational costs ca. 3h (single core)

Bicluster of fingerprints and compounds I

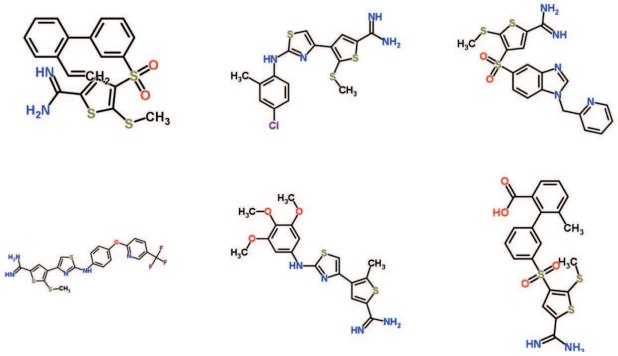


Compounds of the fingerprint bicluster



- All compounds of this bicluster show kinase bioactivity (urokinase-type plasminogen activator)

Compounds of the fingerprint bicluster



Conclusion

- FABIA is a generative model for biclustering of high-dimensional data
- Optimized for sparse big data sets
- Biclustering in drug design can help in selecting compounds with strong on-target effects and might help to impute missing measurements
- Identifies in ChEMBL biclusters that contain compounds with a shared chemical substructure.
- These substructures could be related to bioactivity via the compounds which are screened on bioassays.

Acknowledgments

- All members of the IWT project, especially Hinrich Göhlmann

Open source software

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- FABIA is publicly available as Bioconductor R packages
- Software homepages:
  - http://www.bioinf.jku.at/software/fabia/fabia.html