Two recent applications of graph theory in molecular biology

Debra J. Knisley

25th Clemson Mini-Conference on Discrete Math and Algorithms

October 7, 2010
Outline

- New applications in molecular biology
  - Graphs as Secondary RNA structures
  - Graphs as Amino Acids
  - Graphs as Proteins
DNA to RNA to Protein

\[
\begin{align*}
\text{DNA} & : \text{GTGCATCTGACTCCCTGAGGAGAAG} \\
& \quad \text{(transcription)} \\
\text{RNA} & : \text{CACGTAGACTGAGGA} \\
& \quad \text{(translation)} \\
\text{protein} & : \text{VHLTPEEK} \\
\end{align*}
\]
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-letter</th>
<th>1-letter</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>nonpolar</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>polar</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asp</td>
<td>N</td>
<td>polar</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>polar</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td>nonpolar</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gln</td>
<td>Q</td>
<td>polar</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>polar</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>nonpolar</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>polar</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>nonpolar</td>
</tr>
</tbody>
</table>
Human Genome Project

- How many genes are in the Human Genome?
- Majority of biologists believed a lot more than 50,000
- Rice genome contains about 50,000 genes
- Big surprise, humans only have about half that many!
Major changes in the science of molecular biology

- The number of genes in the genome was just the first of many surprises!
- The belief that the sole function of RNA is to carry the "code" of instructions for a protein is not true.
- It is now known that more than half of the RNA molecules that have been discovered are "non-coding"
The two surprising discoveries are related—

It is not how many genes, it is how they are regulated that is important!

Now everyone is trying to understand regulatory networks - Systems Biology

The study of RNA molecules is at the forefront of the field of Systems Biology
Database of RNA Tree Graphs

List of RNA Tree Graphs by Number of Vertices
- 2 Vertices
- 3 Vertices
- 4 Vertices
- 5 Vertices
- 6 Vertices
- 7 Vertices
- 8 Vertices
- 9 Vertices
- 10 Vertices

List of RNA Tree Graphs by Functional RNA Class
- P5abc
- RNA in Signal Recognition Complex
- Single Strand RNA
- tRNA
- U2 snRNA
- 5S rRNA
- 70S
Secondary RNA structure: Tamar Schlick et.al, web resource RAG (RnaAsGraphs)
Vertices represent internal loops
Tree model of secondary RNA structure

Secondary RNA structure
And corresponding tree.

See [http://www.monod.nyu.edu](http://www.monod.nyu.edu) (Schlick et al.)
They can be large, but most folding algorithms predict smaller tree structures.

The tree representation of an RNA molecule’s secondary structure.
The RAG Database

The six trees of order 6 with their RAG \((n.z)\) index and color classification

6.1:

6.2:

6.3:

RNA in signal recognition complex

6.4:

6.5:

6.6:

tRNA
Modeling RNA Bonding

When two trees $T_1$ and $T_2$ containing $n$ and $m$ vertices respectively are merged, the resulting tree $T_{1+2}$ has a total of $n + m - 1$ vertices.
Modeling RNA Bonding

To accurately model RNA molecule bonding, we must consider all possible vertex identifications between two RNA tree models.

RAG Index: 4.2    RAG Index: 3.1
Modeling RNA Bonding

Merge 1

Merge 2

Merge 3

Merge 4
Modeling RNA Bonding

6.2

6.4

6.5

6.6
Modeling RNA Bonding

- For all 94 graphs on 2 through 9 vertices, we recorded every possible vertex identification resulting in a graph on 9 or less vertices.
- The data from each vertex identification was recorded and translated into a data vector:
  \[ [\langle c_1, c_2, \text{deg}(v_1), \text{deg}(v_2) \rangle, \langle y_1, y_2 \rangle] \]
Observations

Examples of Data Vectors from Red Trees and Black Trees

<table>
<thead>
<tr>
<th>Red Trees 8.5 and 8.15 Data Vectors</th>
<th>Black Trees 8.22 and 8.23 Data Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>[&lt; 1,1,1,1 &gt;, &lt; 1,0 &gt;]</td>
<td>[&lt; 1,1,1,5 &gt;, &lt; 0,1 &gt;]</td>
</tr>
<tr>
<td>[&lt; 1,1,1,2 &gt;, &lt; 1,0 &gt;]</td>
<td>[&lt; 1,1,2,4 &gt;, &lt; 0,1 &gt;]</td>
</tr>
<tr>
<td>[&lt; 1,1,2,1 &gt;, &lt; 1,0 &gt;]</td>
<td>[&lt; 1,1,3,4 &gt;, &lt; 0,1 &gt;]</td>
</tr>
<tr>
<td>[&lt; 1,1,2,2 &gt;, &lt; 1,0 &gt;]</td>
<td>[&lt; 1,0,1,6 &gt;, &lt; 0,1 &gt;]</td>
</tr>
</tbody>
</table>
Artificial Neural Network

Each tree’s final classification was calculated as an average of a linear combination of prediction values from the vertex identifications.

<table>
<thead>
<tr>
<th>Code</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a[8.5] := 2 \ast \text{Classify} (\langle 1, 1, 1, 2 \rangle)$</td>
<td>$\langle 1.99966, 0.00135 \rangle$</td>
</tr>
<tr>
<td>$b[8.5] := 4 \ast \text{Classify} (\langle 1, 1, 1, 1 \rangle)$</td>
<td>$\langle 4.00000, 2.68201 \times 10^{-7} \rangle$</td>
</tr>
<tr>
<td>$c[8.5] := 4 \ast \text{Classify} (\langle 1, 1, 1, 1 \rangle)$</td>
<td>$\langle 4.00000, 2.68201 \times 10^{-7} \rangle$</td>
</tr>
<tr>
<td>$d[8.5] := 4 \ast \text{Classify} (\langle 1, 1, 2, 1 \rangle)$</td>
<td>$\langle 3.99932, 0.00270 \rangle$</td>
</tr>
<tr>
<td>$\text{Class}[8.5] := \frac{a[8.5] + b[8.5] + c[8.5] + d[8.5]}{14}$</td>
<td>$\langle 0.99991, 0.00034 \rangle$</td>
</tr>
</tbody>
</table>
## Results: Values for the Classified RAG Trees

<table>
<thead>
<tr>
<th>RAG Index</th>
<th>Color Class</th>
<th>ANN Prediction</th>
<th>ANN Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Red</td>
<td>1.00000</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>7.10</td>
<td>Black</td>
<td>0.59091</td>
<td>Unclassifiable</td>
</tr>
<tr>
<td>7.11</td>
<td>Black</td>
<td>0.00045</td>
<td>Highly Not-RNA-Like</td>
</tr>
<tr>
<td>7.2</td>
<td>Red</td>
<td>0.99860</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>7.3</td>
<td>Red</td>
<td>0.99990</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>7.9</td>
<td>Black</td>
<td>0.43130</td>
<td>Unclassifiable</td>
</tr>
<tr>
<td>8.10</td>
<td>Red</td>
<td>0.99994</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>8.15</td>
<td>Red</td>
<td>0.56343</td>
<td>Unclassifiable</td>
</tr>
<tr>
<td>8.17</td>
<td>Black</td>
<td>0.36524</td>
<td>Not-RNA-Like</td>
</tr>
<tr>
<td>8.18</td>
<td>Black</td>
<td>0.00595</td>
<td>Highly Not-RNA-Like</td>
</tr>
<tr>
<td>9.6</td>
<td>Red</td>
<td>0.99991</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>9.11</td>
<td>Red</td>
<td>0.99993</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>9.13</td>
<td>Red</td>
<td>0.99740</td>
<td>Highly RNA-Like</td>
</tr>
<tr>
<td>9.27</td>
<td>Red</td>
<td>0.99795</td>
<td>Highly RNA-Like</td>
</tr>
</tbody>
</table>
Two recent applications of graph theory in molecular biology

<table>
<thead>
<tr>
<th>ID: (8,15)</th>
<th>( \lambda_2 = 0.3065 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID: (8,17)</td>
<td>( \lambda_2 = 0.3187 )</td>
</tr>
<tr>
<td>ID: (8,18)</td>
<td>( \lambda_2 = 0.3542 )</td>
</tr>
<tr>
<td>ID: (8,19)</td>
<td>( \lambda_2 = 0.3738 )</td>
</tr>
<tr>
<td>ID: (8,20)</td>
<td>( \lambda_2 = 0.3820 )</td>
</tr>
<tr>
<td>ID: (8,21)</td>
<td>( \lambda_2 = 0.3820 )</td>
</tr>
<tr>
<td>ID: (8,22)</td>
<td>( \lambda_2 = 0.4525 )</td>
</tr>
<tr>
<td>ID: (8,23)</td>
<td>( \lambda_2 = 1.0000 )</td>
</tr>
</tbody>
</table>

Diagram: 5S rRNA
Domination number = 4

Domination number = 3

Domination number = 2
Locating-dominating number

Vertex Dominating Vertex
1 9
2 7
3 4 & 10
5 4
6 7 & 10
8 4, 7 & 9

Combinatorial optimization
locating domination number = 4

locating domination number = 5

locating domination number = 5
Graphical measures

- We used two measures defined by domination invariants

\[ P_1 = \frac{\gamma + \gamma_t + \gamma_a}{n}, \quad P_2 = \frac{\gamma_L + \gamma_D}{n} \]

where \( \gamma \) is the domination number, \( \gamma_t \) is the total domination number, \( \gamma_a \) is the global allied domination number, \( \gamma_L \) is the locating domination number, \( \gamma_D \) is the differentiating domination number, and \( n \) is the number of vertices in the graph.
Results

P1 and Status

P2 and Status
Combinatorial Molecular Biology?

- In combinatorics, we find optimal structural properties by defining graphical invariants based on minimal/maximal achievements such as the domination number, chromatic number, etc.

- Can we utilize the graphical invariants from graph theory as descriptors of biomolecules? That is, do biomolecules exhibit optimal structural properties?
Computational Chemistry- Chemical Graph Theory

- Topological indices together with quantities derived from chemical properties are called molecular descriptors.
- There are over 3,000 molecular descriptors in the Handbook of Molecular Descriptors. Of those, approximately 1,600 can be classified as topological indices.
- A tool of QSAR, the process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity.
The dilemma

- It is generally believed that topological indices are not valid for large graphs.
- Unlike RNA, proteins are very large molecules.
- Graphical invariants are suited for large graphs, but are often intractable.
Determining the function of a protein

- When a new protein is discovered, we may not know its function
- There are two ways to address the question of function
- Compare its primary sequence (sequence of amino acids) to all other known proteins
- Predict (or determine) its 3D conformation and compare its conformation to all other known 3D conformations
Hamming distance

R-T-S-C-A-G-G-W-A-C  target sequence
R-T-S-C-A-G-G-W-A-C  target sequence

What is the Hamming distance in each case.

- Are these “equally close”
- No, some amino acids are more likely to substitute for each other than other amino acids. There exist Scoring Matrices (often based on the likelyhood of one amino acid being replaced by another)
- See: AAindex database
Graphs for amino acids
Graphs for amino acids

- Histidine (His)
  - Chemical structure
  - $pK_a$ values: $1.70$, $9.09$, $6.04$

- Aspartic Acid (Asp)
  - Chemical structure
  - $pK_a$ values: $1.95$, $9.60$, $3.71$
Table of graph invariants from vertex-weighted graphs

<table>
<thead>
<tr>
<th>Row</th>
<th>acid</th>
<th>G</th>
<th>g</th>
<th>d</th>
<th>c</th>
<th>m</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>1.000</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>R</td>
<td>54</td>
<td>36</td>
<td>9</td>
<td>0</td>
<td>1.750</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>42</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>1.600</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>44</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>1.600</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>1.333</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>Q</td>
<td>42</td>
<td>24</td>
<td>6</td>
<td>0</td>
<td>1.667</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>E</td>
<td>44</td>
<td>24</td>
<td>6</td>
<td>0</td>
<td>1.667</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>H</td>
<td>40</td>
<td>36</td>
<td>6</td>
<td>5</td>
<td>2.000</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>I</td>
<td>24</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>1.600</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>L</td>
<td>36</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>1.600</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>K</td>
<td>38</td>
<td>24</td>
<td>8</td>
<td>0</td>
<td>1.667</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>44</td>
<td>24</td>
<td>6</td>
<td>0</td>
<td>1.600</td>
<td>12</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>36</td>
<td>24</td>
<td>7</td>
<td>6</td>
<td>2.000</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>P</td>
<td>24</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>2.000</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>1.333</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>T</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>1.500</td>
<td>14</td>
</tr>
<tr>
<td>18</td>
<td>W</td>
<td>62</td>
<td>48</td>
<td>9</td>
<td>9</td>
<td>2.182</td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>Y</td>
<td>52</td>
<td>24</td>
<td>8</td>
<td>6</td>
<td>2.000</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>V</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>1.500</td>
<td>12</td>
</tr>
</tbody>
</table>
Selecting the invariants

We see that the measure $m$ - average degree changed the relationship of R and K quite a bit.
Selecting the invariants

Four graphical invariants that capture the basic properties of the amino acids

Possible add-on tool for scoring multiple sequence alignments
What is reverse engineering?

- Building a mathematical model from observations
- The model is a representation of the biomolecule, network or whatever system under observation
- The model is adjusted to fit the observations
The DREAM CHALLENGE

What is DREAM?

- Discussions for Reverse Engineering Assessment and Methods
- “The fundamental question for DREAM is simple: How can researchers assess how well they are describing the biological systems?”
Sponsors of DREAM

- Columbian University Center for Multiscale Analysis Genomic and Cellular Networks (MAGNet)
- NIH Roadmap Initiative
- IBM Computational Biology Center
- The New York Academy of Sciences
A Challenge is posted each year: An opportunity for “assessment”

- DREAM4(2009)- Organized by Gustavo Stolovitzky, IBM Computational Biology Center and Andrea Califano, Columbia University

- The PDZ challenge – Given the primary sequence of 5 proteins, almost identical, can you predict the binding target of each one. That is, what is the binding affinity difference of each ”mutant”.

The Challenge

- There is a binding pocket in each PDZ domain
- Mutations in the binding pocket result in a change in the binding affinity
- Given a mutation, can you predict the change in binding affinity?
- A single amino acid change can have "synergistic" consequences
Highlighted amino acids are attracted to different amino acids in the target protein.
Our Approach

1. Use the graph-theoretic models and corresponding graphical invariants of individual amino acids.

2. Replace each alphabet string representing the primary structure of each domain (and corresponding target) with a numerical string. The numerical string was determined by the observation of the properties of the amino acids.

3. Train a neural network using these numerical strings as inputs.
Conclusion

“The idea of optimization is intimately connected with modern science. Pioneers like Galileo, Fermat and Newton, were convinced that the world had been created by a benevolent God who had established the laws of nature as the most efficient way to achieve his purposes: in short, this is the best of all possible worlds and it is the task of science to find out why and how. Gradually this view was overturned, leaving optimization as an important tool for the human-engineered world.”

Taken from the abstract of Ivar Ekeland, University of British Columbia, for his upcoming Math Matters Lecture on The Best of All Possible Worlds: The Idea of Optimization.
Joint work with a number of people at ETSU

- Jeff Knisley- applied mathematics
- Teresa Haynes- graph theory
- Edith Seier-statistics
- Denise Koessler- Mathematics graduate student
- Cade Herron and Jordon Shipley- Mathematics undergraduate students

- Acknowledgement: Tamar Schlick and the web resource folks at NYU
References

References

References

