Projects in Bioinformatics (236524)

Ron Y. Pinter

pinter@cs.technion.ac.il, x4955
Or: Ron.Pinter@gmail.com

Spring 2014
Course Objectives and Method

• Develop and implement bioinformatic tools
• Focus on algorithms, systems, and usability (not application of existing tools)
• Outcome: working tool + report (oral and written)
• Work in pairs (not mandatory)
• Planning and tracking throughout the term
<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/3</td>
<td>Introductory lecture and presentation of topics for projects</td>
</tr>
<tr>
<td>11/3</td>
<td>Project group meetings – Decision on project topic</td>
</tr>
<tr>
<td>18/3</td>
<td>Project group meetings – Literature search results and general outline</td>
</tr>
<tr>
<td>6/4</td>
<td>Deadline for submission of detailed workplan</td>
</tr>
<tr>
<td>6/4</td>
<td>Project group meetings – Discussion and approval of workplan</td>
</tr>
<tr>
<td>2/5</td>
<td>Project group meetings – Update</td>
</tr>
<tr>
<td>9/5</td>
<td>Oral project presentations</td>
</tr>
<tr>
<td>16/5</td>
<td>Deadline for submission of final web report.</td>
</tr>
</tbody>
</table>
Suggested Projects

- Determining the directionality and organism of an interaction in a primary literature derived intercellular network of the immune system (Asst/Prof. Shai Shenorr)
- Coloring dotted interval graphs (with applications to DNA tagging)
- Qualitative comparison of GRN simulation profiles
- Efficient computation of Graphlet Degree Vectors (GDVs) and its application to protein-protein interaction networks
- Cytoscape plug-ins (apps for network search and alignment using partial information biological network queries)
- Select and implement an algorithm from a recent issue:
  - 1. The Journal of Computational Biology (JCB)
  - 2. The ACM/IEEE Transactions on Computational Biology and Bioinformatics (TCBB)
  - 3. Bioinformatics
Coloring of Dotted Interval Graphs (with Applications to DNA Tagging)

• Length polymorphism
  – The result of microsatellite repeats
  – Usually short (2, 3, 4 or 5 nucleotide) tandem repeats
  – Used for genetic mapping (with linkage analysis)

• Amplified by PCR, separated by *e.g.* electrophoresis (slab-gels)
  – Each genotype forms an “arithmetic series” of spots

• Each fragment can be tagged
  – This allows multiplexing
Tags are a critical resource - # must be minimized

Problem modeled as a graph coloring problem
- NP-hard in general, but there are families of graphs, e.g. interval graphs, for which it is polynomial-time solvable

The family of graphs that arise in the above tagging problem is an interesting generalization of interval graphs, but unfortunately it gives rise to arbitrary graphs making the coloring problem for it NP-hard as well
Still, there is hope to find specific approximation algorithms that would be much better than general approximation algorithms.

In this project we will:
- Explore interesting approximation schemes for the coloring problem of perforated graphs
- Apply developed algorithms to both synthetic and real bioinformatic data in order to evaluate them
Efficient computation of Graphlet Degree Vectors (GDVs)

- Characterizing individual nodes in a network by their neighborhoods [T. Milenkovic and N. Przulj, Cancer Informatics 2008]
- Leveraging operations on vector spaces + enrichment with GO terms
- Application to PPI in human, yeast, and E. coli
**GDV definition**

- **Graphlets** – a graphlet $g$ is a small (e.g. 2-4 nodes), connected graph pattern whose occurrence in a larger graph $G=(V,E)$ as a subgraph is a mapping $m$ of the nodes of $g$ to a subset of the nodes in $G$ such that if $(v_1,v_2)$ is an edge in $g$ then $(m(v_1),m(v_2))$ is an edge in $E$.

- **Motifs** – subgraphs that occur in a network at frequencies much higher than expected at random.
Occurrences of Graphlets
Using Profile Lengths as Discriminators