ALGEBRAIC SHORTCUTS FOR CROSS-VALIDATION OF SUPERVISED NETWORK PREDICTION AT LIGHT SPEED

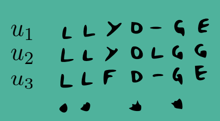
Michiel Stock, T. Pahikkala, A. Airola, J. Meys, W. Waegeman and B. De Baets
michiel.stock@UGent.be
KERMIT, Ghent University
@michielstock

In brief

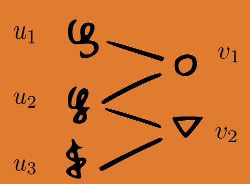
Supervised machine learning methods are important tools to predict biological networks, such as protein-ligand-, protein-protein- or gene regulatory networks. Incorrect validation schemes can lead to an overly optimistic performance estimation. We developed kernel-based framework for light speed training, tuning and validating models for any relevant prediction setting.

Supervised network prediction

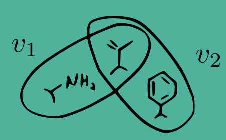
Data required: experimentally determined biological network with two similarity matrices describing the nodes.



e.g. alignment-based similarity matrix K



e.g. protein-ligand adjacency matrix Y



e.g. molecular similarity matrix G

Two-step kernel ridge regression

Our framework learns a pairwise kernel-based predictor of the form:

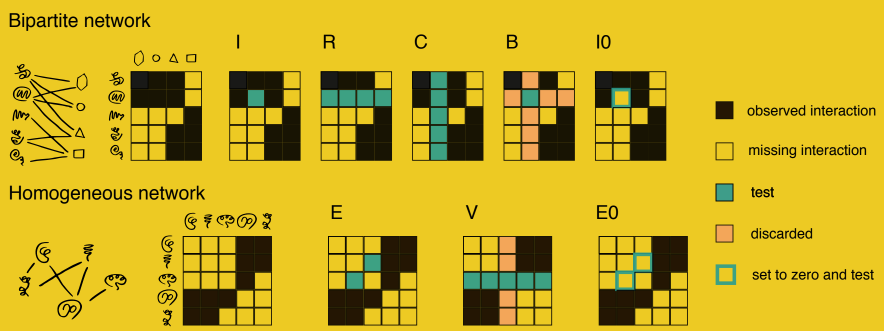
$$f(u,v)=\sum_i\sum_j W_{ij}k(u,u_i)g(v,v_j)\,.$$

The weight vector can easily be obtained by:

$$W=(K+\lambda_u\mathbb{I})^{-1}Y(G+\lambda_v\mathbb{I})^{-1}\,,$$

with λ_u and λ_v two regularization parameters.

Different cross-validation settings



Our algebraic shortcuts can compute leave-one-out cross-validation for any setting with a **constant time complexity**.

More information

The two-step kernel ridge regression is part of a theoretically well-founded pairwise learning framework. An R package **xnet** is currently in development. Scan the QR code on top for the preprint and a link to the software.

Experimental results

Our shortcuts can perform cross-validation several orders of magnitude faster than a naive implementation.