Link Prediction

May 18, 2017

Karsten Borgwardt, ETH-Department BSSE in Basel

Content:

- What is link prediction?
- Which are the most popular link prediction algorithms?
- What are typical pitfalls in link prediction?

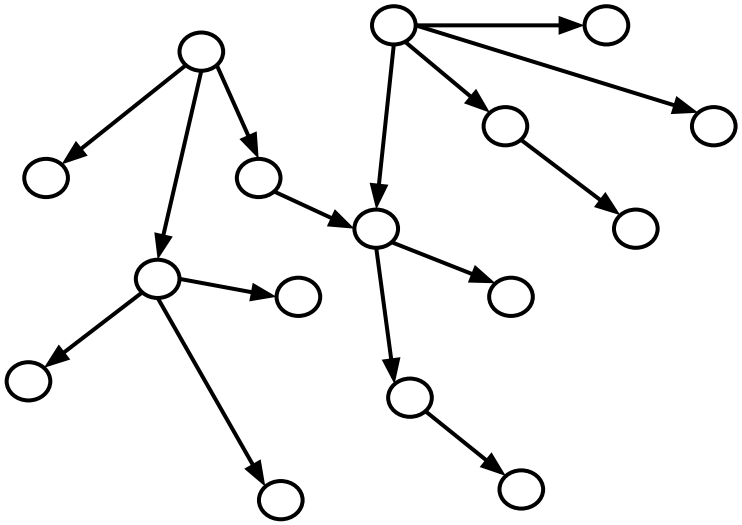
Link Prediction - Definition

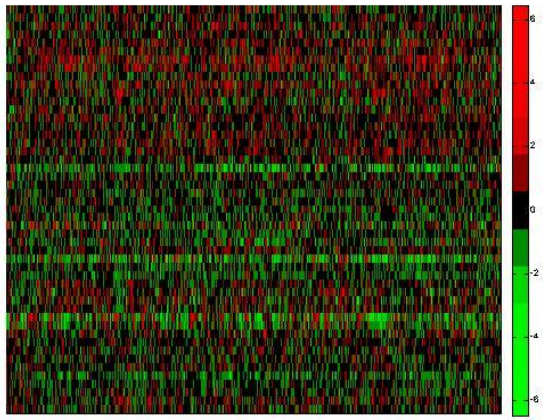
What is link prediction?

Given a network, predict the existence of missing edges in this *incomplete* network.

Link Prediction - Motivation

Discovery of Networks of Gene Regulation and Gene Co-Expression

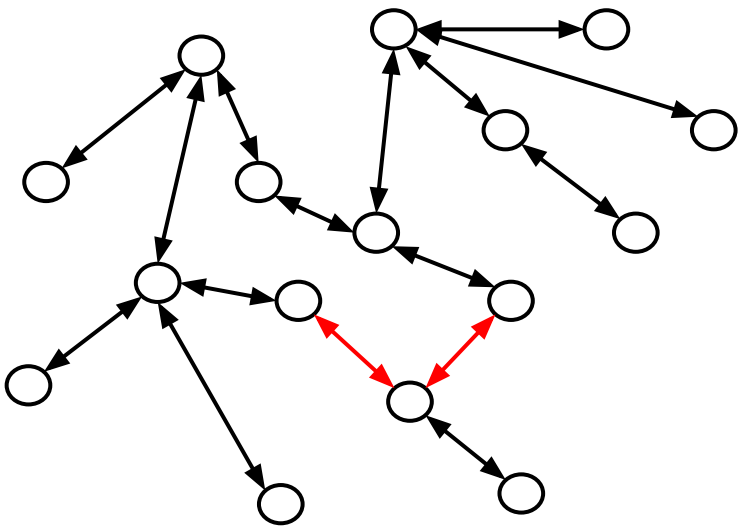


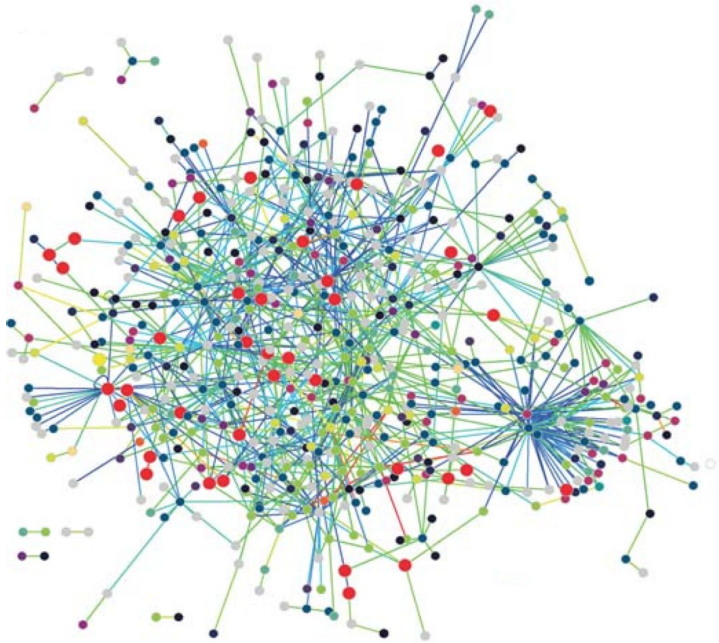


- Which transcription factor regulates which gene?
- Which genes are being co-expressed?

Link Prediction - Motivation

Protein-protein interaction prediction

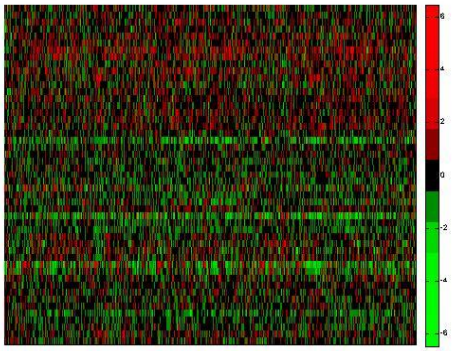




Fitz et al., PLoS One 2008

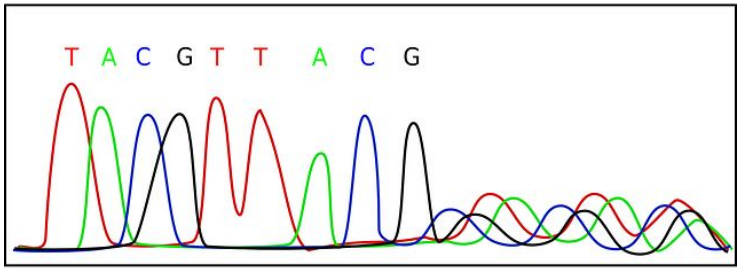
- Which edges are missing from the network?

Link Prediction - Typical Setup

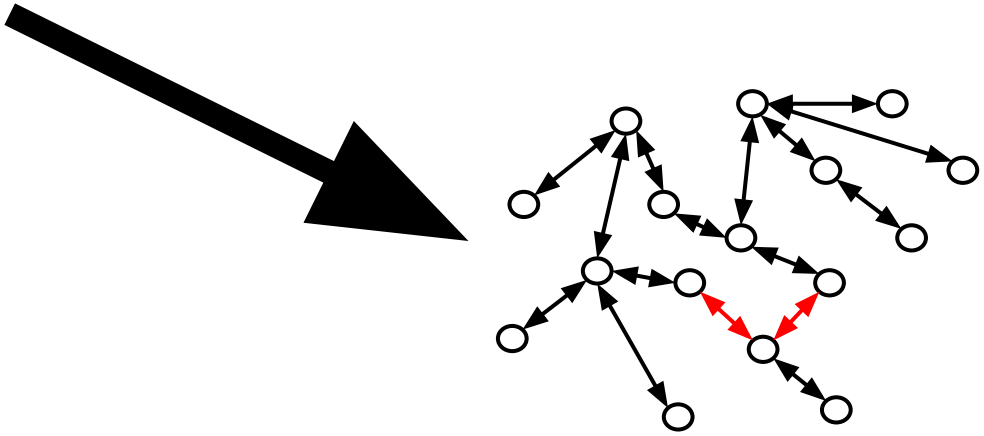


Gene expression data

Known network structure



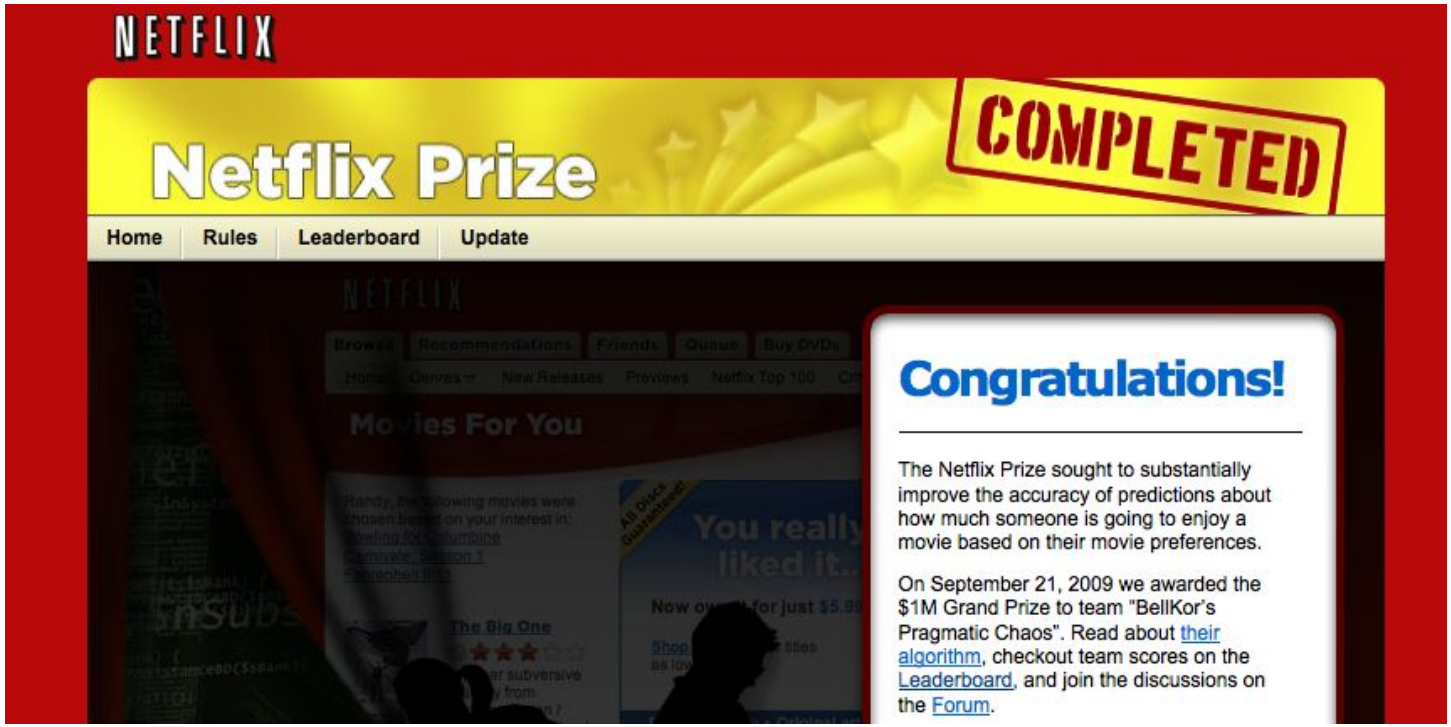
Gene/protein sequences



Link prediction

Link Prediction - Plethora of Methods

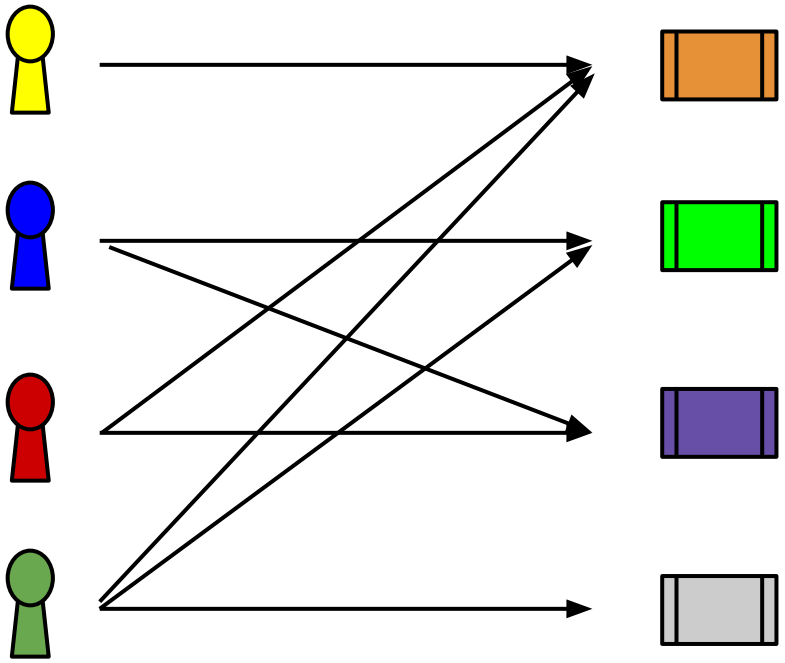
Link prediction or *collaborative filtering* is of huge importance even outside biology.



Source: netflixprize.com

Link Prediction - Plethora of Methods

Link prediction or *collaborative filtering* in many fields:



35 different link predictions methods were used in the DREAM5 Challenge (Marbach et al., 2012).

Link Prediction - DREAM Challenges

DREAM 3 - In Silico Network Challenge (Prill et al., 2010)

Sub-challenge	Network	Nodes	Edges	Regulators
***In Silico* Size 10**	Ecoli1	10	11	5
	Ecoli2	10	15	3
	Yeast1	10	10	7
	Yeast2	10	25	8
	Yeast3	10	22	9
***In Silico* Size 50**	Ecoli1	50	62	13
	Ecoli2	50	82	11
	Yeast1	50	77	26
	Yeast2	50	160	37
	Yeast3	50	173	35
***In Silico* Size 100**	Ecoli1	100	125	26
	Ecoli2	100	119	19
	Yeast1	100	166	60
	Yeast2	100	389	71
	Yeast3	100	551	81

In each of the three sub-challenges the number of nodes was held constant but the number of edges and regulator nodes was not. There were five gold standard networks in each of the three sub-challenges (which were treated as three separate contests).

doi:10.1371/journal.pone.0009202.t002

Link Prediction - DREAM Challenges

DREAM 3 - In Silico Network Challenge (Prill et al., 2010)

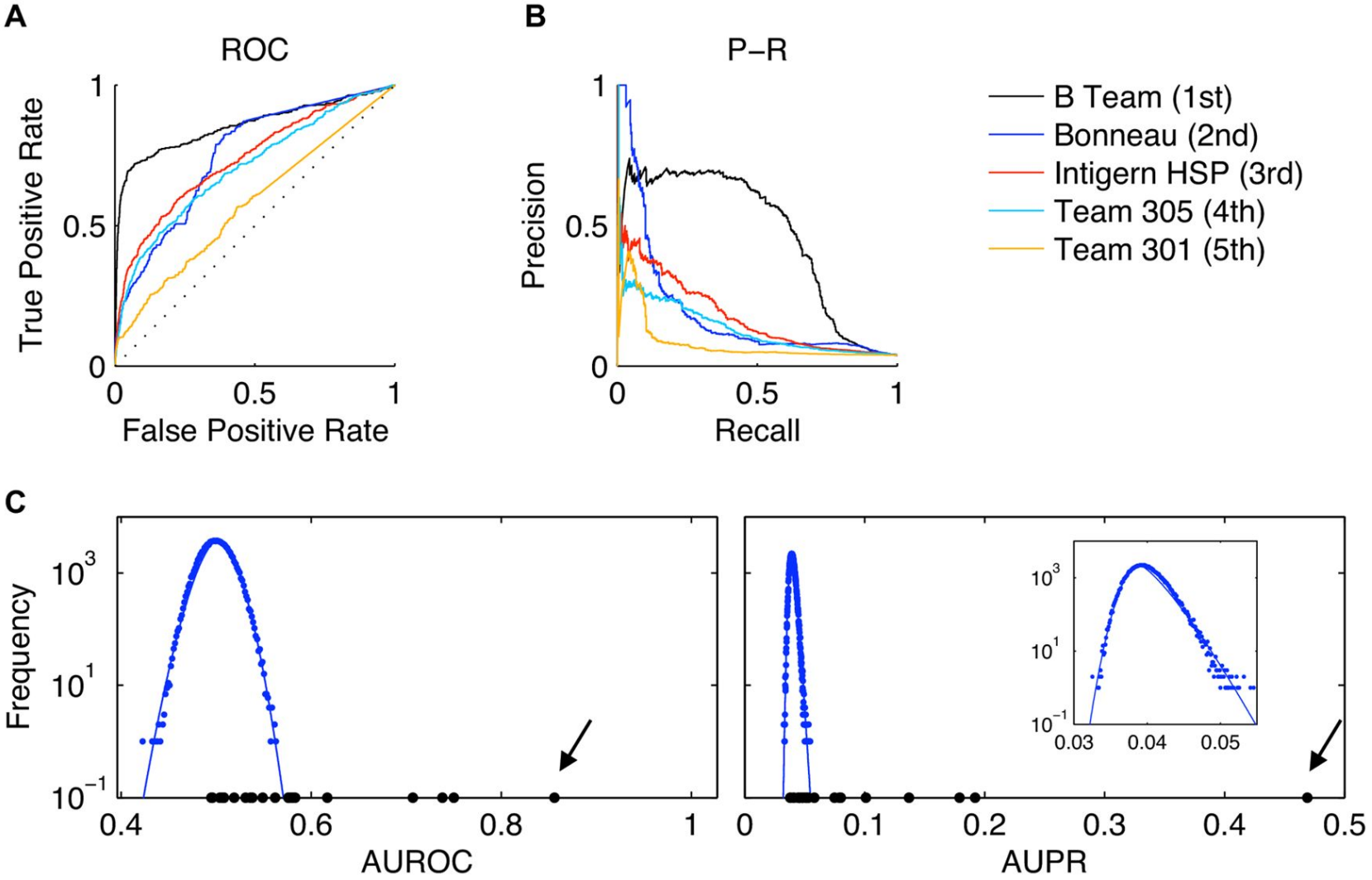
Source Node	Target Node	Confidence	Scoring Cutoff (k)
G85	G1	1.00	1
G85	G10	0.99	2
G10	G85	0.73	3
G99	G52	0.44	4
⋮	⋮	⋮	⋮
G10	G3	0.01	N(N-1)

Predicted edges were to be ranked from most confidence to least confidence that the edge is present in the network. A directed edge is denoted by a source and target node and an arbitrary (non-increasing) score between one (most confidence) to zero (least confidence). Thus, edges that are predicted to exist in the network should be at the top of the list and those predicted not to exist in the network should be at the bottom of the list. To evaluate the predicted network, two metrics—area under the ROC curve and area under the precision-recall curve—were computed by scanning all possible decision boundaries (i.e., $k = 1$, $k = 2$, etc.) up to the maximum number of possible directed edges (excluding self-edges).

doi:10.1371/journal.pone.0009202.t003

Link Prediction - DREAM Challenges

DREAM 3 - In Silico Network Challenge (Prill et al., 2010)



Link Prediction - Mode of Prediction

Which modes of link prediction do exist?

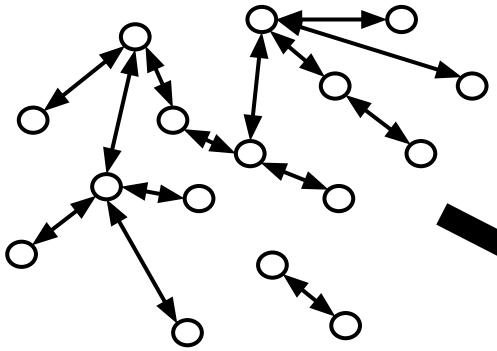
Unsupervised link prediction

Our prediction model is not based on learning from examples, but rather a predefined rule.

Supervised link prediction

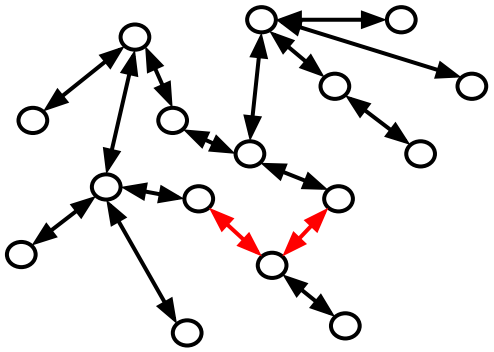
We are given examples of existing links and learn a prediction model based on these examples.

Link Prediction - Unsupervised Link Prediction



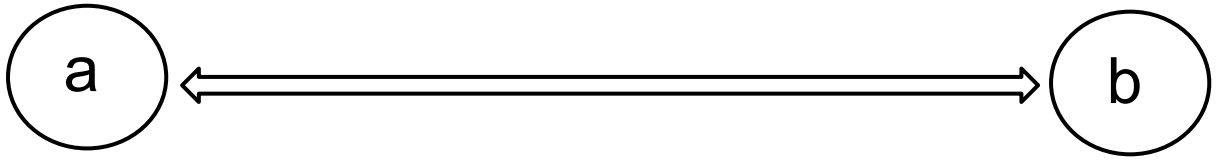
Original network

Prediction based on a rule that is not inferred from the data



Predicted links

Link Prediction - Similarity-Based Approach



Predict an edge between genes a and b if their similarity s(a,b) is above a threshold θ.

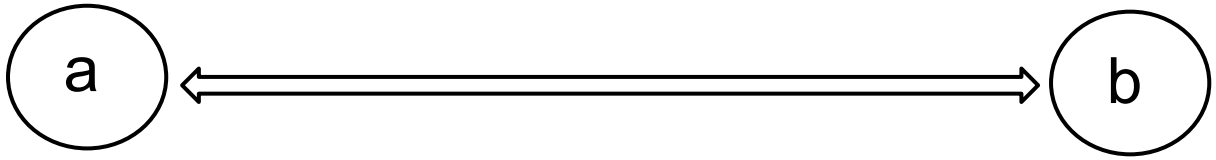
Advantages:

- Straightforward to implement
- Scales to large networks

Disadvantages:

- How to set θ?
- Is similarity really necessarily a condition for interaction?

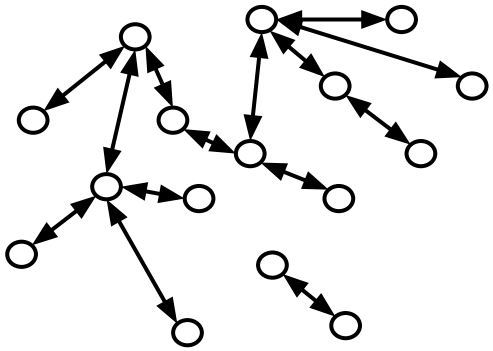
Link Prediction - Similarity-Based Approach



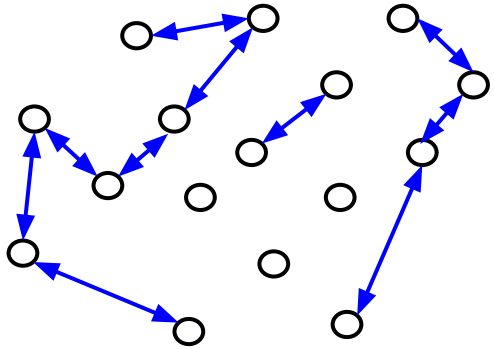
What are typical similarity measures *s(a,b)* in link prediction?

- Pearson's correlation coefficient
- Mutual Information
- String kernels that count common subsequences in two protein sequences (*k*-mers)
- Number of shared neighbors

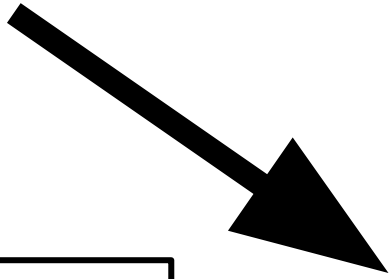
Link Prediction - Supervised Link Prediction



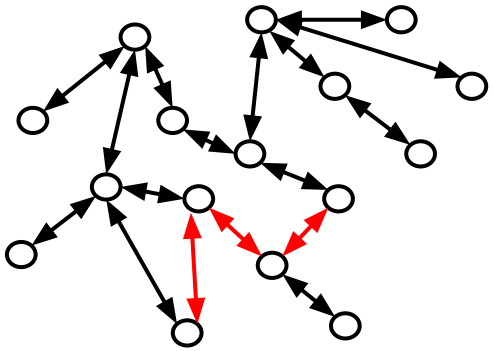
Known interactions



Known ‘non-interactions’

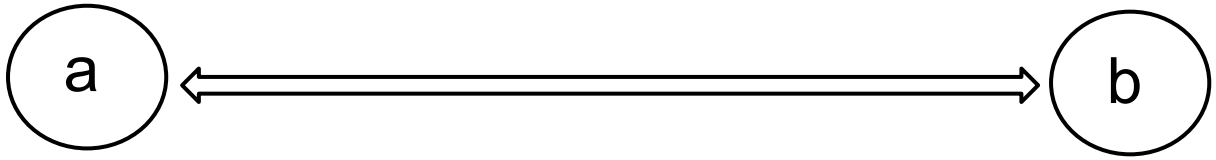


Predictor is learned from known interactions and non-interactions



Predicted links

Link Prediction - Cluster-Based Learning



Predict an edge between genes *a* and *b* if *a* and *b* belong to the same cluster.

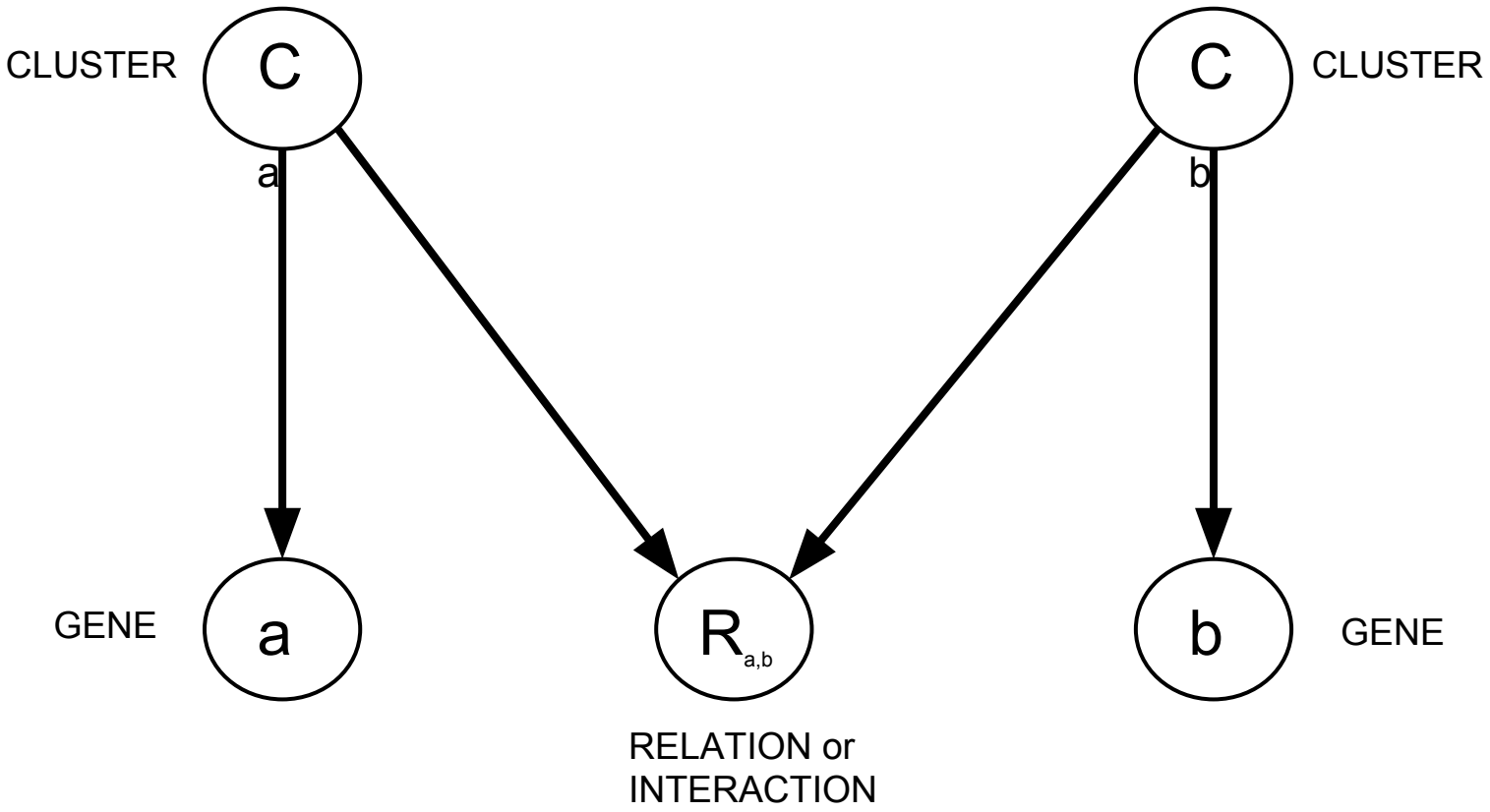
Advantages:

- More general than just links based on pairwise similarity

Disadvantages:

- Biologically unrealistic? Interaction only depends on cluster membership

Link Prediction - Latent Group Models



Link Prediction - Latent Group Models

Goal: Infer interaction probability between genes

Latent group models **(Kemp et al., 2006, Xu et al., 2006)**

Training phase:

1. **Pick a subset S of genes.**
2. **Cluster genes from S into k different groups based on their gene expression profiles.**
3. **For each pair of clusters i and j, determine the empirical interaction probability p_{ij}.**

Link Prediction - Latent Group Models

Goal: Infer interaction probability between genes based on their cluster membership (latent group)

Latent group models **(Kemp et al., 2006, Xu et al., 2006)**

Prediction phase:

1. **Given a new pair of genes a and b.**
2. **Assign a and b to the most similar Clusters C_a and C_b.**
3. **Report interaction probability $p_{a,b}$ learned in the training phase.**

Link Prediction - Latent Feature Models

Cluster-based link prediction:

Links depend on cluster membership, that is, *one latent variable*.

Latent feature-based link prediction

Links depend on a set of latent or hidden features that is, *several latent variables* (e.g. Menon and Elkan, 2011).

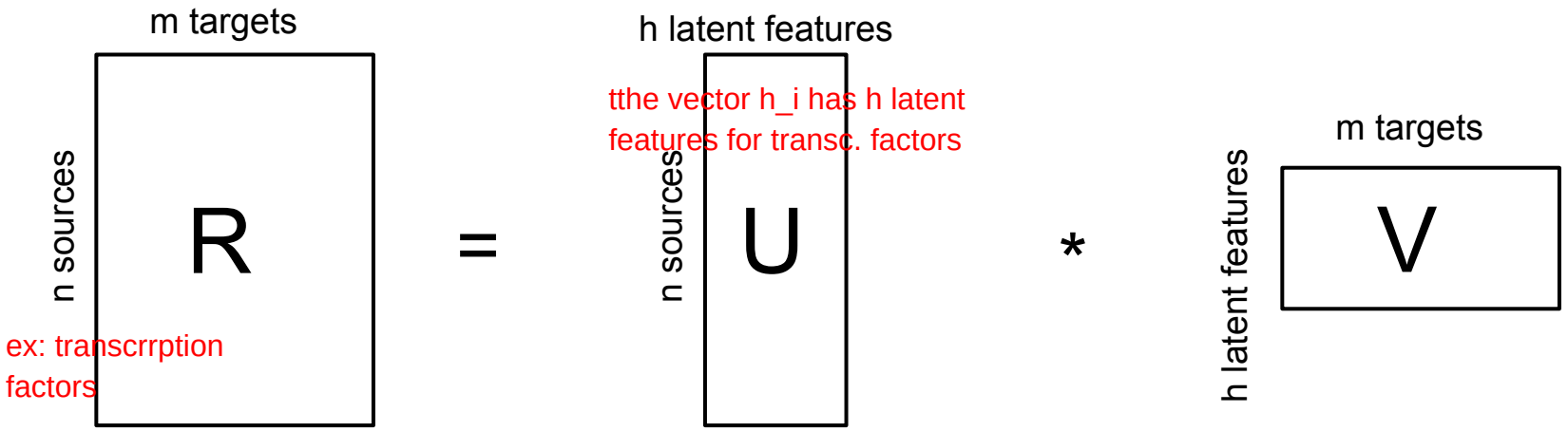
same pathway -> similar functions
around same place in cell
do they form dimers or something like that

consider groups of latent variables, not only one latent variable

some make it more likely for interaction, some make it less likely for interaction

Link Prediction - Latent Feature Models

General idea: Matrix factorization



Relation matrix or Interaction matrix R

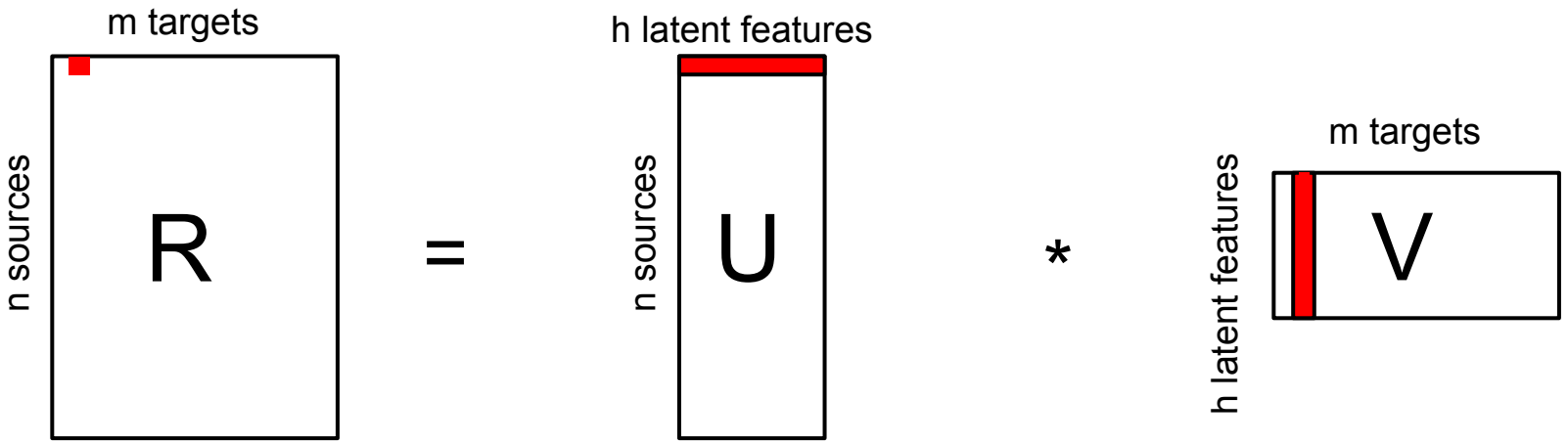
Feature representation of source nodes, U

Feature representation of target nodes, V

Link Prediction - Latent Feature Models

if R has unknown entries, one can still make U and V and approximate the unknown entries in R

General idea: Matrix factorization



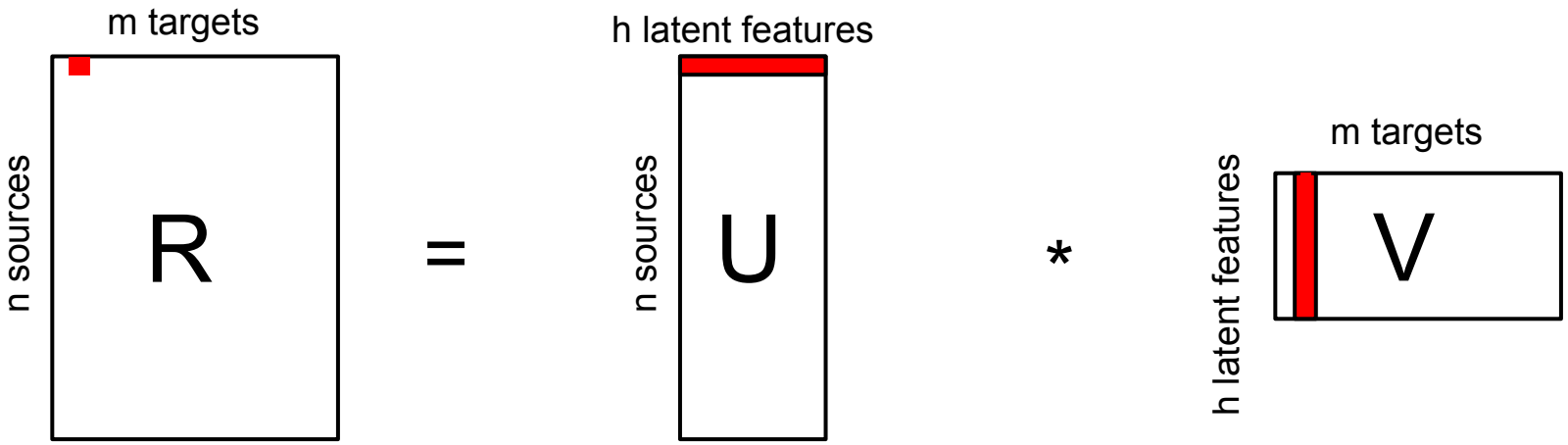
Relation matrix or Interaction matrix R

Feature representation of source nodes, U

Feature representation of target nodes, V

Link Prediction - Latent Feature Models

General idea: Matrix factorization



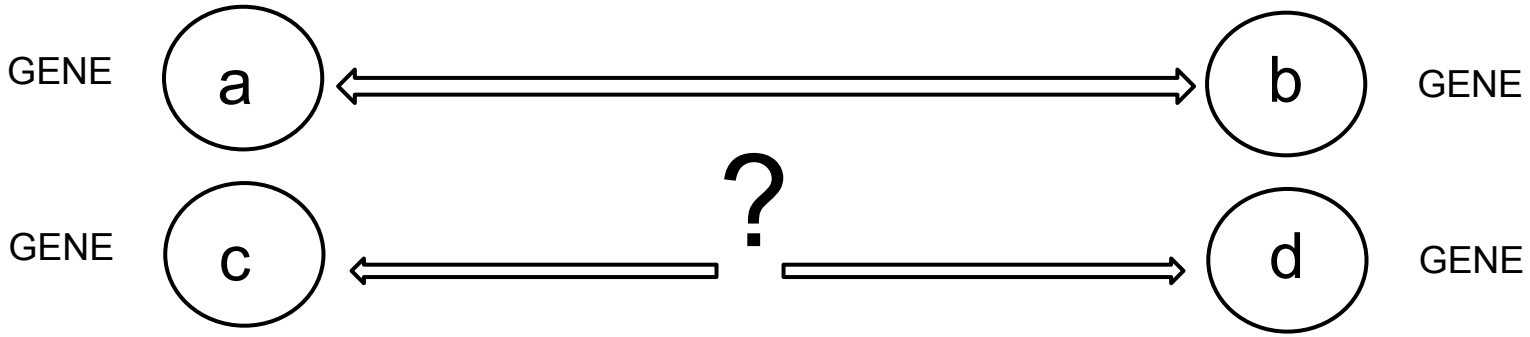
$$R_{a,b} = \sum_{i=1}^{h} U_{a,i} V_{i,b}$$

Link Prediction - Latent Feature Models

Special form of Matrix factorization:

- We do not know all interactions, that is, some entries of *R* are unknown.
- Hence we perform a matrix factorization on a matrix with missing values, to then impute these values.
- Link prediction means *Matrix Completion* here.

Link Prediction - Similarity-based classificiation (Kernel approaches)



Tensor pairwise kernel (Ben-Hur and Noble, ISMB 2005)

Given two pairs of nodes *(a, b)* and (c, d).

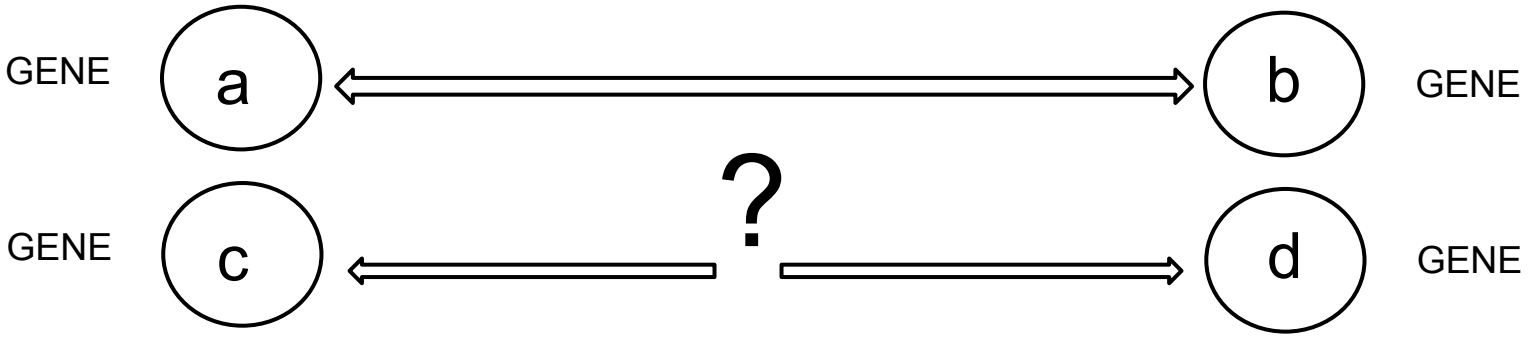
$$k_{tensor}((a, b), (c, d)) = k_{nodes}(a, c)\ k_{nodes}(b, d) + k_{nodes}(a, d)\ k_{nodes}(b, c);$$

kernel-function is big, when the k_nodes are similiar:
k_tensor = N, N big => similarity big

This kernel quantifies the similarity of the source and target nodes in both edges, for both directions.

k_{nodes} is a kernel that measures the similarity of two nodes.

Link Prediction - Similarity-based classificiation (Kernel approaches)



Metric learning pairwise kernel (Vert et al., 2007)

Given two pairs of nodes (a, b) and (c, d).

maybe genes react, when gene A has feature A and gene B has feature B and then a reaction between them occurs

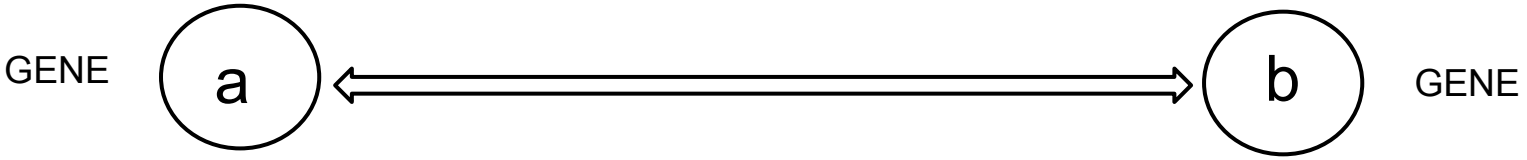
$$k_{ml}((a, b), (c, d)) = [(\varphi(a) - \varphi(b))^{T}(\varphi(c) - \varphi(d)]$$

$\varphi(g)$ is a vector that describes features of gene/protein *g*.

A pair *(a, b)* is similar to a pair *(c, d)* if $a - b$ is similar to $c - d$, or if $a - b$ is similar to $d - c$.

this function is about differences, instead of the similarity of the starting node and end node as in the previous slide

Link Prediction - Similarity-based classificiation - Selection of negative examples



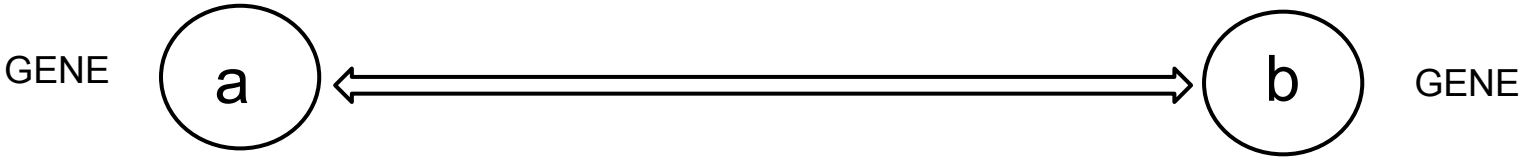
Fair selection of negative examples is tricky:

The common strategy is to pick proteins from different compartments of the cell, as measured by the **Gene Ontology Cellular Component** annotation.

This oversimplifies the problem (Ben-Hur and Noble, 2006), classifiers predict overly well if one only considers proteins from very dissimilar locations.

Dr. Noble is at the ETH
maybe it's wor th to check some literature by him

Link Prediction - Similarity-based classificiation - Selection of negative examples

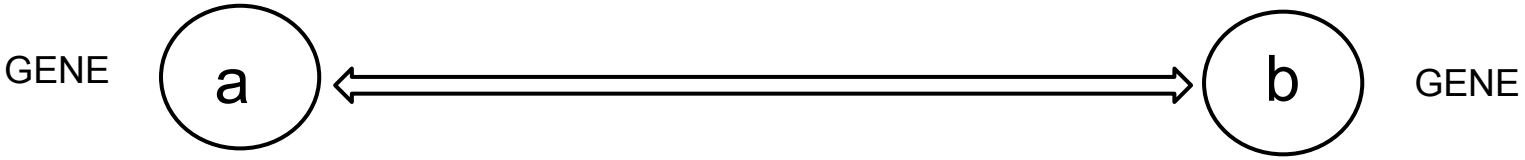


Fair selection of negative examples is tricky:

Another common strategy is to evaluate predictors on datasets which include the same number of positive examples and negative examples for interactions (balanced datasets).

This leads to too pessimistic results (Park and Marcotte, 2011).

Link Prediction - Regression based approaches



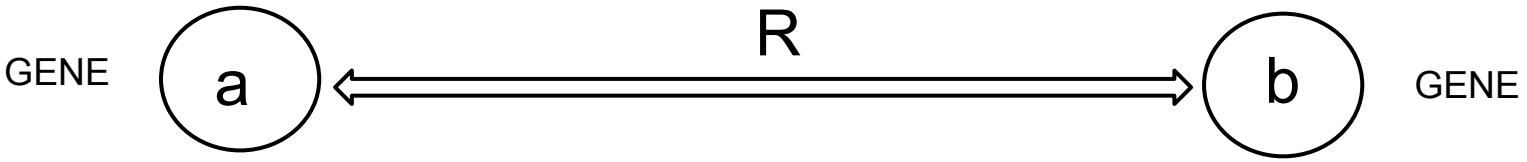
Idea: Predict the state of one gene given all others

Common method: LASSO (Linear regression with sparsity constraints), which represents our assumption that one gene should be interacting with a small set of other genes (e.g. Haury et al., 2012).

Challenge: How to set the regularization parameter that controls the sparsity of the solution?

Link Prediction - Evaluation Criteria

important table



True Label (R) vs. Predicted Label (R*)	R = 1	R = -1	
R* = 1	TP	FP	
R* = -1	FN	TN	
	P	N	

P = positive
TP = true pos.
FN = false neg.

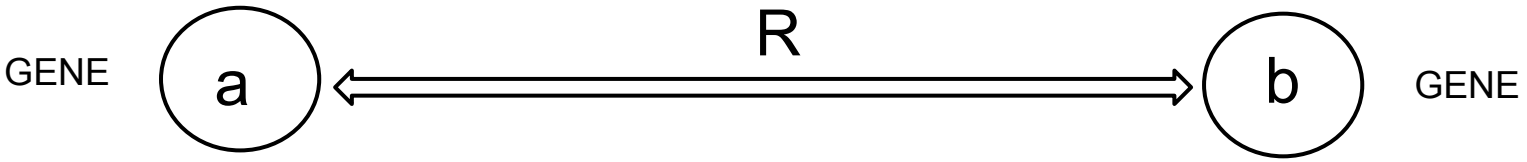
N = negative
FP = false pos.
TN = true neg.

TP+TN is true predictions

accuracy = (TP + TN) / (TP + FP + FN + TN);

Which percentage of predictions is correct?

Link Prediction - Evaluation Criteria



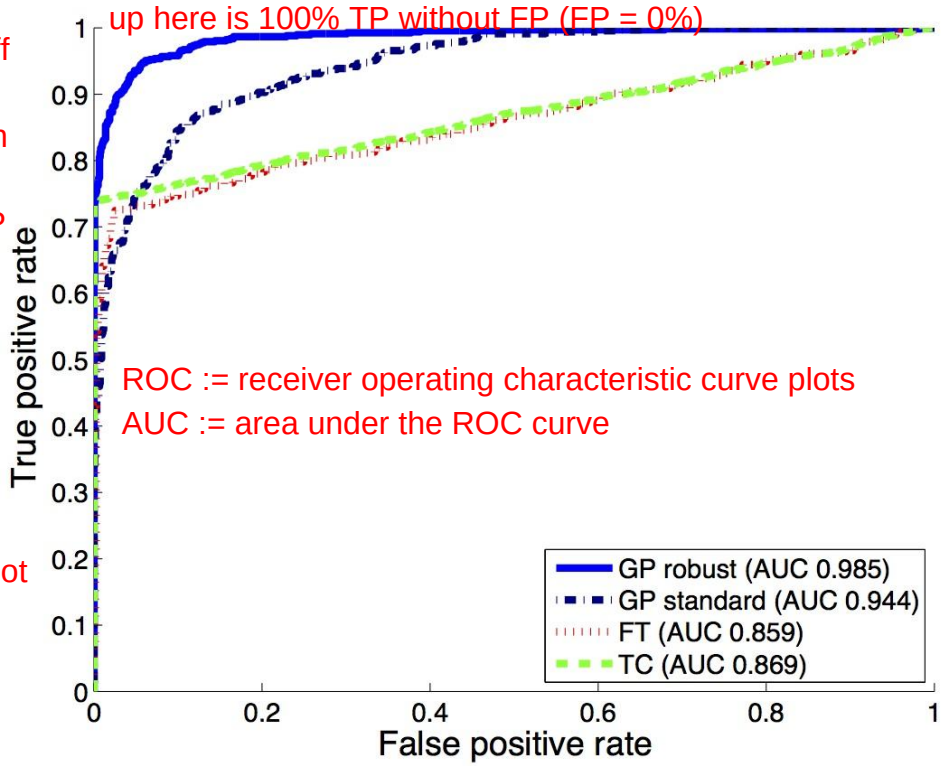
True Label (R) vs. Predicted Label (R*)	R = 1	R = -1	
R* = 1	TP	FP	
R* = -1	FN	TN	
	P	N	

- **sensitivity** (or recall or true positive rate) = TP / P;
- **specificity** = TN / N;
- false positive rate = FP / N = 1 - specificity;

Link Prediction - Evaluation Criteria - ROC

we dont know where to cut off
the rangliste of all possible
and most probable interaction
thats why we do iteratively:
first: only top interaction is TP
all other all negative
2nd: only top 2 intera. is TP,
all other neg.
...
i-th: only top-i intera is TP,
all other neg.

thats why we get a plot and not only a single dot



Receiver **O**perating **C**haracteristic **(ROC)** curve plots

False Positive Rate vs. True Positive Rate.

AUC = **A**rea **u**nder the ROC **c**urve

Link Prediction - Evaluation Criteria - Pitfalls

One interpretation of AUC:

Presented with a positive example (R_{ab}=1) and a negative example (R_{ab}=-1), AUC is the probability that the classifier will correctly discover which one is the positive example.

Pitfall:

Even with a high AUC, a classifier can be rather useless in practice, if one class is much larger than the other one.

Link Prediction - Evaluation Criteria - Pitfalls

What to do if the classes are very unbalanced?

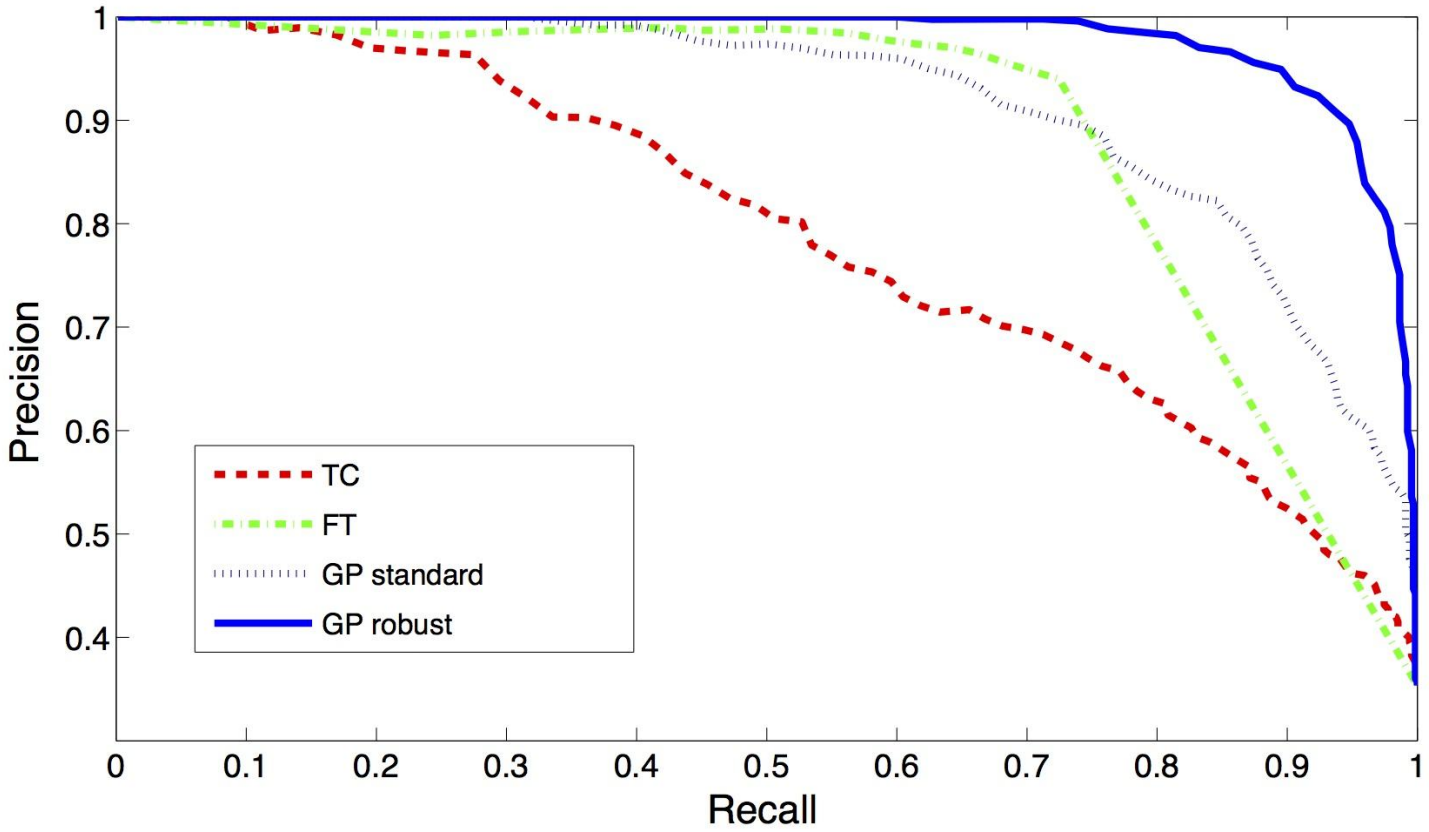
Report precision and recall

precision = TP / TP + FP;
Which percentage of the positive predictions are correct?

recall (or sensitivity or true positive rate) **= TP / P;**
Which percentage of the positive examples did the classifier find?

Link Prediction - Evaluation Criteria - Pitfalls

Precision-Recall-Curve



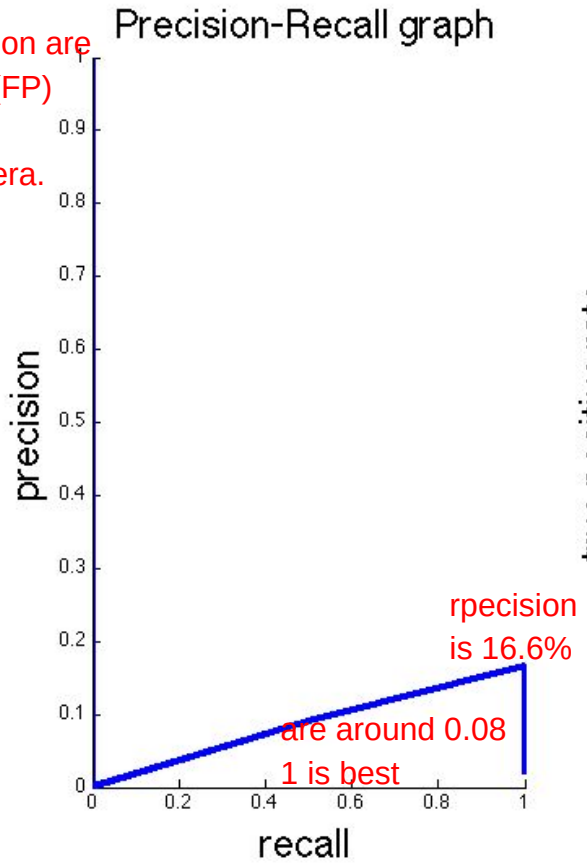
(variation of threshold)

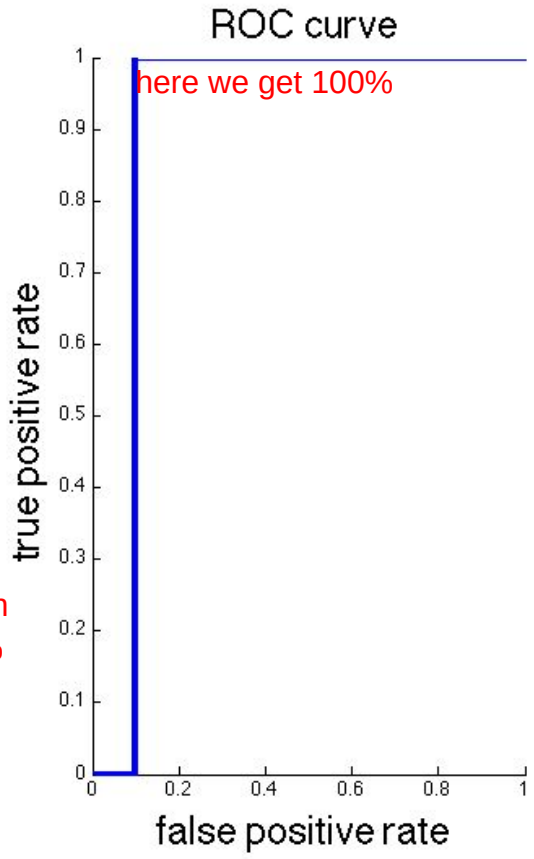
Link Prediction - Evaluation Criteria - Pitfalls

Example

We perform link prediction for 102 pairs of nodes. For 2 pairs, a link actually exists. They are ranked in the 11th and 12th position of the solution.

the ten first interaction are wrong interactions (FP)
11 and 12 are TP
the rest are non-intera.

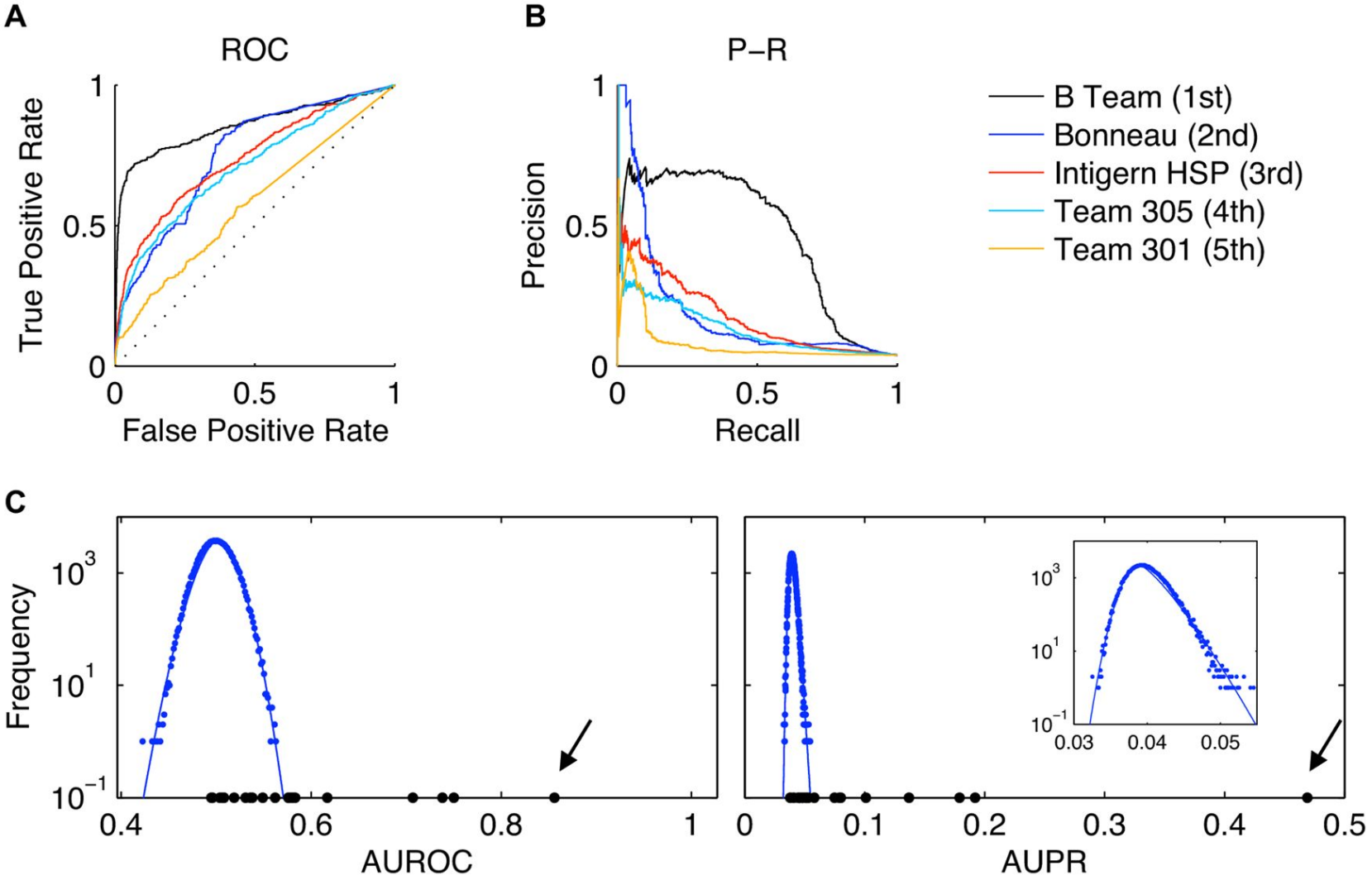




AUC/ROC is misleading here and looks much better than the precision-recall curve.

Link Prediction - DREAM Challenges

DREAM 3 - In Silico Network Challenge (Prill et al., 2010)



Link Prediction: Summary

- Biological networks are inherently incomplete. Link prediction tries to complete these networks by computational predictions.
- A plethora of methods exists. It is important to understand their underlying assumptions, as the links they predict reflect these assumptions.
- Several strategies of choosing negative examples for supervised link prediction may introduce biased results.
- The evaluation of link prediction is complicated by the fact that the positive class (interactions) and negative class (non-interactions) are highly unbalanced. Reporting precision and recall, in addition to AUC, is recommended.

References

- Ben-Hur, A., and Noble, W.S. (2005). Kernel methods for predicting protein-protein interactions. Bioinformatics *21 Suppl 1*, i38–i46.
- Ben-Hur, A., and Noble, W.S. (2006). Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics *7*, S2.
- Haury, Anne-Claire, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. “TIGRESS: Trustful Inference of Gene REgulation Using Stability Selection.” *BMC Systems Biology* 6 (2012): 145.
- Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In AAAI 2006, 381-388
- Marbach, D., Costello, J.C., Küffner, R., Vega, N., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat Methods *9*, 796–804.
- Menon, Aditya Krishna, and Charles Elkan. “Link Prediction via Matrix Factorization.” In *Machine Learning and Knowledge Discovery in Databases*, edited by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, 437–52. Lecture Notes in Computer Science 6912. Springer Berlin Heidelberg, 2011.
- Park, Y., and Marcotte, E.M. (2011). Revisiting the negative example sampling problem for predicting protein-protein interactions. Bioinformatics *27*, 3024–3028.
- Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. (2010) Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. PLoS ONE 5(2): e9202.
- Titz, B., Rajagopala, S.V., Goll, J., Häuser, R., McKevitt, M.T., Palzkill, T., and Uetz, P. (2008). The Binary Protein Interactome of Treponema pallidum – The Syphilis Spirochete. PLoS ONE *3*(5):e2292
- Vert, J.-P., Qiu, J., and Noble, W.S. (2007). A new pairwise kernel for biological network inference with support vector machines. BMC Bioinformatics *8*, S8.