A summary of the week

Leo van Iersel
(CWI)

What do biologists want?

What do biologists want?

- Biologists don't know what they want
- Depends on data and goals
- Changes all the time

Two Types of Networks

- Affinity networks
- Data-display networks
- Implicit networks
- Unrooted networks
- Split networks
- Genealogical networks
- Reticologenies (inference networks)
- Explicit networks
- Rooted networks

We have this ----------> **towards** this

Two extremes

- Not a perfect dichotomy
  - Rooted data-display networks
  - Median networks are inference networks (not pure data-display)
- Biologists use split-networks (Neighbour Net) and augmented tree
- Biologists often incorrectly interpret networks
  - Interpret data-display as genealogical network

Networks in practice

- Biologists use networks to display **complexity, uncertaincy** and **treelikeness**
- … not to display genealogy
- There is a demand for networks that are ``in the middle''
- Difficult to get information out of complex networks

An important distinction

- Networks to quantify complexity (treelikeness)
  - Don't simplify network (obviously)
- Networks to answer (other) questions
  - Might want to simplify network
  - To see more structure

What do biologists want?

Biologists want to play

- Networks should show lots of information
  - How are gene trees embedded in the network?
  - Confidence values
  - Branch lengths
  - What causes the incongruence?
  - All solutions (and which data gives which solution)
- Range of tools
- Range of options
- Easy to add, remove data (e.g. Taxa)

Some more options

- Specify portion of genome you're interested it
- Different restrictions to network space (biologist can choose)
- Representing output networks
  - Cluster
  - Sort
  - Summarize

Biologists want a statistical framework

- Confidence values
- Maximum Likelihood
- Bayesian
- Hypothesis testing

Modelling

- Take many/all possible causes of incongruence into account
  - Incomplete lineage sorting
  - Duplication and loss
  - Reticulation
  - Stochastic reasons/randomness
  - Functional reasons/selection
  - Incorrect rooting, other errors

Modelling (2)

- Specialise
  - Recombination
  - Hybridization
  - LGT
- Complex models that are tractable for small numbers of taxa
- Mixture models
  - Substitutions, indels and reticulation

General inputs

- Nonbinary trees
- Multiple (weighted) trees
- Partial trees
- Sequences

Software

- Easy-to-use
- Fast
  - (Meta)Heuristics
- Many tools in one program
- Tools all connected
- Interactive
  - Clickable
  - Zooming
  - Adding/removing
- Biologist in charge

Validation

- Networks need to be validated
- Network methods need to be validated
- Datasets needed where (properties of) real genealogy are known
  - For each kind of data
- Validate networks also on tree-like data
- Put links on blog (to data+paper)

Split networks

- More data, more messy
  - Hairball networks
  - Problem with whole genomes
  - Filtering/preprocessing
- Biologists over-interpret networks
- Neighbor-Net gives restricted kind of network

Important Computational Problems

- Maximum Likelihood/Parsimony
  - Compute score (fast!)
  - Search network space (local operations)
- Deal with whole genomes
  - Networks from gene-order data
- Generalize everything to
  - Nonbinary trees
  - Partial trees
  - Multiple trees

More Questions for Mathematicians

- Detect if backbone (tree) structure is supported
- Detect fictional branches
- Link foodwebs to networks
- $O(c^n)$ algorithm for minimum reticulation ARG
- ...and for hybridization number of multiple trees
- Is the most probable clade always in the species tree?
- Measures for treelikeness (incl. branch lengths)

Tree of Life

- There exists a tree of life
  - Question is, how resolved is it
- (At least) for prokaryotes there is (almost) no tree-like signal
- Tree-thinking should be replaced by **network thinking**

Networks often more suitable for biologists than trees

Biologists don't know networks

- Need textbooks, workshops, courses
- User-friendly software
- Publications showing usefulness

Transdisciplinary

- Computationalists learn about biology
- “Touch” data
- Become geneticists

Interdisciplinary

- Close collaborations
- Publications showing usefulness

Mathematicians should provide

- What is method trying to reveal/display?
- What are its limitations?
- What is rigorous test?
- What form of data does it need?

Biologists should provide

- Data
- Interpretation of results
- Design of test data

Both sides need to boil down to the essence

For biologists to keep in mind

- First step is exploratory data analysis
  - Search for unexpected patterns (errors)
- Filtering techiques
  - To make sure noise is not expressed by network
- Do not interpret affinity networks as genealogies

Protocol

- Protocol useful for non-specialists wanting to construct networks
- Should be suggested practice
  - Specialists can deviate
- Can potentially prevent missuse of network methods

Some conclusions

- Network methods give hints, not conclusions (something to try / think about for biologist; not necessarily a perfect genealogy)
- Network methods should provide sufficient information for biologist to answer questions
- Biologist in charge, can play with program
- Different tools for different data / questions
- Easy-to-use sofware that works for all data
- Heuristics
- Validation

Phynet Blog

- http://phylonetworks.blogspot.com/
- Please contribute
  - links to validated datasets
  - continue discussions

Thank you all for your active participation!