CSCI1950-Z
Computational Methods for Biology
Lecture 22

Ben Raphael
April 22, 2009
http://cs.brown.edu/courses/csci1950-z/

Outline

Network alignment and querying
  • PathBLAST
  • Color coding and randomized algorithms.
PathBLAST

- Goal: identify conserved pathways (chains)
- Idea: can be done efficiently by dynamic programming if networks are DAGs

Score: match + gap + mismatch + match


Why paths?
PathBLAST
(Kelley, et al. PNAS 2003)

- Find conserved pathways in protein interaction maps of two species
- Model & Scoring: (Whiteboard)
PathBLAST

- Problem: Networks are neither acyclic nor directed
- Solution: Randomize
  Impose random ordering on nodes, perform DP; repeat many times

- On average, highest scoring path preserved in $2/L!$ subgraphs
- Finds conserved paths of length $L$ within networks of size $n$ in $O(L!n)$ expected time
- Drawbacks
  - Computationally expensive
  - Restricts search to specific topology

PathBLAST: Computational Formulation

- **I = \{start vertices\}, e.g. receptors.**
- **Goal:** Find highest scoring paths \(I \rightarrow v\) for all \(v\) in \(G\).

---

PathBLAST: Computational Formulation

- **Given:**
  - Undirected weighted graph \(G = (V, E, w)\)
  - Set of start vertices \(I\), and end vertex \(v\)
- **Find:** a minimum-weight simple path \(P = (\nu_1, e_1, \nu_2, e_2, \ldots, e_{k-1}, \nu_k)\) starting in \(I\) and ending at \(v\):
  - \(\nu_1\) in \(I\) and \(\nu_k = v\).
- Recall: Simple path \(\nu_i \neq \nu_j\) if \(i \neq j\)
- NP-hard in general (reduction from TSP)
- Let \(w_i(\nu) = \text{weight of above}\).
  - Dynamic programming solution (whiteboard)

Scott, et al. JCB 2006
Color-coding
(Alon, Yuster, & Zwick)

• Assign each vertex random color between 1 and k.
• Colorful path: path w/ distinct colors.
• Colorful path $\rightarrow$ simple path.
• Goal: find colorful paths
  – Dynamic programming solution (whiteboard)
• High-scoring path not discovered when two vertices have same color.
• Repeat for many random colorings. (How many?)

Adding extra constraints

• Require a protein: assign it a unique color.
• Require a specific number of proteins from a set T: $W(v, S, c) = \text{min. weight of path ... (same as above) and contains exactly } c \text{ vertices in } T.$
• Order constraint on proteins in path
  – Membrane proteins $\rightarrow$ transcription factors.
Adding extra constraints

Rooted trees:
- Rooted at $v$.
- Every leaf is in $I$.

$$W(v, S) = \min \left\{ \min_{u : c(u) \in S \setminus \{c(v)\}} W(u, S \setminus \{c(v)\}) + w(u, v), \right.$$  
$$\left. \min_{(S_1, S_2) : S_1 \cap S_2 = \{c(v)\}, S_1 \cup S_2 = S} \min W(v, S_1) + W(v, S_2) \right\}$$

Color-coding
(Alon, Yuster, & Zwick)

- Extends to many other cases of subgraph isomorphism problem:
  - Does a graph $G$ have a subgraph isomorphic to graph $H$?
- $H =$ simple path of length $k$.
- $H =$ simple cycle of length $k$.
- $H =$ tree.
- $H =$ graph of fixed (bounded) tree-width
Additional Problems

1. Efficient querying of a network (e.g. QNET)
2. Find conserved subgraphs
   Heavy subgraphs in product graph
3. Multiple network alignment

Sources

