PREDICTION OF MOLECULES ACTIVITY USING GRAPHICAL AND GEOMETRICAL KERNELS

F. Costa    A. Ceroni    P. Frasconi

Dipartimento di Sistemi e Informatica
Università degli Studi di Firenze
via di S. Marta 3, 50139 Firenze, Italy

Joint APrIL/IQ Workshop
Titisee - March 2006
OUTLINE

1. **Motivation**
   - Structured Learning in Bio-Chemistry
   - Open Issues

2. **2D and 3D Decomposition Kernels**
   - 2D Decomposition Kernel
   - 3D Decomposition Kernel

3. **Experimental Results**
   - NCI Cancer Dataset
   - NCI HIV Dataset
THE TASK

- Quantitative Structure Activity Relationship (QSAR) is an important task in bio-chemistry
- QSAR: deal with molecules of variable size/structure and predict their physical/chemical/biological properties
- Molecules are naturally represented as graphs with labels on vertices and edges \( \leadsto \) machine learning over \textit{structured} data
**STRUCTURED LEARNING IN BIO-CHEMISTRY**

**LITERATURE CHEMICAL GRAPH (2D)**
- Descriptor-based QSAR methods use fixed set of features determined by expert \(\rightsquigarrow\) limited to known aspects
- Graph mining (Kramer 01, Deshpande 03) \(\rightsquigarrow\) graph isomorphism
- Graph kernels (Horváth 04, Mahé 04, Swamidass 05)

**LITERATURE 3D**
- Spatial information on atom pair-wise distance (Deshpande 03, Swamidass 05) \(\rightsquigarrow\) less structured information
- Binding energy from interacting atomic/simple molecular probe at grid-points \(\rightsquigarrow\) alignment problem

---

**Motivation**
2D and 3D Decomposition Kernels
Experimental Results
Conclusions

---

F. Costa, A. Ceroni, P. Frasconi
Graphical and Geometrical Kernels
**Issue...& Good News**

- When molecule 3D models are synthesised from 2D chemical information (ex. CORINA Sadowski 03) $\rightsquigarrow$ noise
- Adding 3D information is useful even with noise!

**Proposed Approach**

In order to exploit different sources of information we use combination of:

- Graph kernel on chemical 2D structure
- Graph kernel on physical 3D structure
**Decomposition Kernel Approach**

- Decomposition is a **flexible** approach for inducing a similarity measure over complex objects based on the similarity of objects parts (*Haussler 99*)

**Problem**

- Complex object \( \rightsquigarrow \) consider **all** possible subparts \( \rightsquigarrow \) **many** possible large sub-parts
- Feature space grows **exponentially** (with sub-part size) \( \rightsquigarrow \)
  - objects are very similar to themselves and very different from any other object
  - diagonal-dominant Gram matrix \( \rightsquigarrow \) **poor generalisation**
ISSUES IN DECOMPOSITIONS

APPROACHES FOR FEATURE SPACE EXPLOSION

- Use prior knowledge to select only relevant parts
- Feature selection via mining frequent substructures (Karypis 03)
- Down-weight the contribution of larger/more complex fragments (Collins 01)
- Kernel transformation (polynomial with degree < 1) (Schoelkopf 02)
THE PROPOSED APPROACH

**2D Decomposition Kernel**
- Decompose an object in all its sub-parts, but...
  - use small fragments to identify matching sub-structures \( \leadsto \) reduce feature space explosion
  - use summarised information over correspondent larger fragments \( \leadsto \) increase expressive power

**3D Decomposition Kernels**
- Decompose a 3D object in small sub-graphs, but...
  - use chemical information to avoid combinatorial complexity
  - do not resort to exact graph matching
Decompositions

**Decomposition Kernel (Haussler 99)**

\[ K(x, x') = \sum_{s \in \text{parts of}(x)} \kappa(s, s') \]
\[ s' \in \text{parts of}(x') \]

**Definition**

\[ \kappa(s, s') = \text{kernel over parts} \]
MATCHING KERNEL

\[ K(x, x') = \sum_{s \in \text{parts of}(x)} \sum_{s' \in \text{parts of}(x')} \delta(s, s') \]

\[ \delta(s, s') = \begin{cases} 
1 & \text{if } s = s' \\
0 & \text{otherwise} 
\end{cases} \]
**WEIGHTED DECOMPOSITIONS**

### 2D WEIGHTED DECOMPOSITION KERNEL

\[ K(x, x') = \sum_{(s, z) \in \text{parts of}(x)} \sum_{(s', z') \in \text{parts of}(x')} \delta(s, s') \kappa(z, z') \]

- \( s = \text{selector} \)
- \( z = \text{context} \)

**TERMINOLOGY**
Matching Contexts

Context Kernel

- No exact match of the context but of attribute histogram
- The match between the contexts $z$ and $z'$ is computed as:

$$\kappa_d(z, z') = \sum_{j=1}^{m} p(j)p'(j)$$

where $p(j)$ is frequency count of attribute $j$
3D Decomposition Kernel

**Description**

- Represent a molecule as fully connected graph $G$:
  - vertices are atoms with associated properties
  - edges are labelled with associated properties and pair-wise atom distance information
- Let a **shape** be a fully connected (small) sub-graph of $G$
**3D Decomposition Kernel**

**The Kernel**

- **Idea:** decompose the molecule in a set of shapes
- Let the 3D kernel be a set kernel over the shapes of the decomposed graph:

\[
K(x, x') = \sum_{s \in \text{shape of}(x)} \sum_{s' \in \text{shape of}(x')} \kappa(s, s')
\]
To reduce complexity we do not consider all possible shapes (avoid combinatorial explosion)

Shapes are complete graphs of atoms connected by paths of maximum length $l$ (e.g., $l = 1$) in the chemical graph.
**3D Decomposition Kernel**

**The Shape Kernel**

- **Idea:** Use a decomposition kernel over the shape

- Shape as a set $s = \{ E_i \}$ with $E_i = (t_i, d_i)$ where $t_i$ = string encoding vertices & bond type and $d_i$ = distance between atoms

$$
\kappa(s, s') = \prod_{i=1}^{\vert E \vert} k(E_i(s), E_i(s'))
$$

- where $k(\cdot, \cdot)$ is the edge kernel

$$
k(E_i(s), E_i(s')) = \delta(t_i, t_i') e^{\sigma |d_i - d_i'|^2}
$$
**Problem**

- Dataset from National Cancer Institute (NCI)
- 70,000 compounds screened for suppressing or inhibiting activity on 60 human tumour cell lines
- $\approx 3500$ compounds per cell line
- 3D coordinates generated by CORINA (Sadowski 03) from 2D structure

**Prediction Task**

- **Target**: Balanced binary classification problem
- **Measure**: Accuracy and mean of the ROC area (AUC) on a ten folds cross validation
NCI CANCER DATASET

COMPARISON

- **Swamidass et al 2005 MinMax Kernel** SVM on generalised fingerprints based on depth-first visits
- **Swamidass et al 2005 3D Histogram** SVM over histogram of distance between all pairs of atom type
**FIGURE:** Accuracy and AUC values for 60 cell lines of NCI Dataset. Results sorted by MinMax accuracy. **3DK parameters:** \( l = 3 \), **WDK parameters:** \( l = 3 \), **WDK+3DK** polynomial with \( d = 2 \).
**NCI CANCER DATASET**

**Figure:** AUC values for 60 cell lines of NCI Dataset. Results sorted by WDK AUC. **3DK parameters:** \( l = 3 \), **WDK parameters:** \( l = 3 \), **WDK+3DK** polynomial with \( d = 2 \).
NCI HIV Dataset

Problem
- NCI dataset with 42,687 compounds classified for evidence of anti–HIV activity as:
  - confirmed active (CA): 100% protection
  - moderately active (CM): >50% protection
  - confirmed inactive (CI)

Prediction Task
- Target: Three unbalanced binary classification problems:
  - CA vs CM (1,503 molecules, pos/tot=28.1%)
  - CA+CM vs CI (42,687 molecules, pos/tot=3.5%)
  - CA vs CI (41,606 molecules, pos/tot=1%)
- Measure: mean and standard deviation of the ROC area (AUC) on a five folds cross validation
Datasets and Literature

Comparison

- **Karypis et al 2003** Frequent Sub-Graph Algorithm (FSG) SVM on mined frequent sub-graphs feature space
- **Horváth et al 2004** Cyclic Pattern Kernel (CPK) SVM on cyclic paths feature space
## Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CA vs CM</th>
<th>CA+CM vs CI</th>
<th>CA vs CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSG</td>
<td>78.6%</td>
<td>78.6%</td>
<td>91.4%</td>
</tr>
<tr>
<td>FSG+3D</td>
<td>81.1%</td>
<td>81.9%</td>
<td>94.0%</td>
</tr>
<tr>
<td>$\gamma_{CPK}$</td>
<td>84.0 ± 1.0%</td>
<td>83.7 ± 1.2%</td>
<td>94.7 ± 0.8%</td>
</tr>
<tr>
<td>$\gamma_{WDK}$</td>
<td>85.4 ± 1.9%</td>
<td>84.1 ± 0.6%</td>
<td>94.5 ± 0.9%</td>
</tr>
<tr>
<td>$\gamma_{3DK}$</td>
<td>85.3 ± 4.0%</td>
<td>84.4 ± 0.7%</td>
<td>95.1 ± 0.6%</td>
</tr>
<tr>
<td>$\gamma(\text{WDK+3DK})$</td>
<td>86.1 ± 2.8%</td>
<td>84.8 ± 0.9%</td>
<td>95.1 ± 0.7%</td>
</tr>
</tbody>
</table>

**Table**: AUC values for NCI Anti-HIV Dataset. **3DK parameters**: $l = 3$, **WDK parameters**: $l = 4$ & graph complement, **WDK+3DK** gaussian with $\gamma = 1$. 
CONCLUSIONS

SUMMING UP..

- It is possible to exploit 3D information (even if noisy) for QSAR tasks
- We showed a simple (but effective) method for controlling feature space high dimensionality side-effects