Instructor: Sael Lee

CS549 Spring – Computational Biology

Random Walk and Graph Kernel Applications to Bioinformatics

RWR Application: Disease Gene Association

Disease Gene Association

- Disease Gene Association is identification of causal genes of a disease.
- Useful for:
  - Preventing and curing the disease.
  - Understanding the biological functions of genes
- Traditional method popular in the early 2000
  - Genome-wide association studies (GWAS)
  - Relies on testing several hundred thousand common genetic variants found throughout the human genome in large-control cohorts (patients with same disease/phenotype).
  - Problem: Due to lack of the ability to detect ‘common disease by rare variants’ explains only portion of genetic risk

Random Walk Based Methods

- ❑ Random walk based methods are one popular alternative approach for associating genes with disease.
- ❑ General idea:
  - ❑ '<u>guilt by association</u>' principle (Wolfe et al., 2005) with respect to a set of known genes related to the given disease.

Kohler,S. et al. (2008)

Kohler et al.’s Approach

- Gene–disease associations by using a global network distance measure for the definition of similarities in protein–protein interaction
  - a random walk analysis

- Data sets
  - Disease-Gene Family Information
  - Protein-Protein Interaction Network
- Disease-Gene Prediction Methods
  - Random Walk
  - Diffusion Kernel
  - Other methods

Kohler,S. et al. (2008) Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet., 82, 949–958.

Kohler,S. et al. (2008)

Disease-Gene Family Information

- A total of 110 disease-gene families defined as follows:
  - Online Mendelian Inheritance inMan (OMIM) database
    - Extract genetically heterogeneous disorders - selecting mutations in distinct genes associated with similar or even indistinguishable phenotypes
    - Cancer syndromes comprising genes associated with hereditary cancer, increased risk, or somatic mutation in a given cancer type;
    - Complex (polygenic) disorders that are known to be influenced by variation in multiple genes.
  - Domain knowledge and literature or database searches
    - Select all genes clearly associated with the disorder at hand
- Summary of extracted 110 disease-gene families
  - Contains 783 genes with 665 distinct genes (Some genes were members of more than one disease family),
  - Largest family contained 41 genes and the smallest only three genes.
  - On average, each family contained **seven genes**.

Kohler,S. et al. (2008)

Protein-Protein Interaction Data

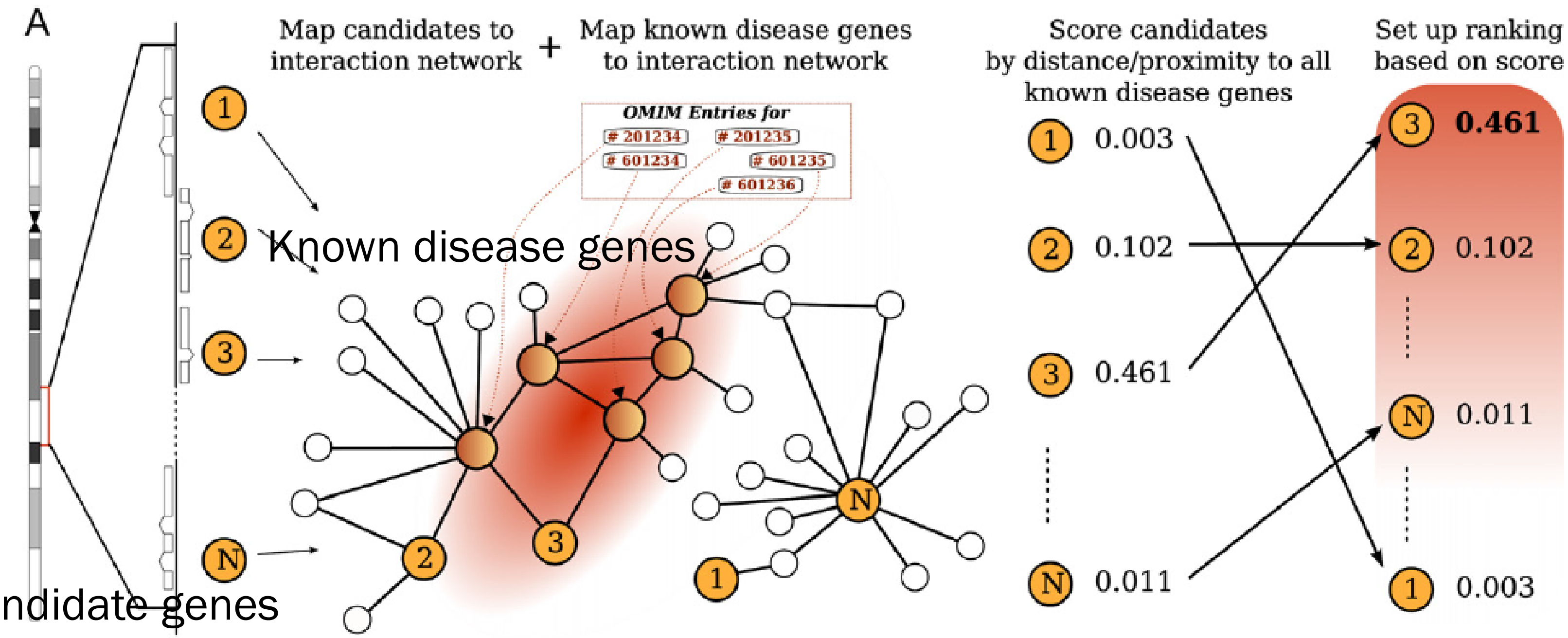
- PPI graph structure
  - Undirected graph: nodes representing the genes and edges representing the mapped interactions of the proteins encoded by the genes.
- PPI construction
  - Entrez Gene & IntACT and DIP
    - Five networks from species comprises interaction from HPRD, BIND, and BioGrid.
    - human, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, and Saccharomyces cerevisiae
    - Protein interactions mapped to the genes coding for the proteins, and redundant interactions removed
  - Mapping nonhuman interactions to human
    - map to homologous human genes identified by Inparanoid analysis with a threshold Inparalog score of 0.8.
    - If both interaction partners could be mapped to human proteins, the interaction was used.
  - STRING
    - STRING: contains functional links between proteins on the basis of both experimental evidence for protein-protein interactions as well as interactions predicted by comparative genomics and text mining.
    - STRING uses a scoring system that is intended to reflect the evidence of predicted interactions.
    - Included interactions with a score of at least 0.4, (medium-confidence)

Kohler,S. et al. (2008)

Disease-Gene Prediction

assigns a score to each of the candidate genes,

Figure 1. Disease-Gene Prioritization


A
Map candidates to interaction network
+
Map known disease genes to interaction network
OMIM Entries for
# 201234
# 201235
# 601234
# 601235
# 601236
Score candidates by distance/proximity to all known disease genes
Set up ranking based on score
1
2
3
N
1 0.003
2 0.102
3 0.461
N 0.011
3 0.461
2 0.102
N 0.011
1 0.003

Known disease genes

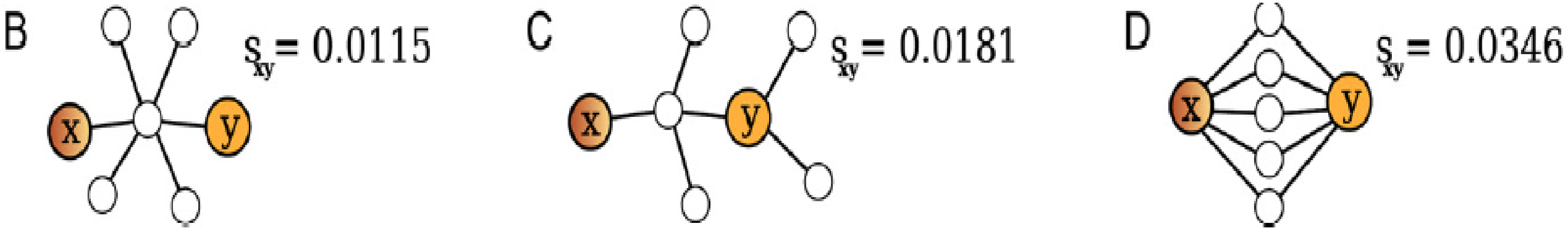
Candidate genes

1. Extract candidate genes contained in the linkage interval
2. Map candidate genes previously known disease genes and to the interaction network
3. Assigns a score to each of the candidate genes
    1. Base on relative location of the candidate to all of the known ‘‘disease genes’’ by the use of global network-distance measures.
4. Rank genes in the linkage interval

Kohler,S. et al. (2008)

Disease-Gene Prediction Steps

Example: Different net configuration consisting of the same number of nodes



- The global distance between a hypothetical disease gene (x) and a candidate gene (y) is different in each case.
- In (B), proteins x and y are connected via a hub node with many other connections, so that the global similarity is less than in (C),
- In (C) x and y are connected by a protein with fewer connections than those of the hub.
- Nodes that are connected by multiple paths (D) receive a higher similarity than do nodes connected by only one path.
- NOTE: shortest path between x and y is identical in each case (B–D),
    - distance measures relying on shortest path cannot differentiate between these three types of connection.

Random Walk with Restart

$$\mathbf{p}^{t+1} = (1 - r)\mathbf{W}\mathbf{p}^{t} + r\mathbf{p}^{0}$$

Where **W** is the column-normalized adjacency matrix of the graph and p_t is a vector in which the i-th element holds the probability of being at node i at time step t and restart probability of r

Initialization: $p_0 = 1/|G_d|$ where $|G_d|$ number of known genes assigned to disease This is equivalent to letting the random walker begin from each of the known disease genes with equal probability.

Iteration: Candidate genes were ranked according to the values in the steady-state probability vector p^{inf}. This was obtained at query time by performing the iteration until the change between p^{t+1} and p^{t+1} (measured by the L1 norm) fell below 10^{-6}.

Diffusion Kernel

- The diffusion kernel K of a graph G is defined as

$$\mathbf{K} = e^{-\beta \boldsymbol{L}}$$

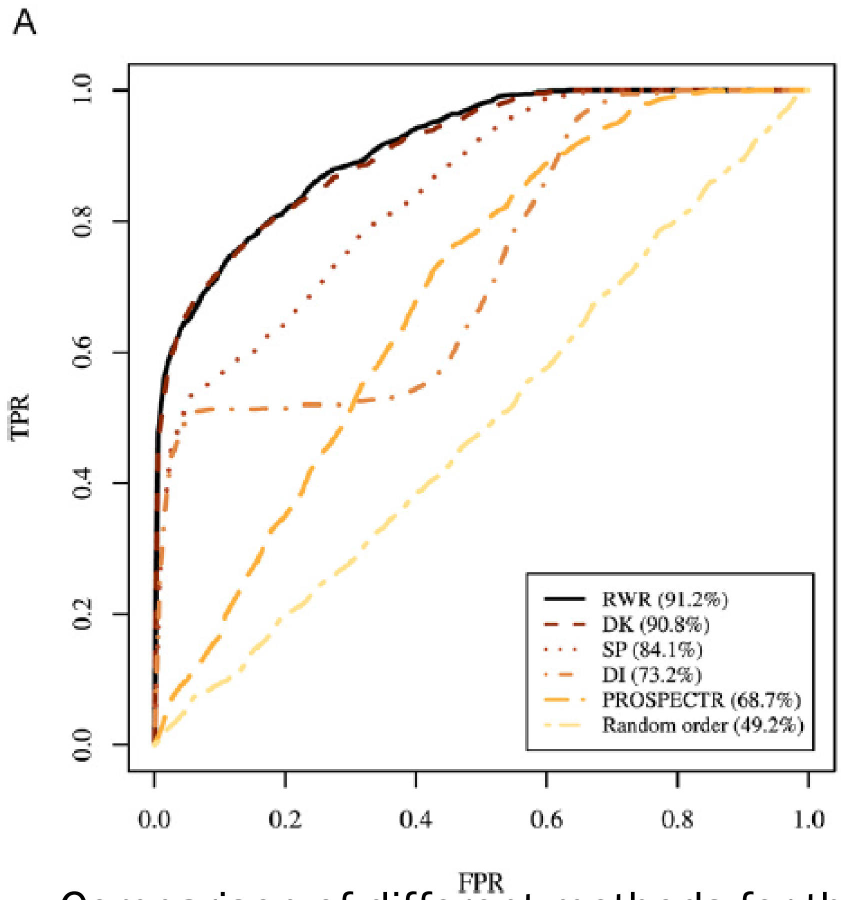
where, β controls the magnitude of the diffusion.

The matrix L is the Laplacian of the graph, defined as D - A, where A is the adjacency matrix of the interaction graph and

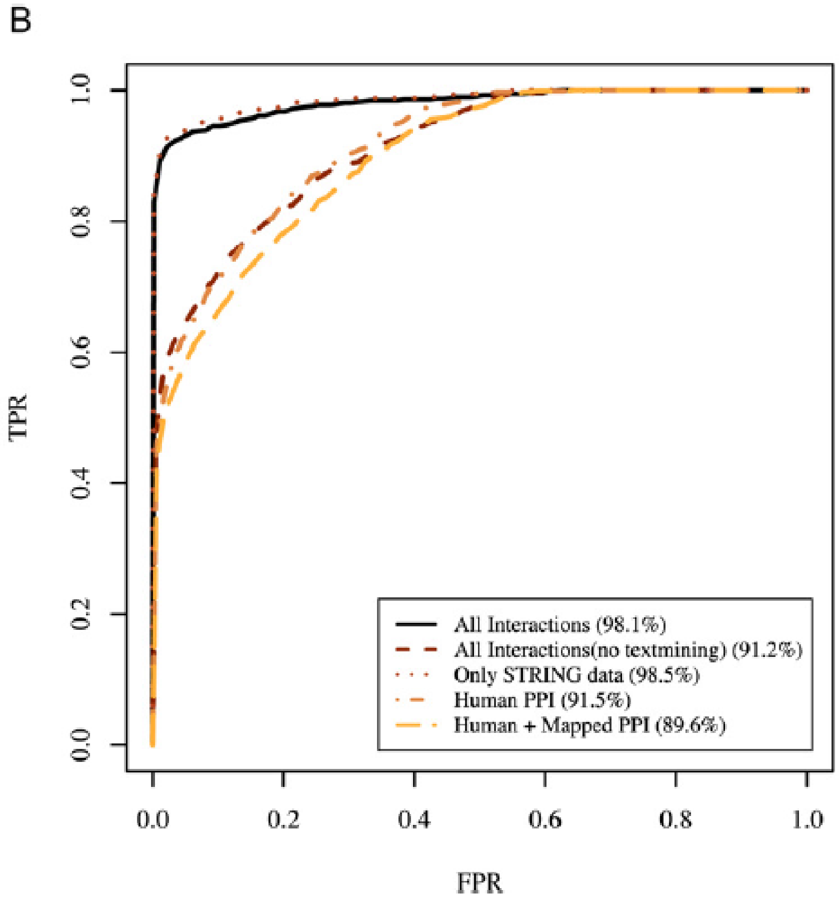
D is a diagonal matrix containing the nodes’ degrees.

- For small β,
  - The column vector j of the matrix K represents the steady-state probability vector of the random walk when starting at node j.
- Diffusion Score
  - for each candidate gene j was assigned in accordance with its s core defined as score(j) $= \sum_{i \; in \; disease \; gene \; family} \boldsymbol{K}_{ij}$

Results



Comparison of different methods for the all-interactions network without STRING text-mining data

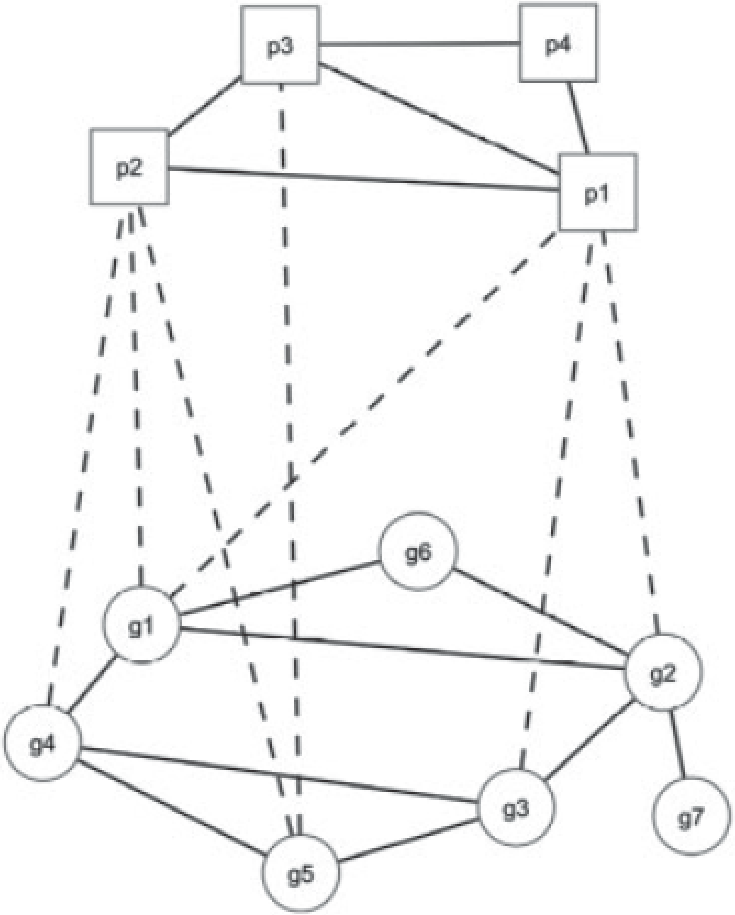


Comparison of different data sources with RWR analysis

(Li and Patra, 2010)

Li,Y. and Patra,J.'s Heterogeneous network

- An extension of the random walk approach on a heterogeneous network that includes protein–protein interaction, disease–disease and gene–disease networks.



Li,Y. and Patra,J. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 26, 1219–1224.

The upper subnetwork is phenotype network, and the lower network is gene network.

(Li and Patra, 2010)

Data source

- The **protein–protein interaction** (PPI)
  - Human Protein Reference Database (HPRD).
  - HPRD contains manually curated scientific information pertaining to the biology of most of the human proteins.
- **Disease-related phenotype**
  - Interpreted as a textual description of a disease’s detectable outward manifestations. (van Driel *et al*., 2006; Wu *et al*., 2008),
  - Phenotype entry was defined as an MIM record.
  - Excluded the records with the prefix ‘∗’ and ‘∧’. Because the prefix ‘∗’ refers to the record of disease gene, and ‘∧’ refers to the obsoleted record.
- The **phenotypic similarity**
  - Calculated using MimMiner (van Driel *et al*., 2006).
- **Gene–phenotype relationship**
  - OMIM database (Hamosh *et al*., 2005), extracted using BioMart (Smedley *et al*., 2009).
- **Disease category** information
  - Manual classification concerning the physiological system affected (Goh *et al*., 2007).

(Li and Patra, 2010)

Construction of the heterogeneous network

- ❑ Gene network:
  - ❑ two genes are connected if the proteins they encode interact with each other according to the HPRD database.
- ❑ Phenotype network:
  - ❑ Each phenotype entity is connected with its five nearest neighbors, and the edge is weighted by the corresponding similarity score using MimMiner.
- ❑ Gene–Phenotype network: bipartite graph
  - ❑ phenotype entity with the relevant genes

- ❑ Heterogeneous network
  - ❑ $A_{G(n\times n)}$ adjacency matrix for gene network
  - ❑ $A_{P(m\times m)}$adjacency matrix for ork, phenotype network
  - ❑ $B_{(n\times m)}$adjacency matrix for bipartite graph,.

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix},$$

RWRH

- Let $\mathbf{p}_o$ be the initial probability vector and $\mathbf{p}_s$ be a vector in which the *i*-th element holds the probability of finding the random walker at node *i* at step *s*.
- The probability vector at step *s*+1 can be given by

$$\mathbf{p}_{s+1} = (1-\gamma)M^T\mathbf{p}_s + \gamma\mathbf{p}_0,$$

where *M* is the transition matrix of the graph. M_{ij} is the transition probability from node *i* to node *j*. The parameter γ ∈(0,1) is the restart probability. At each step, the random walker can return to seed nodes with probability γ.

- Initialization

$$\mathbf{p}_0 = \begin{bmatrix} (1-\eta)\mathbf{u}_0 \\ \eta\mathbf{v}_0 \end{bmatrix}$$

The parameter η∈(0,1) is used to weight the importance of each subnetwork
u0 and **v**0 represent the initial probability of gene network and phenotype network

Transition matrix

❑ Transition matrix of the heterogeneous network

$$M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_P \end{bmatrix}$$

Let *λ* be the jumping probability: probability of the random walker jumping from gene network to phenotype network or vise versa

$$(M_{GP})_{i,j} = p(p_j|g_i) = \begin{cases} \lambda B_{ij}/\sum_j B_{ij}, & \text{if } \sum_j B_{ij} \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

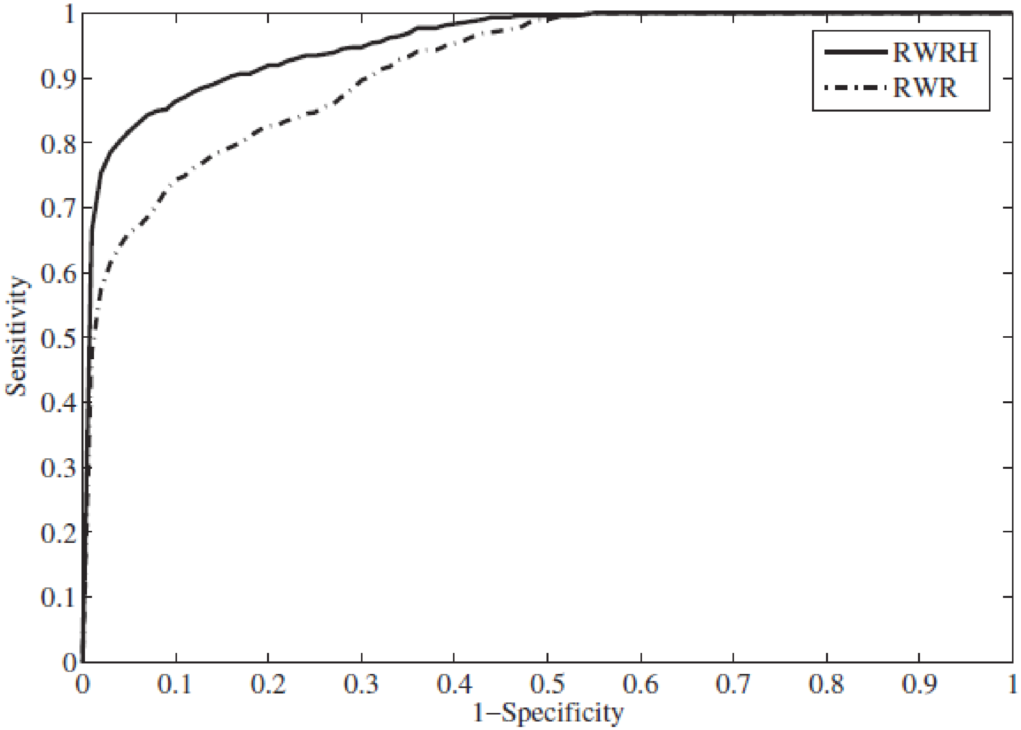
$$(M_{PG})_{i,j} = p(g_j|p_i) = \begin{cases} \lambda B_{ji}/\sum_j B_{ji}, & \text{if } \sum_j B_{ji} \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$(M_G)_{i,j} = \begin{cases} (A_G)_{i,j}/\sum_j (A_G)_{i,j}, & \text{if } \sum_j B_{ij} = 0 \\ (1-\lambda)(A_G)_{i,j}/\sum_j (A_G)_{i,j}, & \text{otherwise.} \end{cases}$$

$$(M_P)_{i,j} = \begin{cases} (A_P)_{i,j}/\sum_j (A_P)_{i,j}, & \text{if } \sum_j B_{ji} = 0 \\ (1-\lambda)(A_P)_{i,j}/\sum_j (A_P)_{i,j}, & \text{otherwise.} \end{cases}$$

Results

- ❑ Disclose hidden disease–disease associations
- ❑ Accuracy comparison

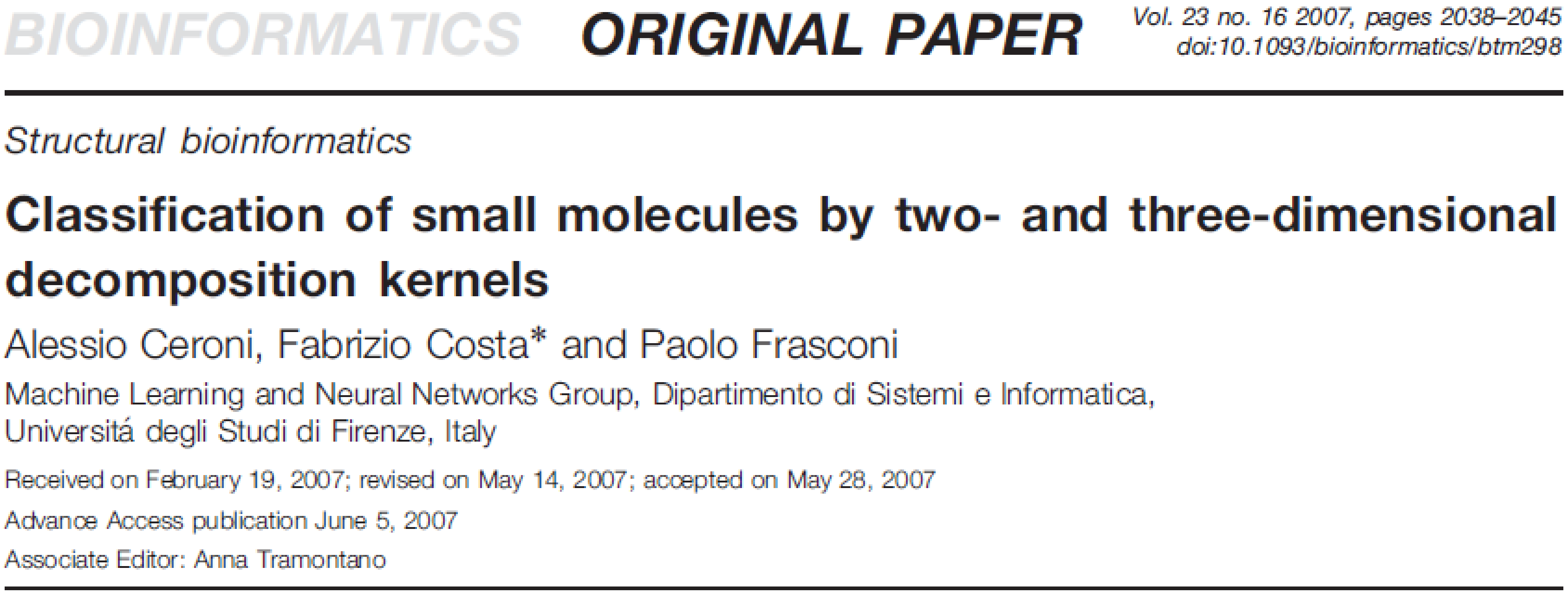


RWR on gene network

Fig. 2. ROC curve of RWR and RWRH.

Classification of small molecules by two- and three-dimensional decomposition kernels

Ceroni, A., Costa, F., & Frasconi, P. (2007). *Bioinformatics*, *23*(16), 2038–45.



BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 16 2007, pages 2038–2045
doi:10.1093/bioinformatics/btm298

Structural bioinformatics

Classification of small molecules by two- and three-dimensional decomposition kernels

Alessio Ceroni, Fabrizio Costa* and Paolo Frasconi

Machine Learning and Neural Networks Group, Dipartimento di Sistemi e Informatica,
Universitá degli Studi di Firenze, Italy

Received on February 19, 2007; revised on May 14, 2007; accepted on May 28, 2007

Advance Access publication June 5, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Several kernel-based methods have been recently introduced for the classification of small molecules. Most available kernels on molecules are based on 2D representations obtained from chemical structures, but far less work has focused so far on the definition of effective kernels that can also exploit 3D information.

Results: We introduce new ideas for building kernels on small molecules that can effectively use and combine 2D and 3D information. We tested these kernels in conjunction with support vector machines for binary classification on the 60 NCI cancer screening datasets as well as on the NCI HIV data set. Our results show that 3D information leveraged by these kernels can consistently improve prediction accuracy in all datasets.

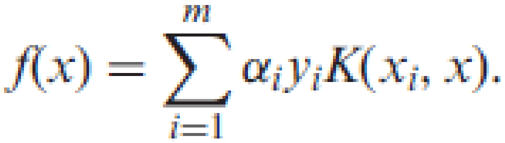
Availability: An implementation of the small molecule classifier is available from http://www.dsi.unifi.it/neural/src/3DDK

Methods:
Background on kernel methods for structured data

A major challenge is to **define an effective quantitative measure of similarity**

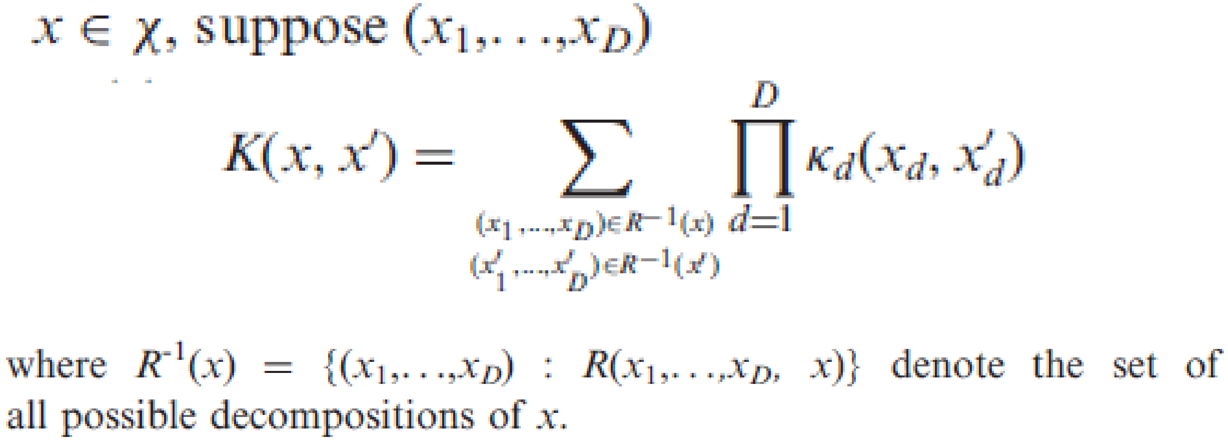
Base Learner: **SVM**

classification function f(x) is obtained from data



$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x).$$

Kernels on structured data



$x \in \chi$, suppose $(x_1,\ldots,x_D)$

$$K(x, x') = \sum_{\substack{(x_1,\ldots,x_D)\in R^{-1}(x) \\ (x'_1,\ldots,x'_D)\in R^{-1}(x')}} \prod_{d=1}^{D} \kappa_d(x_d, x'_d)$$

where $R^{-1}(x) = \{(x_1,\ldots,x_D) : R(x_1,\ldots,x_D, x)\}$ denote the set of all possible decompositions of x.

A weighted decomposition kernel (WDK) for 2D chemical structures

Idea: each substructure in which a graph is decomposed is **enriched with its graphical context** characterized by a **decomposition** R(s,z,x) where s is a subgraph of x called the **selector** and z is a subgraph of x called the **context** of occurrence of s in x (generally a subgraph containing s).

This setting results in the following general form of the kernel:

$$K_{2D}(x, x') = \sum_{\substack{(s,z)\in R^{-1}(x)\\(s',z')\in R^{-1}(x')}} \delta(s, s')\kappa(z, z')$$

where, δ is the **exact matching kernel** applied to selectors and
κ is a kernel on contexts.

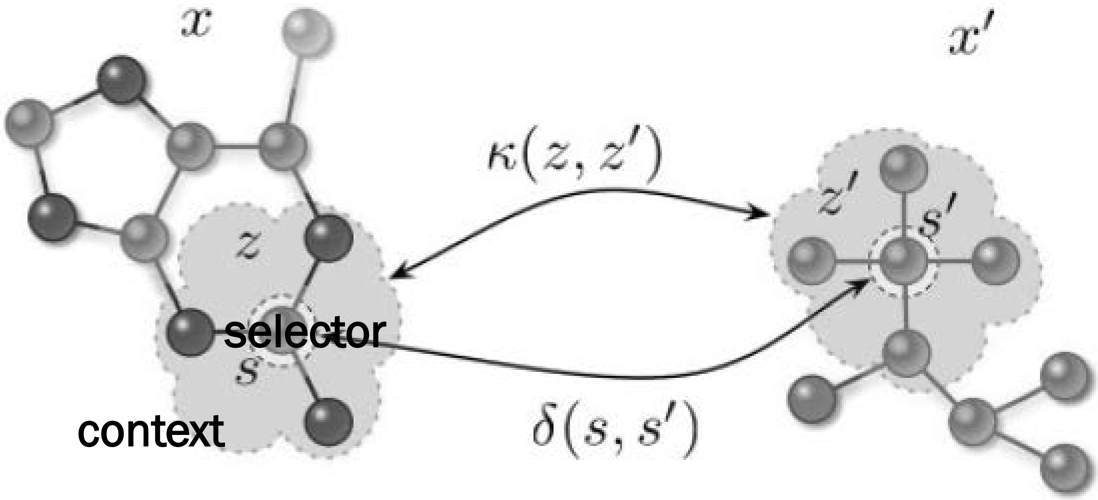


Fig. 1. Comparing substructures in a weighed decomposition kernel.

$$K_{2D}(x, x') = \sum_{\substack{(s,z)\in R^{-1}(x)\\(s',z')\in R^{-1}(x')}} \delta(s, s')\kappa(z, z')$$

WDK parameter used in the article

IDEA

- **Selectors** are always single atoms and the match
  - $\delta(s, s')$ is defined by the coincidence between the type of s and s'.
- The **context kernel** κ is based on soft match between substructures, defined by the distributions of label contents after discarding topology.

Context Kernel Specifics

1. Let L denote the total number of attributes labeling vertices and edges and for $l = 1, \ldots, L$
2. Let $p_l(j)$ be the observed frequency of value j for the $l - th$ attribute in a substructure z.
3. Then compare substructures by means of a **histogram intersection kernel**

$$\kappa_\ell(z, z') = \sum_{j=1}^{m_\ell} \min\{p_\ell(j), p'_\ell(j)\}$$

$$\kappa(z, z') = \sum_{\ell=1}^{L} \kappa_\ell(z, z').$$

Where m_l is the number of possible values of attribute l.
shall use $L =$ 3: 1) **atom type, 2) atom charge** and **3) bond type**, while atom coordinates are discarded for computing the WDK.

Contexts are formed as follows

Given a vertex $v \in V$ and an integer $r \geq 0$,

- Let $x(v, r)$: substructure of x obtained by retaining all the vertices that are reachable from v by a path of length at most r, and all the edges that touch at least one of these vertices.

The **decomposition relation** R_r, dependent on the context radius r, is then defined as

$$R_r = \{(s, z, x) : x \in \chi, s=\{v\}, z=x(v, r), v \in V\}.$$

where s is the selector and z is the context for vertex v.

Weighted Decomposition Kernel (WDK)

$$K_{2D}(x, x') = \sum_{\substack{(s,z)\in R^{-1}(x) \\ (s',z')\in R^{-1}(x')}} \delta(s, s')\kappa(z, z')$$

Three-dimensional decomposition kernels

A 3D molecular structure is interpreted as a special kind of relational data object where atoms are related by chemical bonds but also by their spatial distances

The molecule is first decomposed into a set of overlapping 3D substructures of varied geometry, called **shapes**.

Given a molecule $x = (V, E)$, a **shape** of **order n** is a set of n distinct vertices

$$\sigma = \{u_1, u_2, \ldots, u_i\}, \;\; u_i \in V, \text{for } i = 1, \ldots n.$$

kernel between two molecules

$$K_{3D}(x, x') = \sum_{\sigma \in S_r(x)} \sum_{\sigma' \in S_r(x')} k_{\text{shapes}}(\sigma, \sigma')$$

kernels between all pairs of shapes

$$k_{\text{shapes}}(\sigma, \sigma') = \prod_{i=1}^{n(n-1)/2} \delta(e_i, e'_i) e^{-\gamma(d_i - d'_i)^2}$$

Kernels between all pairs of shapes

Given a **shape of order n** and two vertices $u, v \in \sigma$,

- let $e = (t[u], t[v], b[u, v])$ denote a **labeled edge of the shape**, formed by considering the two atom types t[u] and t[v] and the bond type b[u,v].

Then, let $< e_1, \ldots, e_{n(n-1)/2} >$ denote the **lexicographically ordered sequence of all labeled edges** in .

For example, the shape (C1,C2,C3,O1) for the molecule NSC_1027 yields **lexicographically ordered sequence of all labeled edges** (C.2,C.3,1) (C.2,C.3,1) (C.2,O.2,2) (C.3,C.3,0) (C.3,O.2,0) (C.3,O.2,0).

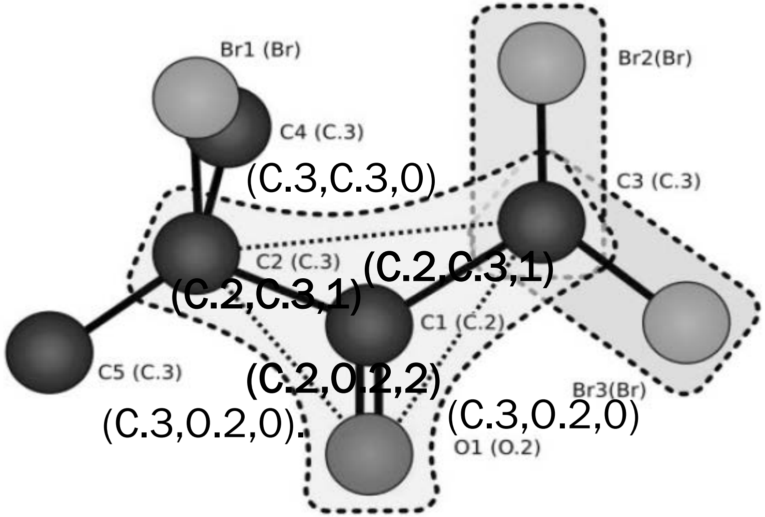


Fig. 2. Illustration of the definition of 2D-supported shapes. The three 2D-supported shapes of radius 1, anchored to atom C3 in the molecule NSC_1027 are (Br3,C3), (Br2,C3) and (C1,C2,C3,O1). Atom identifiers and types (in parentheses) are formatted according to the Tripos Sybyl MOL2 conventions.

Kernels between a pair of shapes

The kernel between two shapes $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ of equal order n is defined as:

$$k_{\text{shapes}}(\sigma, \sigma') = \prod_{i=1}^{n(n-1)/2} \delta(e_i, e'_i) e^{-\gamma(d_i - d'_i)^2}$$

$$\begin{array}{rl|c|l}
\cline{3-3}
\sigma: & < & e_1 & , \ldots, e_{n(n-1)/2} > \\
\sigma': & < & e'_1 & , \ldots, e'_{n(n-1)/2} > \\
\cline{3-3}
\end{array}$$

Where γ is a kernel hyperparameter and $d_i = ||\vec{\zeta}[u_i] - \vec{\zeta}[v_i]||$ is the length of edge e_i, i.e. the Euclidean distance between atoms ui and vi.

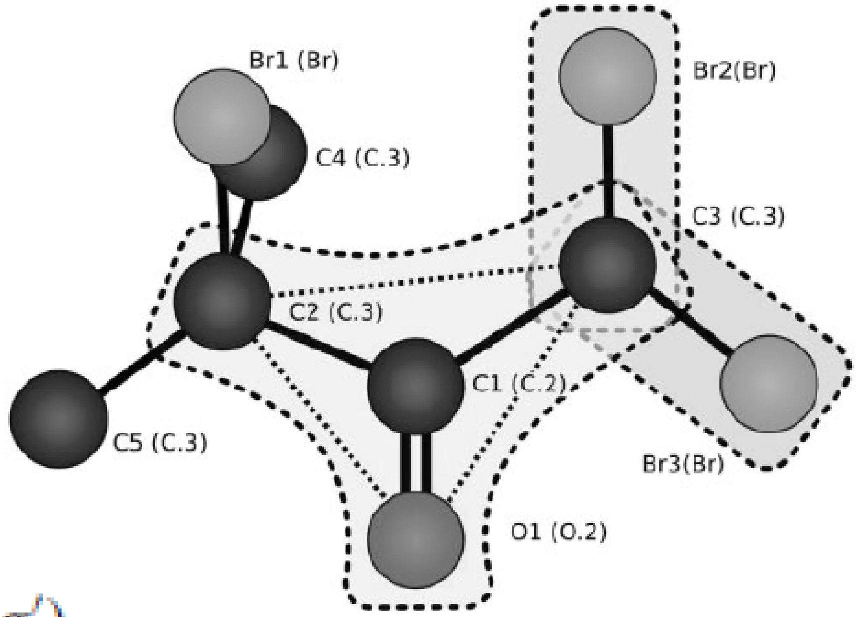
* The kernel between two shapes of different order is null.

kernels between all pairs of shapes cont.

3D decomposition kernels (3DDK)

Select just the adjacent list of vertices that are **within distance r** from x.

Given a vertex $v \in V$ and an integer r, a **2D-supported shape** anchored in v is a set of vertices $\sigma = \{v, w\} \cup$ adj$[w]$ such that $w \in x(v, r)$ and adj[w] is the adjacency list of w. Let $S_r(x)$ denote **shape set of radius r of x.**



$$K_{3D}(x, x') = \sum_{\sigma \in S_r(x)} \sum_{\sigma' \in S_r(x')} k_{\text{shapes}}(\sigma, \sigma')$$

Data Set: &

NCI cancer dataset

National Cancer Institute public dataset of screening results for the ability of more than 70,000 compounds to suppress or inhibit the growth of a panel of 60 human tumor cell lines.

Subset of NCI dataset corresponding to the parameter GI50, the concentration that causes 50% growth inhibition is used.

Binary classification: cancer-inhibiting (1) or not (-1).

NCI HIV dataset

Contains 42,687 compounds evaluated for evidence of anti HIV activity from the DTP AIDS antiviral screen program of the National Cancer Institute.

Compounds are divided in **three classes: 1)** 422 compounds are confirmed active (CA), **2)** 1081 are moderately active (CM) and **3)** 41 184 are confirmed inactive (CI).

Three Class classification with SVM

Three classification problems are formulated on this dataset:

1. (CA verses CM): positive examples are confirmed active compounds, while moderately active compounds forms the negative class;
2. (CA+CM verses CI): the positive class is formed by the combination of moderately active and confirmed active compounds and in
3. (CA verses CI): the positive examples are confirmed active compounds and the negative class is formed by confirmed inactive compounds.

Combining Kernels– WDK & 3DDK

NCI cancer dataset

The WDK and 3DDK used in this experiment had both the radius r = 3 and no graph complement was used for the WDK. γ parameter in pair-wise shape kernel was set to 2.5

$$K(x, x') = (1 + \kappa(x, x'))^2$$

$$\kappa \text{ is either } K_{2D} \text{ or } K_{3D} \text{ or } \kappa(x,x') = K_{2D}(x,x') + K_{3D}(x,x').$$

These measures were estimated by a 10-folds cross-validation

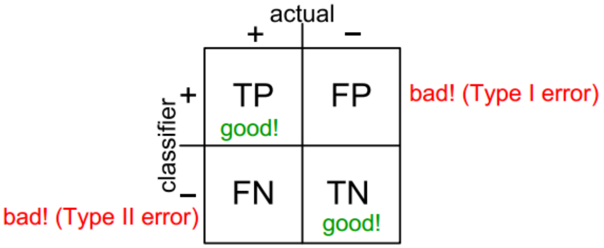
NCI HIV dataset

For the WDK, graph complement and context radius r = 4 is used.
For the 3DDK, the radius to r = 3 is used. γ parameter in pair-wise shape kernel was set to 2.5.

$$K(x, x') = e^{-\frac{1}{2}(\kappa(x,x)+\kappa(x',x')-2\kappa(x,x'))}$$

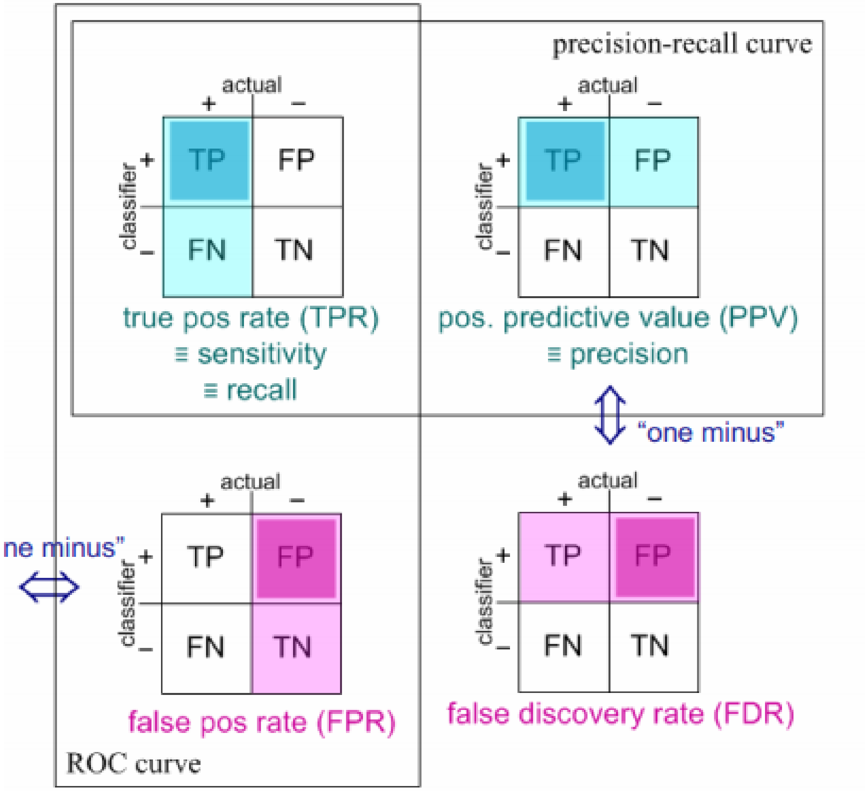
AUC performance was estimated by a 5-folds cross-validation

ROC vs Precision Recall



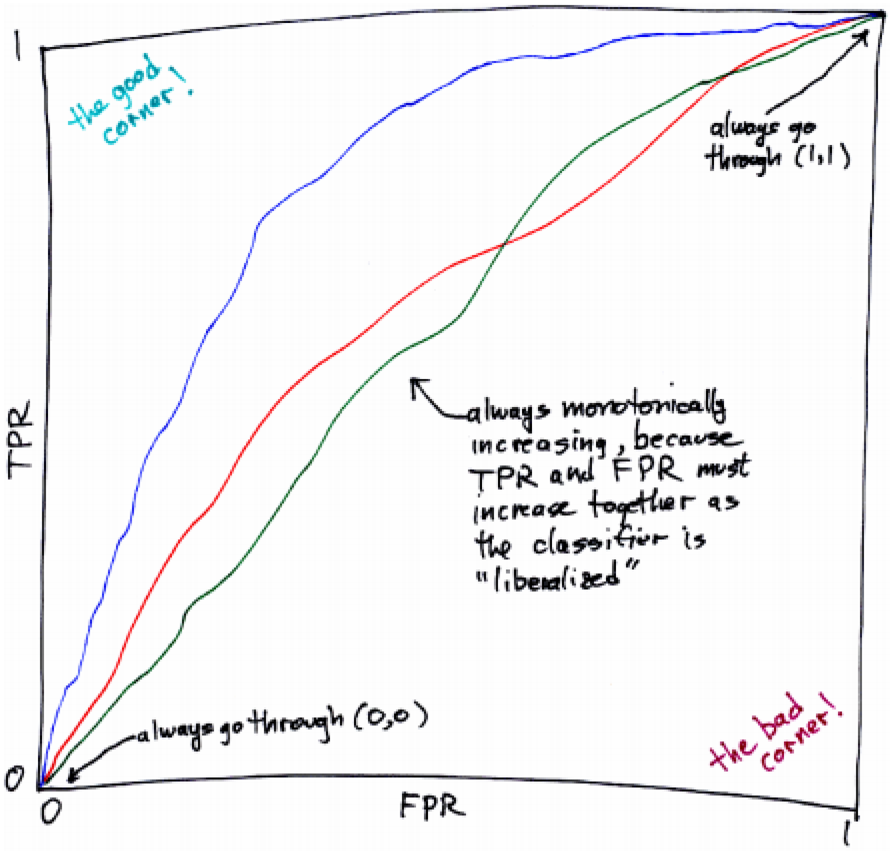
precision recall curve

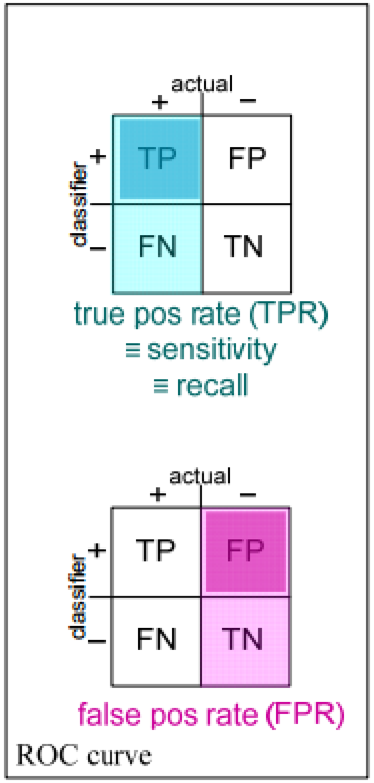
ROC (AUC)



Slide from The UT Austin, CS 395T, Spring 2008, Prof. William H. Press

ROC (“Receiver Operating Characteristic”) curves plot TPR vs. FPR as the classifier goes from “conservative” to “liberal”





true pos rate (TPR)
≡ sensitivity
≡ recall

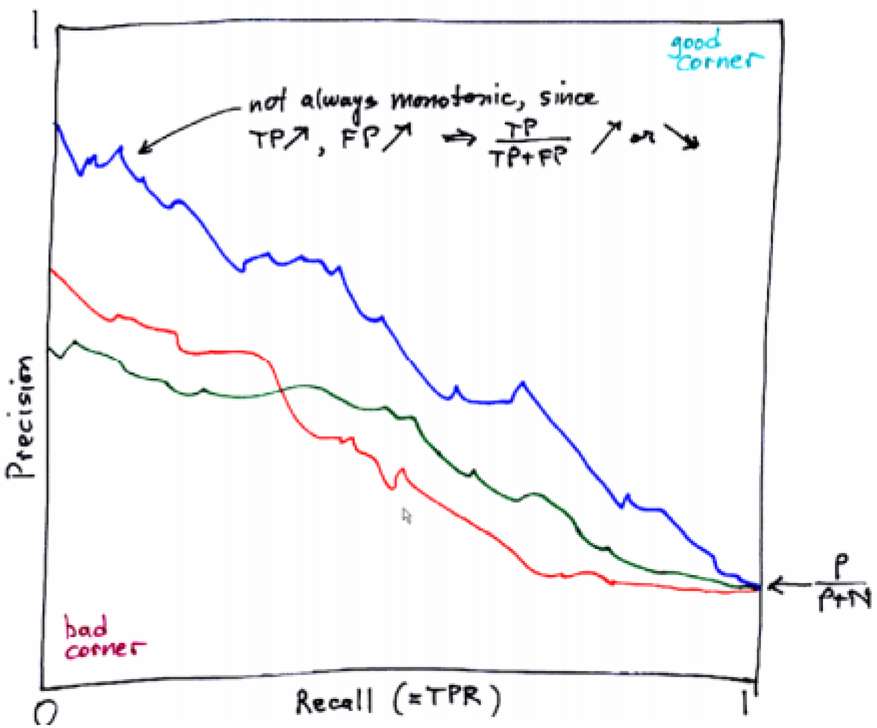
false pos rate (FPR)

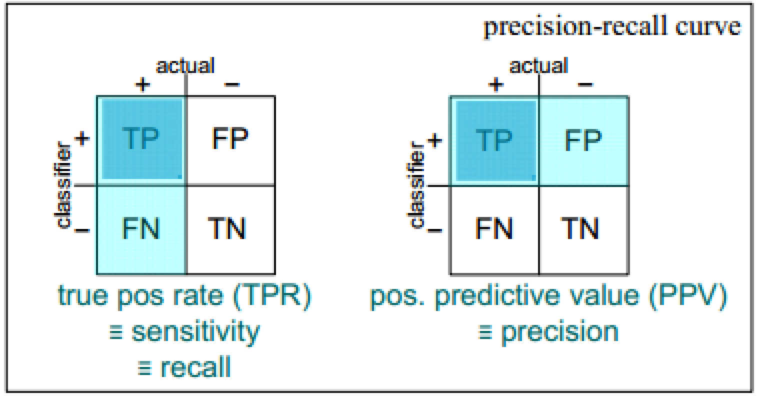
ROC curve

blue dominates red and green
neither red nor green dominate the other

You could get the best of the red and green curves by making a hybrid or “Frankenstein” classifier that switches between strategies at the cross-over points.

Precision-Recall curves overcome this issue by comparing TP with FN and FP

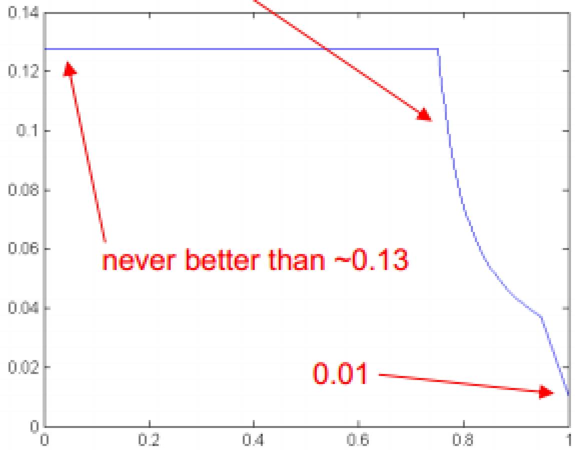




true pos rate (TPR)
≡ sensitivity
≡ recall

pos. predictive value (PPV)
≡ precision

By the way, this shape "cliff" is what the ROC convexity constraint looks like in a Precision-Recall plot. It's not very intuitive.

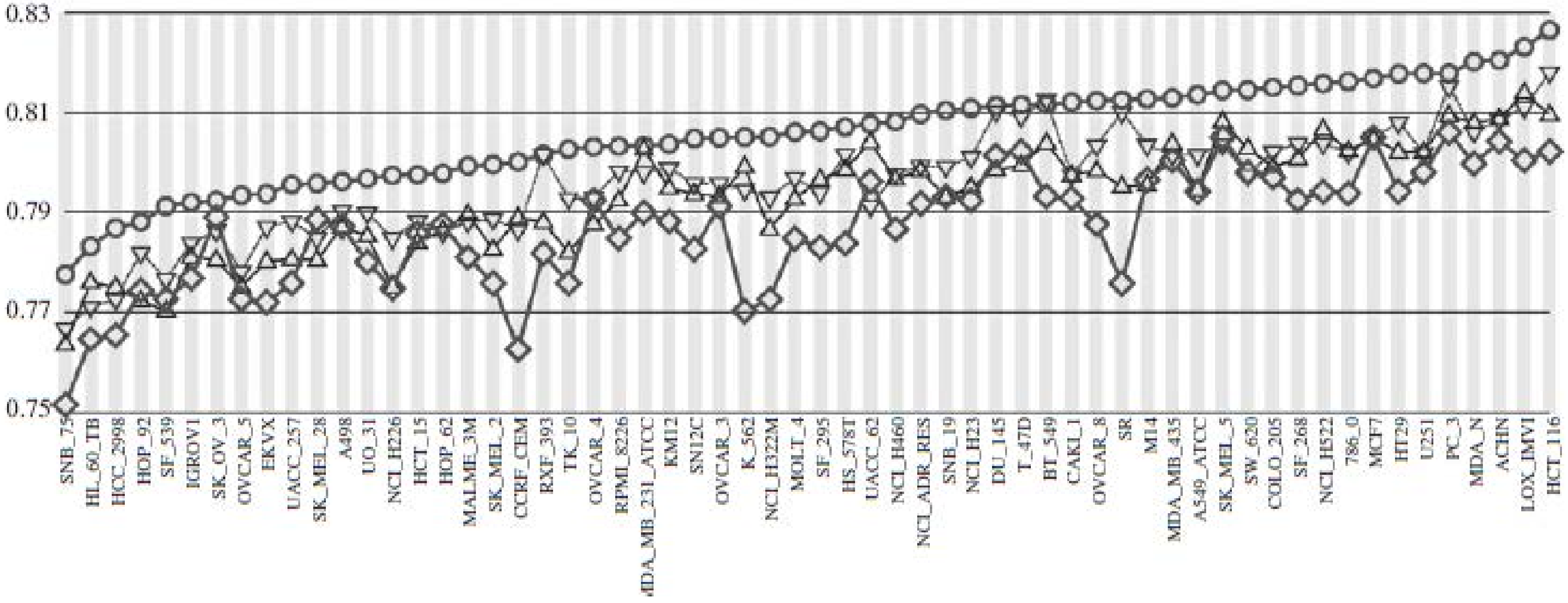


Continue our toy example:

note that P and N now enter

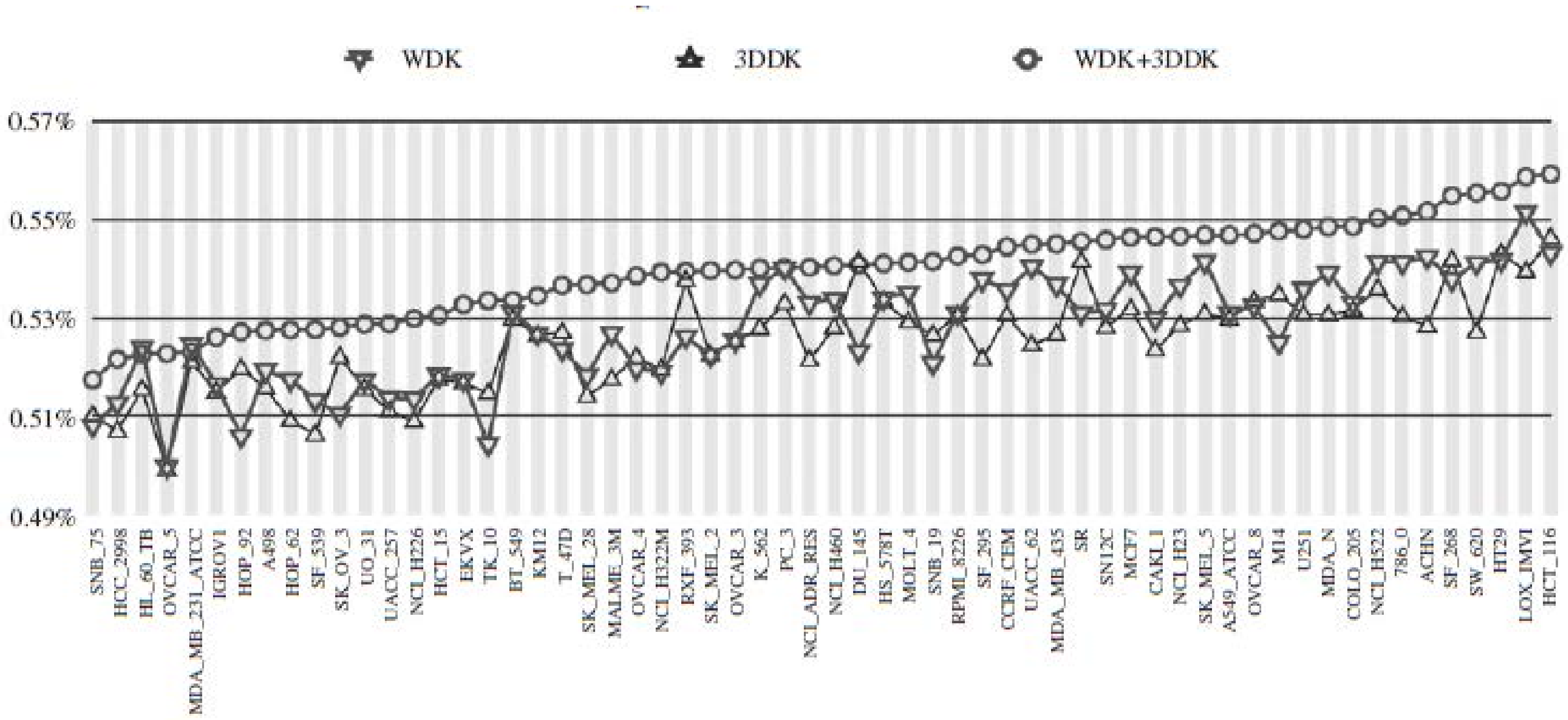
```
prec = tpr*100./(tpr*100+fpr*9900);
prec(1) = prec(2); % fix up 0/0
reca = tpr;
plot(reca,prec)
```

Results: NCI Cancer screening dataset.



ROC AUC values

Results: NCI Cancer screening dataset.



precision/recall curve values

Result: NCI Anti-HIV screening dataset

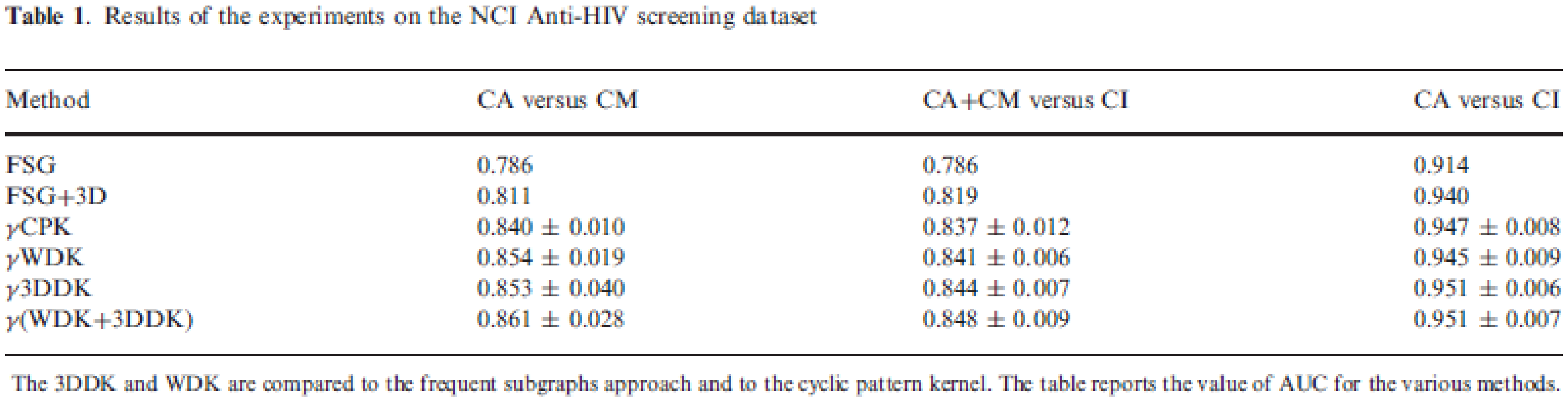


Table 1. Results of the experiments on the NCI Anti-HIV screening dataset

Method	CA versus CM	CA+CM versus CI	CA versus CI
FSG	0.786	0.786	0.914
FSG+3D	0.811	0.819	0.940
γCPK	0.840 ± 0.010	0.837 ± 0.012	0.947 ± 0.008
γWDK	0.854 ± 0.019	0.841 ± 0.006	0.945 ± 0.009
γ3DDK	0.853 ± 0.040	0.844 ± 0.007	0.951 ± 0.006
γ(WDK+3DDK)	0.861 ± 0.028	0.848 ± 0.009	0.951 ± 0.007

The 3DDK and WDK are compared to the frequent subgraphs approach and to the cyclic pattern kernel. The table reports the value of AUC for the various methods.