High-Throughput Screening with Two-Dimensional Kernels

Chloé-Agathe Azencott and Pierre Baldi

Institute for Genomics and Bioinformatics
University of California, Irvine

IJCNN07 - Agnostic Learning vs. Prior Knowledge Challenge

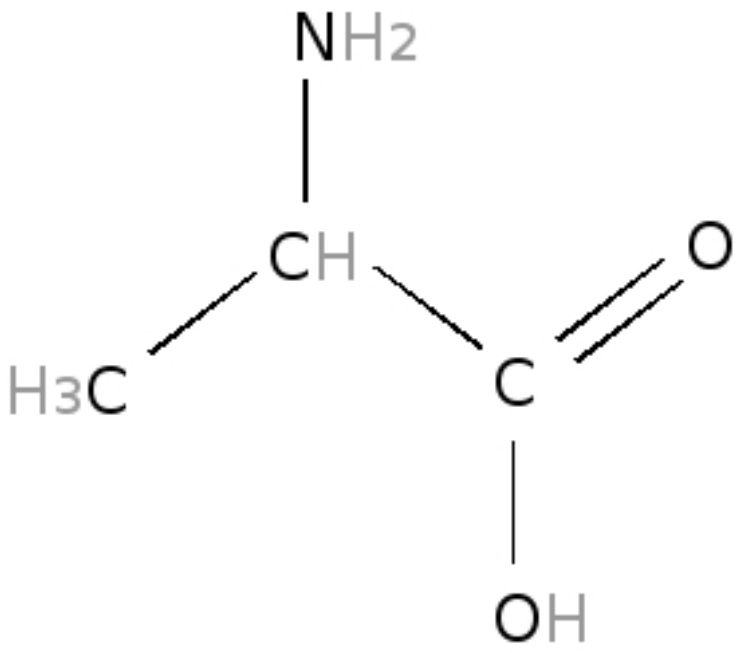
High-Throughput Screening

- Drug discovery
- Quickly test thousands of molecules to identify possible drug candidates
- *in silico* (Virtual Screening):
  - less resources (time and money)
  - ability to test virtual compounds (not yet synthetized)

Outline

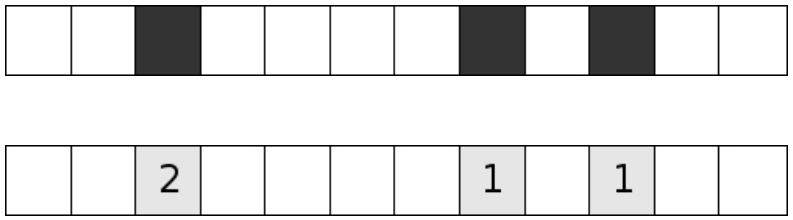
Similarity Between Two Molecules

How can we define the similarity between two molecular compounds?



- each node = atom
- each edge = bound
- Graph can be labeled (e.g. with atom name and type of bound)

Molecular Fingerprints



- Presence/Absence or count of each feature
- Sparse and long ($\simeq 100,000$)
  (=> compression)

Extended-Connectivity Fingerprints (ECFP)

- Assign a label L to each atom (node)
- At each iteration: update the label of each node

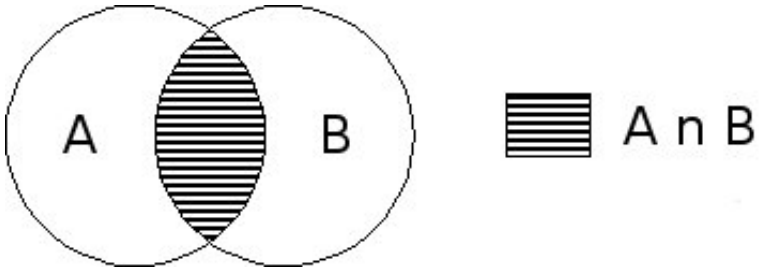
$$L_{new} = h(L_{old}, L_1, \ldots, L_k)$$

  where $L_1, \ldots, L_k$ are the labels of the neighboring nodes and h is a hashing function
- Each final label is a feature

Labeling Schemes

- Atoms (Nodes):
  - Element (or atomic number): Carbon, Oxygen, etc...
  - Element-Connectivity: Element + # connected atoms
  - Sybyl typing of atoms: Element + property (amide, hybridization, ...)
  - etc...
- Bonds (Edges):
  - Type of bond (single, double, triple, aromatic...)
  - No label
  - etc...

Tanimoto



- Tanimoto (binary):

$$\frac{A \cap B}{A \cup B}$$

- MinMax (counts):

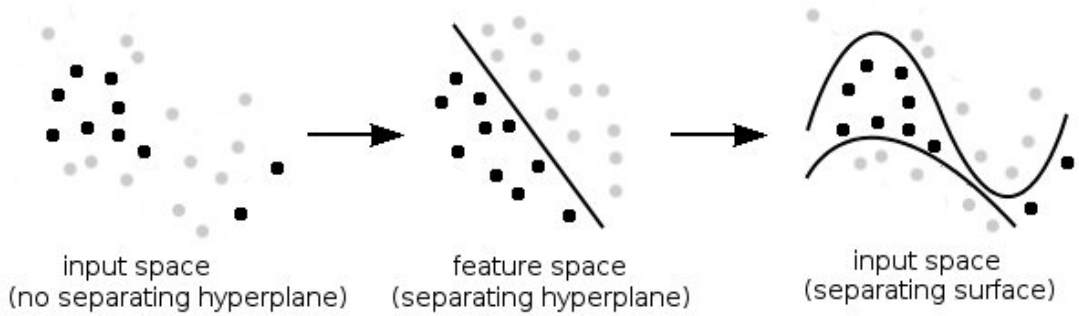
$$\frac{\sum_i \min(A_i, B_i)}{\sum_i \max(A_i, B_i)}$$

reduces to Tanimoto in the binary case

Support Vector Machines

How can we use these similarities in the framework of SVM for the classification of molecular compounds?

Feature Space Matching



- Φ: map the input space ($\mathcal{X}$) to a feature space ($\mathcal{H}$) where the data is linearly separable
- kernel k: $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

Support Vector Machines

- $$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x) + b$$

  where:
  - k = kernel
  - $x_1, \ldots, x_n$ = training examples
- Computation of the optimal manifold in $\mathcal{H}$
- Size of the dataset => need for an online implementation (SVMTorch)

MinMax is a kernel

Theorem

Tanimoto and MinMax are kernels

- We can map molecules to a linear space
- We can use a SVM for problems where data = molecules

Optimization

- Grid optimization: hyperparameters C (error-margin tradeoff) and ϵ (insensitivity) of the SVM (so as to minimize the BER)
- Dealing with unbalance: Oversampling
  - divide negative set in k subsets of the size (roughly) of the positive set
  - train each of the k sub-classifiers independantly
  - final decision:
    - each sub-classifier has a vote
    - threshold optimized by cross-validation

Performance of the method on the HIVA dataset of the AL vs. PK challenge

HIVA dataset

- Set of molecular compounds together with their activity towards HIV
- HTS problem
- 3.5% active compounds

10-fold Cross-Validated Results

- Labels: Atomic number, Bound type
- Number of iterations (while building ECFP): 2
- Similarity measure: MinMax (i.e. counts representation)
- 10-fold Cross-Validated BER = **0.2238**
- BER on the train set = **0.000**

Final Results

- BER = **0.2693**
- Winning BER in the Prior Knowledge track
- Best Agnostic Learning BER = **0.2741**

Conclusion

- Small training set but limited **overfitting** => hope of (relatively) **good prediction performance** on the testing set
- Method can be applied to **HTS** but also to **various problems** in the **chemistry domain** and also to any domain where the data can be represented by **graphs**
- **Fast** enough to be applied to large datasets

References I

- L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi
  Graph Kernels for Chemical Informatics
  Neural Networks, 18(8):1093–1100, 2005.
- M. Hassan, R. D. Brown, S. Varma-O'Brien and D. Rogers
  Cheminformatics analysis and learning in a data pipelining environment
  Molecular Diversity, 10:283–299; 2006.
- R. Collobert and S. Bengio
  SVMTorch: Support Vector Machines for Large-Scale Regression Problems
  J. Mach. Learn. Res., 1:143–160, 2001.
  http://www.idiap.ch/learning/SVMTorch.html

Acknowledgments I

- Dr. S. Joshua Swamidass
- Pr. Pierre F. Baldi

Thank you!

Is BER a good performance measure for HIVA?

- $$BER = 1 - \frac{1}{2}\left(\frac{TP}{p} + \frac{TN}{n}\right)$$
- BER quantifies for good separation
- But HTS: Find as many hits as possible very early (early recognition)
- BEDROC (Boltzmann-enhanced discrimination of receiver operating characteristic)

BEDROC

- Generalization of AUC (Area under the ROC curve)
- α: $\alpha.Ra << 1$ and $\alpha \neq 0$, where Ra is the ratio of active compounds ($Ra = \frac{p}{N}$)
- Compare the classifier to an exponential probability distribution of parameter α (instead of uniform)
- $$BEDROC \approx \frac{1}{Ra.\alpha}\left(\frac{\sum_{i=1}^{p} e^{-\alpha.(r_i/N)}}{\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}}\right) + \frac{1}{1+e^{-\alpha}}$$

J.-F. Truchon and C. I. Bayly
Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem
J. Chem. Inf. Model., 47(2):488–508, 2007.