Spectral graph theory and its applications to molecular graphs

NCATS Informatics Seminar
Febuary 17, 2016

Introduction

- Spectral graph theory
- Well-known matrices
  - Adjacency
  - Laplacian
  - Normalized Laplacian
- Spectral properties
- Molecular applications
  - Layout
  - Canonicalization
  - Invariant
  - Descriptor
- What's next?
- Implementation challenges

Spectral graph theory

- Graph G consists of a set of n vertices $V = \{v_1, v_2, \ldots, v_n\}$ and m edges $E = \{e_1, e_2, \ldots, e_m\}$ where $e_k = v_i \sim v_j$
- Let M be a matrix that encodes G based on V, E, or combinations thereof
- Spectral graph theory is about understanding the properties of G in terms of eigenvalues and eigenvectors of M, i.e.,

$$M\mathbf{v}_i = \lambda_i \mathbf{v}_i,$$

  where λ_i is the ith eigenvalue and $\mathbf{v}_i$ is the corresponding eigenvector.
- The eigenvalues $\{\lambda_i\}$ define the *spectrum* of G
- Outstanding problem: Which graphs are determined by their spectrum?
  - Under what conditions do non-isomorphic graphs have the same spectrum?

Graph spectrum

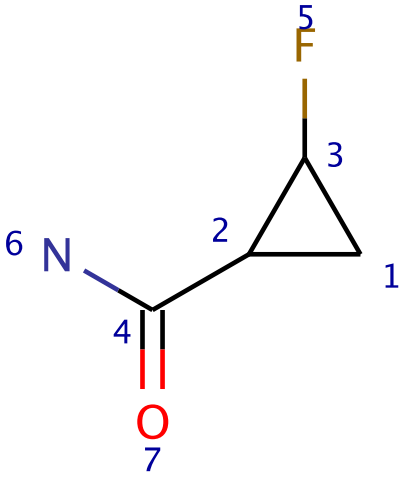
Adjacency

The adjacency A representation of G is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

Foundation of Hückel theory

The topology of a molecule, rather than its geometry, determines the form of the Hückel molecular orbitals.



$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

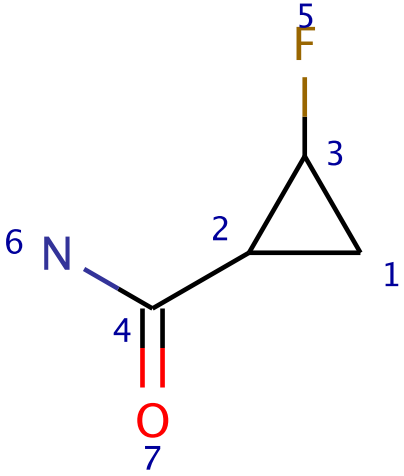
Graph spectrum

Laplacian

Let D be the degree matrix of G, i.e., $D_{ii} = \text{degree}(v_i)$ and 0 elsewhere, we have the Laplacian L defined as follows

$$L = D - A,$$

where A is the adjacency matrix.



$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 3 & 0 & -1 & -1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

Graph spectrum

Normalized Laplacian

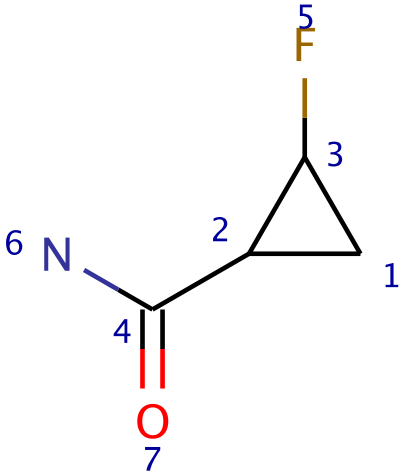
The normalized Laplacian is defined as

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}},$$

or

$$\tilde{L}_{ij} = \begin{cases} 1 & i = j \\ -\frac{1}{\sqrt{d_i d_j}} & i \neq j \end{cases}$$

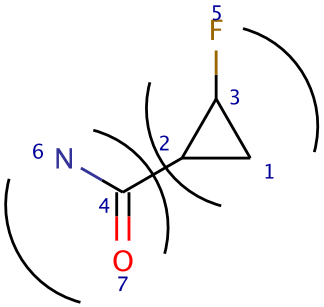
where d_i and d_j are the degrees of v_i and v_j, respectively.



$$\tilde{L} = \begin{pmatrix} 1.0 & -0.4 & -0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.4 & 1.0 & -0.3 & -0.3 & 0.0 & 0.0 & 0.0 \\ -0.4 & -0.3 & 1.0 & 0.0 & -0.6 & 0.0 & 0.0 \\ 0.0 & -0.3 & 0.0 & 1.0 & 0.0 & -0.6 & -0.6 \\ 0.0 & 0.0 & -0.6 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.6 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.6 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$

Spectral properties

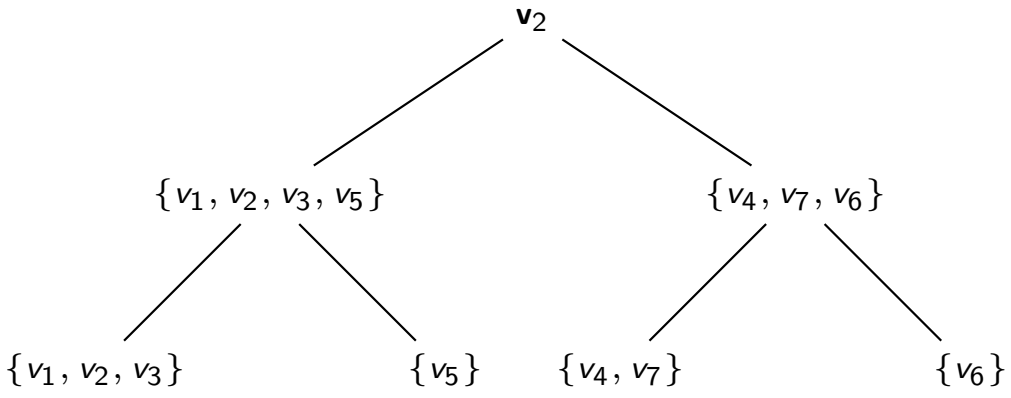
- The spectrum of A is bounded by the maximum degree in G, i.e., $|\lambda_i| \leq \max_k d(v_k)$ for $k = 1, 2, \ldots, n$. For organic molecules, $|\lambda_i| \leq 4$.
- L and $\tilde{L}$'s spectra are non-negative, i.e., $\lambda_i \geq 0$. L and $\tilde{L}$ are semidefinite.
- Multiplicity of $\lambda_i = 0$ in L and $\tilde{L}$ is the number of connected components in G.
- The spectrum of $\tilde{L}$ is bounded by 2, i.e., $0 \leq \lambda_i \leq 2$.
- Let $\lambda_1 = 0 \leq \lambda_2 \leq \cdots \leq \lambda_n$ for L and $\tilde{L}$. The first non-zero λ_i is the *algebraic connectivity* index with the corresponding eigenvector known as the *Fiedler* vector. This vector provides near-optimal 2-partition of G. The Fiedler vector is the foundation of many spectral clustering algorithms.



$$\mathbf{v}_2(\tilde{L}) = \begin{bmatrix} 0.29255 \\ 0.11507 \\ 0.44483 \\ -0.53342 \\ 0.32870 \\ -0.39416 \\ -0.39416 \end{bmatrix}$$

Molecular applications

- Layout — for high symmetry graphs, the eigenvectors $\mathbf{v}_2$ and $\mathbf{v}_3$ of L can be used as coordinates for 2-D embedding.
- Canonicalization — a simple algorithm can be derived based on the eigenvectors of L or $\tilde{L}$. Start with the Fiedler vector and recursively partition the vertices until each partition contains only one vertex. The canonical ordering is the depth first traversal of the binary partition tree.



Molecular applications (cont'd)

Invariant — which graphs are determined by their spectrum?

tripod.nih.gov/?p=522

Tripod Development Cheminformatics proving ground

Home GInAS NPC browser Siphonify Overview Screenshots Jobs About

Do structurally similar molecules have similar hash codes?

Molecular hash codes are fixed-length alphanumeric encoding of molecular graphs. They play a key role within chemical data management systems in facilitating (among other things) structural identity and uniquess validation. While an important component of any hash code generation approach is the structure standardization procedure (something which we briefly discussed previously), the focus of this post is on the encoding step, i.e., that of generating the hash code given a standardized structure. In the remainder of this post, we first give a brief overview of **spectral hash code**, our proposed multi-resolution hash code based on InChI. We then highlight the utility of the hash code through a number of examples. We conclude with the description of a web resource for converting between PubChem CID, InChIKey, and **spectral hash code**. The complete source code is readily available from our bitbucket repository. As always, we welcome comments and feedback!

A **spectral hash code** is a 30-character (150-bit) alphanumeric hash string that uniquely encodes an InChI. The hash code has three logical blocks—denoted as h_1, h_2, and h_3—that progressively encode the InChI graph at finer resolution where $h_1 \subseteq h_2 \subseteq h_3$ and $|h_1| = 9$ (45-bit), $|h_2| = 19$ (95-bit), and $|h_3| = 30$ (150-bit). Consider, for

Molecular applications (cont'd)

- Molecular descriptor — the spectrum can be directly used as molecular descriptors via Chebyshev polynomial expansion around each non-zero eigenvalues. Preliminary results correlate well with other molecular descriptors in RDKit:

Descriptor	Correlation
NumHeavyAtoms	0.887
LabuteASA	0.854
kappa1	0.837
MQN1	0.829
SMR	0.823
Chi1n	0.820
MQN26	0.807
MQN30	0.806
⋮	⋮

What's next?

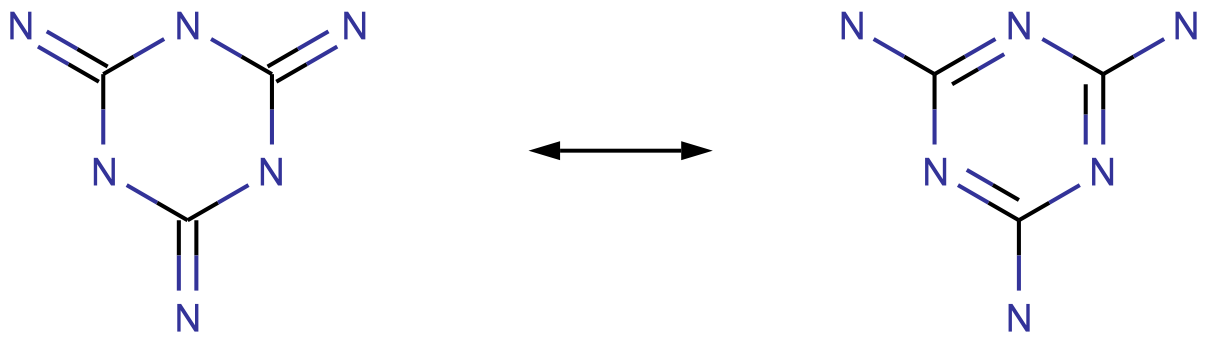
- Going beyond molecular topology with weighted graph based on experimental parameters; e.g.,

$$w(v_i, v_j) = \alpha \frac{m_{v_i} m_{v_j}}{r_{ij}^2},$$

  where m_{v_i} and m_{v_j} are the exact masses of atoms v_i and v_j, respectively, and r_{ij} is the measured bond length between the atoms. Other measurements are possible; e.g., partial charges, electronegativity, electron affinity, ionization energy, etc.
- Can the canonicalization algorithm be extended for automorphism and isomorphism detection?
- Can matrix perturbation theory be used to extend the graph invariant beyond cospectral?

Implemenation challenges

- Self-contained source code in C at `https://spotlite.nih.gov/ncats/spectral_hk`
- De novo InChI parsing
  - Bond order assignment
  - Tautomers



- Three different eigensolvers available: native Jacobi (slow), GNU scientific library (fast), and Intel's MKL (very fast).