Biomarker Discovery:

Controlling false discoveries in high dimensional situations

Benjamin Hofner

Institut für Medizininformatik, Biometrie und Epidemiologie
Friedrich-Alexander-Universität Erlangen-Nürnberg

35th Annual Conference of the ISCB
Vienna - 2014

in cooperation with
Markus Göker, DSMZ, Germany
Luigi Boccuto, GCC, USA

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Identification of Biomarkers for ASD patients (yes/no)

Autism Spectrum Disorders (ASD):

- relatively common neurodevelopmental disease
- biological basis incompletely determined
- no laboratory test for these conditions
- (relatively) hard to diagnose

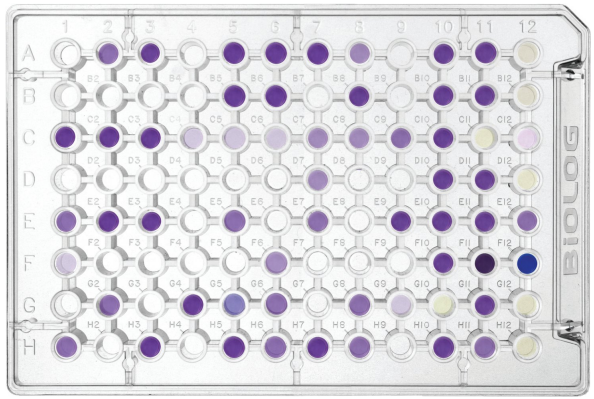
Aim:

Detect differentially expressed amino acid pathways,

i.e., amino acid pathways that differ between healthy subjects and ASD patients.

Identification of Biomarkers for ASD patients (yes/no)

- **Available Data:**
  - Cell lines of $n = 35$ subjects (17 ASD patients and 18 controls)
- **Measurements:**
  - Phenotype Microarrays (PM)
    - 96-well array per patient
    - Each well has a different carbon energy source
    - Maximum reaction ($=$ cellular activity) per well is measured (by a color reaction)

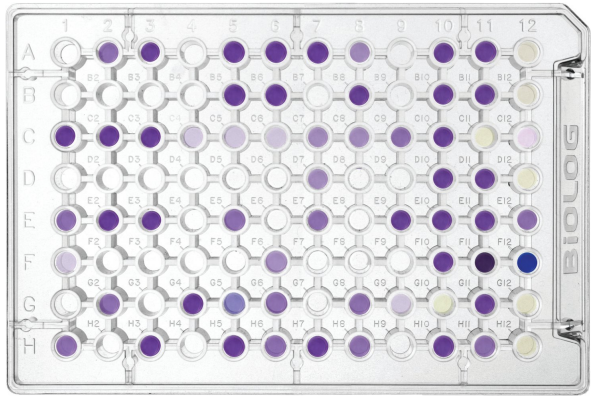


(Source: Biolog Inc., http://www.biolog.com)

- Measurements describe metabolism of subjects (on cell basis)

Identification of Biomarkers for ASD patients (yes/no)

- **Available Data:**
  - Cell lines of $n = 35$ subjects (17 ASD patients and 18 controls)
- **Measurements:**
  - Phenotype Microarrays (PM)
    - 96-well array per patient
    - Each well has a different carbon energy source
    - Maximum reaction ($=$ cellular activity) per well is measured (by a color reaction)



(Source: Biolog Inc., http://www.biolog.com)

- Measurements describe metabolism of subjects (on cell basis)

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Differential expressions modeled using interactions
- Use boosting methods with stability selection for variable selection

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

In short:

- The fitting process is iterative.
- It optimizes for example the (negative) log-likelihood.
- In each step of the algorithm only the effect $\hat{\beta}_{j*}$ of the best fitting variable is updated by adding a fraction $\nu \cdot \hat{\beta}_{j*}$ to the model (with e.g., $\nu = 0.1$).
- The main tuning parameter is the number if iterations.

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

In short:

- The fitting process is iterative.
- It optimizes for example the (negative) log-likelihood.
- In each step of the algorithm only the effect $\hat{\beta}_{j*}$ of the best fitting variable is updated by adding a fraction $\nu \cdot \hat{\beta}_{j*}$ to the model (with e.g., $\nu = 0.1$).
- The main tuning parameter is the number if iterations.

- ▶ Use cross-validation to find the optimal stopping iteration.
- ▶ With "early stopping" variable selection is achieved.

Practical notes

- We get an interpretable model, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via variable selection and shrinkage.

Practical notes

- We get an interpretable model, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via variable selection and shrinkage.

Yet,

- in high-dimensional settings, i.e., with many predictors, we might select a lot of uninformative variables.
- In many situations a **formal selection procedure with error control** seems advisable.

Stability Selection [Meinshausen and Bühlmann 2010]

- ... is a versatile approach, which can be combined with (all) high-dimensional variable selection approaches.
- ... is based on subsampling (▶ draw samples without replacement).
- ... controls the per-family error rate PFER $= \mathbb{E}(V)$, where V is the number of false positives.

Insertion

Overview of Error Rates [see e.g. Dudoit et al. 2003]

per-family error rate (PFER): $\mathbb{E}(V)$

per-comparison error rate (PCER): $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR): $\mathbb{E}(\frac{V}{R})$

	Keep H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
	$m - R$	R	m

Insertion

Overview of Error Rates [see e.g. Dudoit et al. 2003]

per-family error rate (PFER): $\mathbb{E}(V)$

per-comparison error rate (PCER): $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR): $\mathbb{E}(\frac{V}{R})$

Note: The PFER is very conservative

- For fixed α, PFER is more conservative than FWER
  FWER is more conservative than PCER
- For fixed α, FWER is more conservative than FDR
  and thus PFER is more conservative than FDR

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

- Conservative upper bound for the per-family error rate (PFER):

$$\text{PFER} = \mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

(if exchangeability assumption holds for all noise variables)

Improved Stability Selection [Shah and Samworth 2013]

- Tighter, i.e., less conservative error bounds can be derived under certain conditions.

  a) **If distribution of (simultaneous) selection probabilities is unimodal:**

$$\mathbb{E}(V) \leq \frac{q^2}{c(\pi_{\text{thr}}, B) \cdot p} \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

  b) **If distribution of (simultaneous) selection probabilities is r-concave:**

$$\begin{aligned} \mathbb{E}(V) &\leq \min\left\{ D\left(2\pi_{\text{thr}} - 1; \frac{q^2}{p^2}, B, -\frac{1}{2}\right), D\left(\pi_{\text{thr}}; \frac{q}{p}, 2B, -\frac{1}{4}\right) \right\} p \\ &\leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \end{aligned}$$

Improved Stability Selection [Shah and Samworth 2013]

- Tighter, i.e., less conservative error bounds can be derived under certain conditions.

  a) **If distribution of (simultaneous) selection probabilities is unimodal:**

$$\mathbb{E}(V) \leq \frac{q^2}{c(\pi_{\text{thr}}, B) \cdot p} \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

  b) **If distribution of (simultaneous) selection probabilities is r-concave:**

$$\begin{aligned} \mathbb{E}(V) &\leq \min\left\{ D\left(2\pi_{\text{thr}} - 1; \frac{q^2}{p^2}, B, -\frac{1}{2}\right), D\left(\pi_{\text{thr}}; \frac{q}{p}, 2B, -\frac{1}{4}\right) \right\} p \\ &\leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \end{aligned}$$

Condition b) is stronger than a) and might not always hold, especially for larger numbers of subsamples B. Thus a) is usually recommended.

Implementation

- Stability selection is implemented in the *R* package **mboost** [Hothorn et al. 2014, Hofner et al. 2014] in the function

  `stabsel()`

- **mboost** also implements

  `stabsel_parameters(cutoff, q, PFER)`

  to compute error bounds for combinations of two of the three parameters without running the resampling algorithm.

- Stability selection can also be used with other fitting functions.

Implementation

- Stability selection is implemented in the *R* package **mboost** [Hothorn et al. 2014, Hofner et al. 2014] in the function

  `stabsel()`

- **mboost** also implements

  `stabsel_parameters(cutoff, q, PFER)`

  to compute error bounds for combinations of two of the three parameters without running the resampling algorithm.
- Stability selection can also be used with other fitting functions.

Practical recommendation:

- Choose an upper bound for the PFER and specify *either* q or π_{thr}.
- Check that the computed value is sensible (e.g., that is q large enough if π_{thr} and PFER were specified).

Identification of Biomarkers for ASD patients (yes/no)

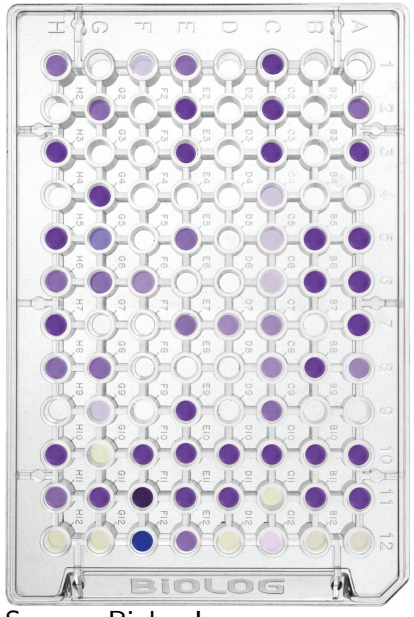
1) Obtain amino acid pathway annotation for each well
   using the R package **opm** [Vaas et al. 2013]
2) Fit main effects model for maximum reaction (`y`),
   given disease status (`group`), pathway annotation (`pathway`), and patient (`id`)

```
y ~ group + pathway + (1 | id)
```

   - Each well constitutes one observation!
   - Each well can belong to multiple pathways!
3) Use 2) as offset model and add group specific effects:

```
y ~ ... + pathway:group
```

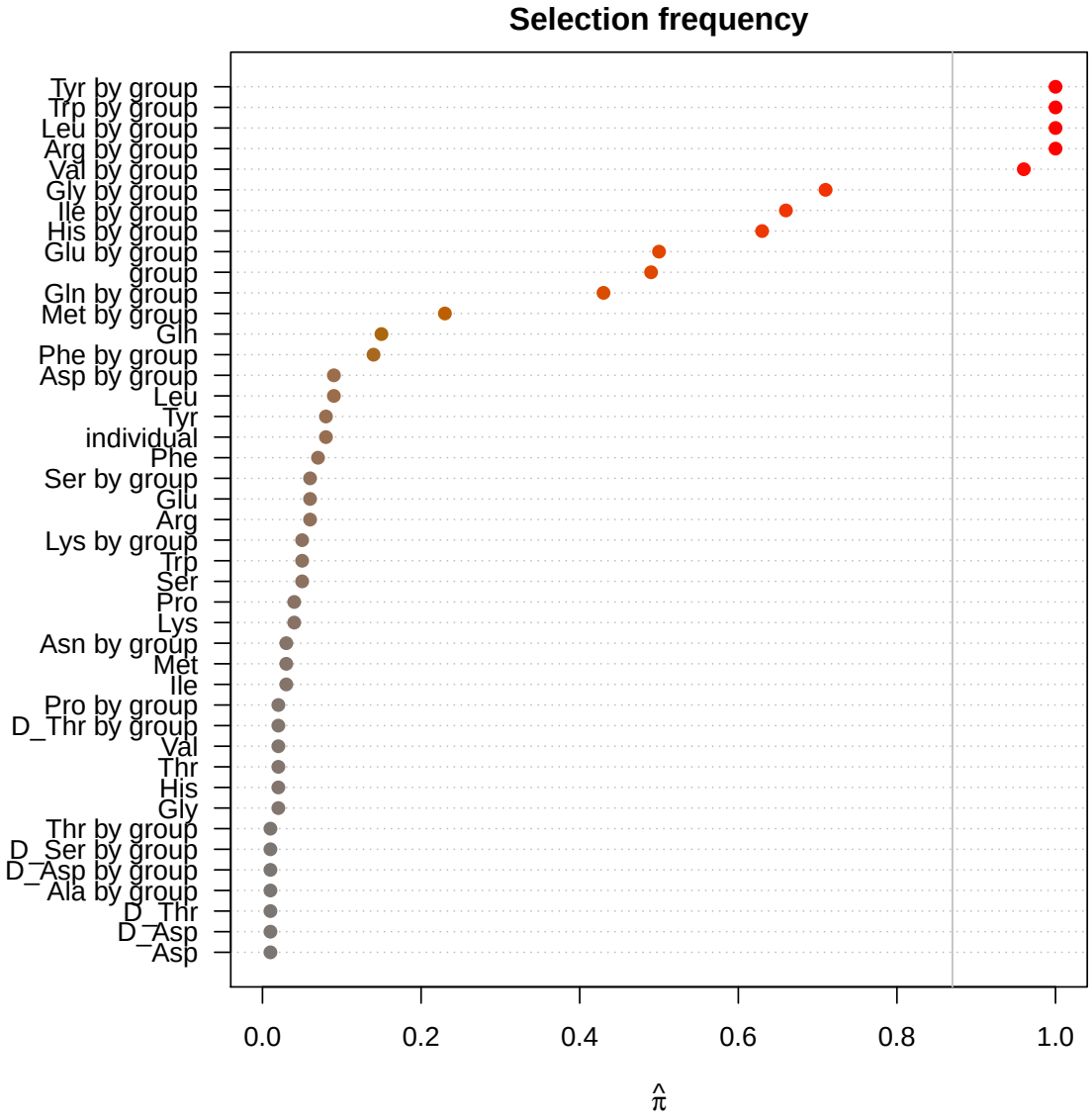
4) **Which of the group-specific pathway effects is selected additionally to the offset model (with PFER ≤ 1)?**



Source: Biolog Inc.

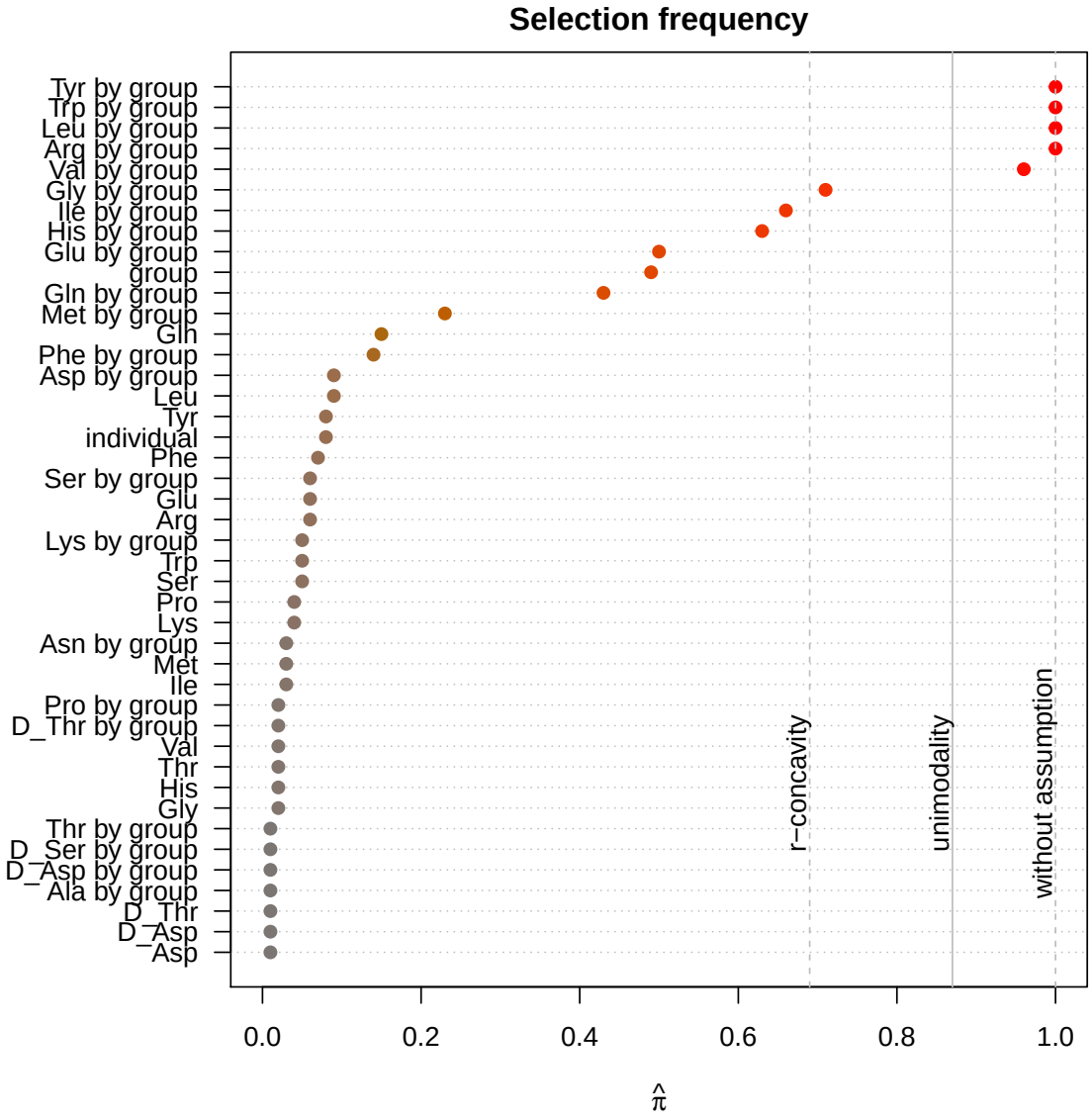
Results: Biomarkers for ASD

Stability selection with PFER ≤ 1 and $q = 10$



Results: Biomarkers for ASD

Stability selection with PFER ≤ 1 and $q = 10$



Results: Biomarkers for ASD

Differentially Expressed Amino Acids

- tyrosine (`Tyr`), tryptophan (`Trp`), leucine (`Leu`), arginine (`Arg`)
  - selection frequency $\hat{\pi} = 100\%$
- valine (`Val`)
  - selection frequency $\hat{\pi} = 96\%$
- glycine (`Gly`)
  - selection frequency $\hat{\pi} = 71\%$

Results: Biomarkers for ASD

Differentially Expressed Amino Acids

- tyrosine (`Tyr`), tryptophan (`Trp`), leucine (`Leu`), arginine (`Arg`)
  - selection frequency $\hat{\pi} = 100\%$
- valine (`Val`)
  - selection frequency $\hat{\pi} = 96\%$
- glycine (`Gly`)
  - selection frequency $\hat{\pi} = 71\%$

Biomedical Conclusion

- Confirms abnormal metabolism of tryptophan in ASD cells [see Boccuto et al. 2013]
- \+ Additional amino acids seem to be affected, although on a milder level
- Suggest an abnormal metabolism of large amino acids

Summary and Outlook

- Stability selection works well in conjunction with boosting.
- It controls the PFER and is especially useful in sparse, high-dimensional settings.
- Stability selection results in a fundamentally new solution, which cannot be recreated elsewise

  (e.g. by selecting a certain regularization parameter, here the stopping iteration m_{stop}).
- Stability selection can also be used for boosted GAM models and other fitting approaches.

- Yet, stability selection is quite conservative
  - as it controls the PFER
  - and as even this control seems to be conservative (at least for the standard error bound).
- Higher selection numbers (i.e. higher TPR) can be obtained by tighter error bounds, yet, sometimes error bounds do not hold any more.

Slides and further information available from
`http://benjaminhofner.de`

References

Luigi Boccuto, Chin-Fu Chen, Ayla Pittman, Cindy Skinner, Heather McCartney, Kelly Jones, Barry Bochner, Roger Stevenson, and Charles Schwartz. Decreased tryptophan metabolism in patients with autism spectrum disorders. *Molecular Autism*, 4(1):16, 2013.

Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

Benjamin Hofner, Andreas Mayr, Nicolay Robinzonov, and Matthias Schmid. Model-based boosting in R – A hands-on tutorial using the R package mboost. *Computational Statistics*, 29:3–35, 2014.

Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. *mboost: Model-Based Boosting*, 2014. URL `http://CRAN.R-project.org/package=mboost`. R package version 2.3-0.

Nicolai Meinshausen and Peter Bühlmann. Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 417–473, 2010.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:55–80, 2013.

Lea A. I. Vaas, Johannes Sikorski, Benjamin Hofner, Nora Buddruhs, Anne Fiebig, Hans-Peter Klenk, and Markus Göker. opm: An R package for analysing Omnilog®Phenotype MicroArray data. *Bioinformatics*, 29(14):1823–1824, 2013.