Chemical Structure Generation

Richard Wilson

With contributions David White
Dept. of Computer Science
University of York
Generative Models

- Generative models are well known in Pattern Recognition

- Used for many applications
  - Classification
  - Clustering
  - Creating new samples
Graphs in Pattern Recognition

• Graphs are a common representation in PR
  – Many data types defined as relations between parts
Generative Models of Graphs

- Generative models of graphs are more difficult
  - Discrete, strong structural component
  - Order of vertices not known apriori

- White and Wilson [ICIAP 07, David White’s thesis]
  - **Alignment** to create a common space
  - **Representation**: Adjacency matrix, Laplacian, Eigendecomposition
  - **Model**: Multivariate normal
  - **Sampling**: simple procedure with post-threshold
Chemical Structure Graphs

• Graphs can naturally be used to represent chemical structure

```
\begin{align*}
S_1 &= \text{chemical structure 1} \\
S_2 &= \text{chemical structure 2} \\
S_3 &= \text{chemical structure 3} \\
S_4 &= \text{chemical structure 4} \\
S_5 &= \text{chemical structure 5} \\
S_6 &= \text{chemical structure 6}
\end{align*}
```

• Can we build a generative model of chemical structure?
The plan:

**Input** → Set of compounds with interesting properties (chemical structure graphs)

1. **Alignment** of structures
2. **Represent** as weighted adjacency matrices
3. **Model** as a probability distribution
4. **Sample** from the model to get new chemical structures

→ **Output**: Set of novel compounds from the same distribution as the input

“Generative models for Chemical Structures” D. White and R. C. Wilson, JCIM July 2010
Alignment

• We need to align the graphs into a common space
• Problem: the graphs can be structurally very different
  – If they are too different, alignment will be wrong or meaningless
  – Also avoid aligning a large graph to a small graph
• Example

![Diagram showing alignments between different graph structures.](image-url)
• We use a hierarchical alignment procedure
• Coarse and quick measure of similarity (fingerprints)
  – Must be quick to compute
• Align most similar first, keep the largest
  – All descendents are aligned similarly
  – Tree of alignments:
Representation

- Simple representation as weighted adjacency matrix
  - Edges are 0.5 for a single bond, 1.0 for a double
  - Vertices are weighted by periodic group
  - Captures limited information about the properties
- Then vectorised to into a vector space representation

\[
A = \begin{pmatrix}
0.6 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.5 & 0.5 & 0.5 & 0 & 0 & 0 & 1 & 0 \\
0 & 0.5 & 0.5 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0.5 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.5 & 0.5 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0.5 & 0.5 & 0 \\
0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0.5 \\
0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.7 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.7
\end{pmatrix}
\]

\[
a = \begin{pmatrix}
0.6 \\
0.5 \\
0 \\
0 \\
\vdots
\end{pmatrix}
\]
• **Modelling:** the distributions are complex, so we model them using a Gaussian Mixture Model (GMM)
  – Cannot fit a GMM in such high-dimensional space
  – Use PCA to reduce the dimensionality while maintaining main variations
  – Algorithm of Figueiredo & Jain[2002] to fit GMM parameters
**Sampling** from the GMM is straightforward

- Problem: The GMM generates new *graphs* not new chemical structures
  - Not valid chemical structures; do not respect valency or geometric considerations

\[
\begin{pmatrix}
0.5 & 0.5 & 0.05 & 0.02 & 0.14 & 1 \\
0.5 & 0.5 & 1 & 0.4 & 0.1 & 0.01 \\
0.05 & 1 & 0.5 & 0.5 & 0.02 & 0.3 \\
0.02 & 0.4 & 0.5 & 0.5 & 1 & 0.15 \\
0.14 & 0.1 & 0.02 & 1 & 0.5 & 0.5 \\
1 & 0.01 & 0.3 & 0.15 & 0.5 & 0.5
\end{pmatrix}
\]
Sampling

• We solve this problem by creating a projection set of valid structures
• The projection set covers the local chemical structure space
  – We use it to find a similar graph which is also a chemical structure
• Method: Generate a set of chemical structures in the neighbourhood of the input set
  1. Decompose the input set into fragments
  2. Combine fragments to new compounds (the projection set)
  3. Align projection set to input set via similarity
• Then project each generated graph onto the closest chemical structure in the projection set
Some results

• We have used the Directory of Useful Decoys (DUD)
  – Evaluation domain is protein docking affinity
  – DUD also includes examples with similar physical properties which do not dock well (‘decoys’)
• Example 1: COX2 protein
Results

Cox2

+ Input set
Projection set
Generated set
COX2 Input

\[ S_1, S_2, S_3, S_4, S_5, S_6 \]

COX2 Generated

\[ G_1, G_2, G_3, G_4, G_5, G_6 \]
• COX2 binding affinities
Example 2: EFGR protein
Mixture model and generated graphs
Input set
Projection set
Generated set
efgr
EFGR
Some areas for improvement

- The two stage process for generating structure is inefficient (graph model and projection set)
  - Can we directly model chemical structure graphs?
- There is no reference to the 3D shape
  - Shape models
- The resulting molecules may not easy to synthesise
  - Need a model of chemical synthesis to identify interesting compounds