Graph similarities
(through the lens of bioinformatics)

Florian Sikora

LAMSADE, Université Paris Dauphine, CNRS – France
Outline

**Biological Networks**
- Network Query
- Network Alignment

**Gene Tree Reconstruction**
From sequences...to biological networks

Biological sequence comparison

- First protein sequences by Sanger, others
- Dayhoff, Jukes/Cantor
- Needleman/Wunsch
- PAM, BLOSUM matrix
- Swiss-Prot, GenBank, EMBL-Bank
- Stormo
- Haussler, Borodovsky, Churchill
- BLAST

A new type of data becomes routinely available
- Mathematical models of evolution
- Scoring via transition probabilities
- Public genome-scale databases
- Mining for motifs and domains
- Hidden Markov models

Interaction detection with two-hybrid mass spec.
- Interologs: evolutionary models
- Ogata/Kanehisa
- MaWish
- BIND, DIP, MINT, GRID
- Alon's network motifs

Biological network comparison

- Path BLAST
- Scale-free property; robustness
- Sharan/Karp/Ideker

[Sharan, Ideker, Nature Biotech]
What is a biological network

- PPI network
  - vertices ↔ proteins
  - edges ↔ interactions
- Metabolic network, gene regulation networks...
What to look for?

Three main problems for network comparison:

- Network Query
- Network Alignment
- Network Integration
Outline

Biological Networks
   Network Query
   Network Alignment

Gene Tree Reconstruction
Network Query

Topological view:
- find a small graph in a big graph
- ⇒ subgraph isomorphism problems
Network Query

- **Topological view:**
  - find a small graph in a big graph
  - $\Rightarrow$ subgraph isomorphism problems

- **Functional view:**
  - topology is less important
  - **functionalities** of network vertices $\rightarrow$ governing principle
Topology based - ISI/MCIS definition
Topology based - ISI/MCIS definition
Known

- Clique as a special case (W[1]-hard).
- FPT in H-minor free graphs, planar graphs, bounded degree graphs.
- Remains NP-hard on restricted graph classes incl. forests.
  - Not in XP for treewidth/fvs.
  - Look at bigger structural parameters!
Parameter sum of the VCs

- $O^*(k^k)$ algorithm [Abu Khzam 2014].
  - No polynomial kernel (cross-composition) [ABS 2017]
  - No $c^k$ algorithm under ETH [ABS 2017]
    - Very rare for this parameter.

→ very hard
Parameter sum of the VCs

- $O^*(k^k)$ algorithm [Abu Khzam 2014].
  - No polynomial kernel (cross-composition) [ABS 2017]
  - No $c^k$ algorithm under ETH [ABS 2017]
    - Very rare for this parameter.

→ very hard

- What about FVS+VC?
- Cutwidth, rankwidth, modular-width...?
- Dropping topology requirement?
Functionally based - Topology-free

- Functional view:
  - topology is less important
  - functionalities of network vertices $\rightarrow$ governing principle

- motif whose topology is not completely known
- noisy networks (missing connections)
- query between well and poorly annotated species
Functionnaly based - Topology-free

- **Functional view:**
  - topology is less important
  - functionalities of network vertices $\rightarrow$ governing principle

- motif whose topology is not completely known
- noisy networks (missing connections)
- query between well and poorly annotated species

- function $\leftrightarrow$ color
- $\Rightarrow$ graph is vertex-colored (but *not properly*)

- motif (query): *multiset* of colors
Functionally based - Topology-free

- **Functional view:**
  - topology is less important
  - **functionalities** of network vertices $\rightarrow$ governing principle

- motif whose topology is not completely known
- noisy networks (missing connections)
- query between well and poorly annotated species

- function $\leftrightarrow$ color
- $\Rightarrow$ graph is vertex-colored (but **not properly**)!

- motif (query): **multiset** of colors
- motif **occurs** (and thus “accepted”) if **connected** in graph
Find a connected subgraph colored by

\[ M = \{ \textcolor{yellow}{\circ}, \textcolor{red}{\circ}, \textcolor{blue}{\circ}, \textcolor{black}{\circ}, \textcolor{green}{\circ}, \textcolor{magenta}{\circ}, \textcolor{cyan}{\circ}, \textcolor{orange}{\circ}, \textcolor{olive}{\circ}, \textcolor{lightblue}{\circ}, \textcolor{gray}{\circ}, \textcolor{brown}{\circ} \}. \]
Find a connected subgraph colored by $M = \{\, \bullet, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \circ, \bullet\, \}$. 
Find a connected subgraph colored by

\[ M = \{ \color{orange}, \color{red}, \color{pink}, \color{blue}, \color{gray}, \color{green}, \color{cyan}, \color{black}, \color{magenta}, \color{purple}, \color{yellow}, \color{gray} \} \].
Known

- LOT of studies.
- Parameterized complexity
  - Many different parameters (natural, colors, dual, structural,...)
  - Lower bounds
- Many optimization variants for approximation
- Many variants (connectivity, list of colors...)
Outline

Biological Networks
   Network Query
   Network Alignment

Gene Tree Reconstruction
Multiple network comparison

[Sharan, Ideker, Nature Biotech]

- How to extend to more than few networks?
Graph Dictionary Learning

Graph dataset

Dictionary of subgraphs

Vector representation of the graphs

Graph coding on the dictionary

Graph dictionary learning

PEPS 2017
Outline

Biological Networks
  Network Query
  Network Alignment

Gene Tree Reconstruction
Genes

Chromosome \((10^7 - 10^{10} \text{ bp})\)

Gene \((10^3 - 10^6 \text{ bp})\)

Function
History of a family of homologous genes

Represented with a **gene tree**.

Two genes are

- **Orthologs**: lowest common ancestor is a **speciation** event
- **Paralogs**: lowest common ancestor is a **duplication** event
Tree-free methods: Relation graph

Given a set $R$ of relations (sequence similarity, functional annotation...)
Build relation graph $G_R = (V_R, E_{red}, E_{blue})$

- Orthologs: $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, $\{C, D\}$
- Paralogs: $\{A, D\}$, $\{B, D\}$
From one to the other

Easy
From one to the other

How?
Satisfiability and Consistency

A set of orthology/paralogy relations is defined as:

- **Satisfiable**: exists a tree that displays the set of relations

- **Consistent**: satisfiable and agrees with a given tree
Satisfiability and Consistency

A set of orthology/paralogy relations is defined as:

- **Satisfiable**: exists a tree that displays the set of relations
  - Iff blue (orthology) edges induce a \( P_4 \)-free graph (cograph)  
    (Checkable in poly-time) [Hellmuth et al. 2013]

- **Consistent**: satisfiable and agrees with a given tree
Satisfiability and Consistency

A set of orthology/paralogy relations is defined as:

- **Satisfiable**: exists a tree that displays the set of relations
  - Iff blue (orthology) edges induce a $P_4$-free graph (cograph) (Checkable in poly-time) [Hellmuth et al. 2013]

- **Consistent**: satisfiable and agrees with a given tree
  - Iff satisfiable and $P_3$'s are in agreement with $S$ (Checkable in poly-time) [Lafond et al. 2014]
Correction of Orthology and Paralogy Relations

- If it is not satisfiable (consistent): **correct in a minimal way**
- Minimize the number of **relation corrections** to get a satisfiable (consistent) set of relations
Correction of Orthology and Paralogy Relations

- If it is not satisfiable (consistent): correct in a minimal way
- Minimize the number of relation corrections to get a satisfiable (consistent) set of relations
  → Minimum number of edges to edit to have a cograph (...).
Correction of Orthology and Paralogy Relations

- If it is not satisfiable (consistent): correct in a minimal way
- Minimize the number of relation corrections to get a satisfiable (consistent) set of relations

→ Minimum number of edges to edit to have a cograph (...).
Correction of Orthology and Paralogy Relations

- If it is not satisfiable (consistent): correct in a minimal way
- Minimize the number of relation corrections to get a satisfiable (consistent) set of relations
  → Minimum number of edges to edit to have a cograph (...).

![Diagram](image-url)
Correction of Orthology and Paralogy Relations

- If it is not satisfiable (consistent): correct in a minimal way
- Minimize the number of relation corrections to get a satisfiable (consistent) set of relations

→ Minimum number of edges to edit to have a cograph (…).
Satisfiability

- **NP-hard** [El-mallah et al. 1988]
- Solvable in $O^*(4.612^k)$ [Liu et al. 2012]
- $O(k^3)$ kernel [Guillemot et al. 2013]
Satisfiability

- NP-hard [El-mallah et al. 1988]
- Solvable in $O^*(4.612^k)$ [Liu et al. 2012]
- $O(k^3)$ kernel [Guillemot et al. 2013]
- 3Δ-approx [Dondi et al. 2017]
- No constant approx for the weighted variant under UGC [Dondi et al. 2016]
- **Maximization** variant admits a PTAS [Dondi et al. 2017]

Belongs to a larger class of problems ($P_\ell$-free edition/deletion...)
Satisfiability

- **NP-hard** [El-mallah et al. 1988]
- Solvable in $O^*(4.612^k)$ [Liu et al. 2012]
- $O(k^3)$ kernel [Guillemot et al. 2013]
- $3\Delta$-approx [Dondi et al. 2017]
- No constant approx for the **weighted** variant under UGC [Dondi et al. 2016]
- **Maximization** variant admits a **PTAS** [Dondi et al. 2017]

Belongs to a larger class of problems ($P_\ell$-free edition/deletion...)
- Lower bound FPT?
- Constant approx ?
- Synthetic and real data?
Thanks!