



Biomarker discovery in cancer transcriptomic data using network-based regularization

Marta Belchior Lopes

marta.lopes@fct.unl.pt

ML4Microbiome Workshop University of Tirana, May 24, 2022







Precision Medicine

Precision Medicine



Adapted from "Advancing Biomarker Development Through Convergent Engagement", *Mol Imaging Biol*, DOI:10.1007/s11307-019-01361-2

Precision Medicine



From "Advancing Biomarker Development Through Convergent Engagement", *Mol Imaging Biol*, DOI:10.1007/s11307-019-01361-2

From "Data integration and predictive modeling methods for multi-omics datasets", *Molecular Omics*, 2018, DOI:10.1039/C7MO00051K

Genome

•••

methylated

Genome

'OMICS data

- Multiple sources of data
- ► High dimentionality
- ► Data heterogeneity

Classification

N≪p

Feature selection

Sparse Logistic regression

$$P(Y_i = 1 | \mathbf{X}_i) = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}$$

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i \log P(Y_i = 1 | X_i) + (1 - y_i) \log[1 - P(Y_i = 1 | X_i)] \right\}$$



Regularized classification



Feature selection

Sparse Logistic regression

$$P(Y_i = 1 | \mathbf{X}_i) = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}$$



$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log P(Y_i = 1 | \mathbf{X}_i) + (1 - y_i) \log[1 - P(Y_i = 1 | \mathbf{X}_i)] \right\} + \boxed{F(\boldsymbol{\beta})}$$
$$F(\boldsymbol{\beta}) = \lambda \left\{ \alpha \| \boldsymbol{\beta} \|_1 + (1 - \alpha) \| \boldsymbol{\beta} \|_2^2 \right\}$$

Elastic net (Zou & Hastie, 2005), lasso (Tibshirani, 1986)

Regularized classification

N≪p

Feature selection

Sparse Logistic regression

$$P(Y_{i} = 1 | \mathbf{X}_{i}) = \frac{\exp(\mathbf{X}_{i}^{T} \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_{i}^{T} \boldsymbol{\beta})}$$
$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_{i} \log P(Y_{i} = 1 | \mathbf{X}_{i}) + (1 - y_{i}) \log[1 - P(Y_{i} = 1 | \mathbf{X}_{i})] \right\} + \boxed{\mathbb{F}(\boldsymbol{\beta})}$$
$$F(\boldsymbol{\beta}) = \lambda \left\{ \alpha \| \mathbf{W} \boldsymbol{\phi} \boldsymbol{\beta} \|_{1} + (1 - \alpha) \| \mathbf{W} \boldsymbol{\phi} \boldsymbol{\beta} \|_{2}^{2} \right\} \quad \begin{array}{l} \text{Weighted} \\ \text{Elastic net} \end{array}$$

Elastic net (Zou & Hastie, 2005), lasso (Tibshirani, 1986)



From "Network hub-node prioritization of gene regulation with intra-network association", *BMC Bioinformatics* **21**, 101 (2020). DOI:10.1186/s12859-020-3444-7

Network-based regularized classification

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log P(Y_i = 1 | \mathbf{X}_i) + (1 - y_i) \log[1 - P(Y_i = 1 | \mathbf{X}_i)] \right\} + F(\boldsymbol{\beta})$$
$$F(\boldsymbol{\beta}) = \lambda \left\{ \alpha \| \mathbf{w} \circ \boldsymbol{\beta} \|_1 + (1 - \alpha) \| \mathbf{w} \circ \boldsymbol{\beta} \|_2^2 \right\} \quad \substack{\text{Weighted} \\ \text{Elastic net}}$$



Twin Networks Recovery (TWINER) penalty

Promotes the selection of similarly correlated nodes in two gene networks

Network-based regularized classification



$$\Sigma_A = \left[\boldsymbol{\sigma}_1^A, ..., \boldsymbol{\sigma}_p^A\right]$$
 and $\Sigma_B = \left[\boldsymbol{\sigma}_1^B, ..., \boldsymbol{\sigma}_p^B\right]$

$$d_j(A,B) = \arccos \frac{\langle \boldsymbol{\sigma}_j^A, \boldsymbol{\sigma}_j^B \rangle}{\|\boldsymbol{\sigma}_j^A\| \cdot \|\boldsymbol{\sigma}_j^B\|}, \quad j = 1, \dots, p$$

$$w_j = \frac{d_j(A,B)}{\max_k d_k(A,B)}, \quad j,k = 1,\ldots,p$$

 $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log P(Y_i = 1 | \mathbf{X}_i) + (1 - y_i) \log[1 - P(Y_i = 1 | \mathbf{X}_i)] \right\} + F(\boldsymbol{\beta})$ $F(\boldsymbol{\beta}) = \lambda \left\{ \alpha \left[\mathbf{w} \circ \boldsymbol{\beta} \right]_1 + (1 - \alpha) \left[\mathbf{w} \circ \boldsymbol{\beta} \right]_2^2 \right\} \quad \begin{array}{c} \text{Weighted} \\ \text{Elastic net} \end{array}$

Network-based regularized classification



Lopes *et al. BMC Bioinformatics* (2019) 20: 356 https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Check for updates

Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes^{1,2*} ^(D), Sandra Casimiro³ and Susana Vinga^{2,4}

- Breast and Prostate cancers
- Transcriptomic data
- Regularized logistic regression twiner penalty

Lopes *et al. BMC Bioinformatics* (2019) 20: 356 https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Check for updates

Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes^{1,2*} ^(D), Sandra Casimiro³ and Susana Vinga^{2,4}

		#\/arc	AL	AUC		# Miscl	
Breast and Prostate cancers Transcriptomic data Regularized logistic regression – twiner penalty			Train	Test	Train	Test	
	EN	58	0.98	0.98	17	9	
	twiner	69	0.99	0.98	15	9	

Lopes *et al. BMC Bioinformatics* (2019) 20: 356 https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE



Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes^{1,2*} ⁽¹⁾, Sandra Casimiro³ and Susana Vinga^{2,4}

- Breast and Prostate cancers
- Transcriptomic data
- Regularized logistic regression twiner penalty



Lopes *et al. BMC Bioinformatics* (2019) 20: 356 https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Check for updates

Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes^{1,2*} ^(D), Sandra Casimiro³ and Susana Vinga^{2,4}

- Breast and Prostate cancers
- Transcriptomic data
- Regularized logistic regression twiner penalty



Lopes et al. BMC Bioinformatics (2019) 20:356 https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

updates

Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes^{1,2*} ^(D), Sandra Casimiro³ and Susana Vinga^{2,4}

- Breast and Prostate cancers
- Transcriptomic data
- Regularized logistic regression twiner penalty



Gliomas – the most common brain tumors

► Many subtypes (**glioblastoma**, the most aggressive)



Gliomas – the most common brain tumors

► Many subtypes (**glioblastoma**, the most aggressive)

Heterogeneity

intertumoral heterogeneity



► intratumoral heterogeneity





The glioblastoma case

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59 https://doi.org/10.1186/s12859-020-3390-4

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data

Marta B. Lopes^{1*} ^(D) and Susana Vinga²

Intratumoral heterogeneity

Single-cell RNA-seq data



Glioblastoma (GBM)

- ▶ Single-cell RNA-seq data (Darmanis et al., 2017)
- ► Four primary GBM patients
- ▶ 3,589 cell and 23,368 genes



The glioblastoma case



Lopes and Vinga BMC Bioinformatics (2020) 21:59 https://doi.org/10.1186/s12859-020-3390-4

The glioblastoma case





Lopes and Vinga *BMC Bioinformatics* (2020) 21:59 https://doi.org/10.1186/s12859-020-3390-4



Projects





Multi-Omic NETworks in gliomas

(SFRH/PTDC/CCI-BIO/4180/2020)





Gliomas – the most common brain tumors

► Many subtypes (**glioblastoma**, the most aggressive)

Heterogeneity

intertumoral heterogeneity



► intratumoral heterogeneity







GOAL

Identification of molecular **biomarkers** associated with tumor **heterogeneity** in **gliomas**, aiming at improving patient **diagnosis**, **prognosis**, and **therapeutic** decisions.

































Part of the Computational Biology book series (COBO,volume 31)



Gene Module Detection





Gene Module Detection

Differential Network Analysis





Gene Module Detection

Differential Network Analysis

Network Regularization





Gene Module Detection

Differential Network Analysis

Network Regularization

Causal Inference





Gene Module Detection

Differential Network Analysis

Network Regularization

Causal Inference

Deep Learning





DATABASES









National Cancer Institute at the National Institutes of Health







Team



Thanks!

marta.lopes@fct.unl.pt