BIG4
Biosystematics, Informatics, Genomics

BIG4 field workshop

June 5-11 2016, Havraníky, Czech Republic

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

BIG4 Field worklshop, June 5-11 2016, Havraníky, Czech Republic

BOLD algorithm

Viktor Senderov (@vsenderov)
Bulgarian Academy of Sciences/ Pensoft, Sofia Bulgaria
Advisor: Prof. L. Penev

PhD Financed through the EU Marie-Sklodovska-Curie Program
Grant Agreement Nr. **642241**

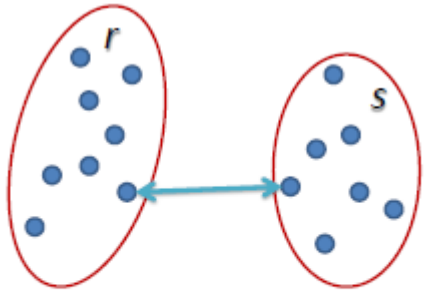
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

BOLD Algorithm

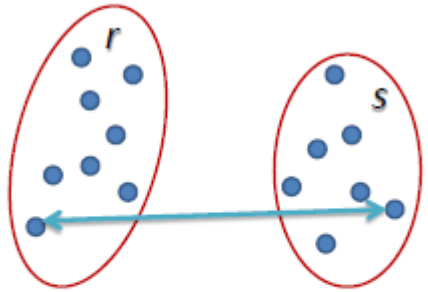
- Cythocrhome oxidase I (**COI**) gene : 648 bp
  - **More than 95%** of animal species posses a diagnostic COI array
  - **COI divergence** rarely exceeds 2% within a named species, while members of different species typically show higher divergence
- RESL algorithm : "**Refined Single Linkage Analysis**"
  - jMotu, ABHD, CROP, GMYC
  - Design of RESL was driven by the need to create **a fast algorithm** (1.8 Mio barcode sequences as of 2013, 10 000 new each week)

Steps of RESL

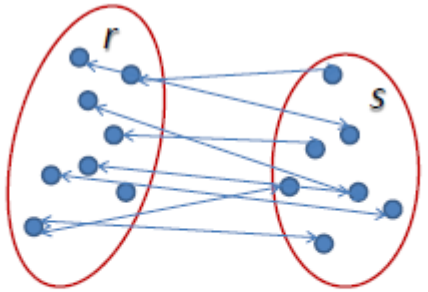
1. Alignment
2. Single Linkage clustering (t = 2.2 %)



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

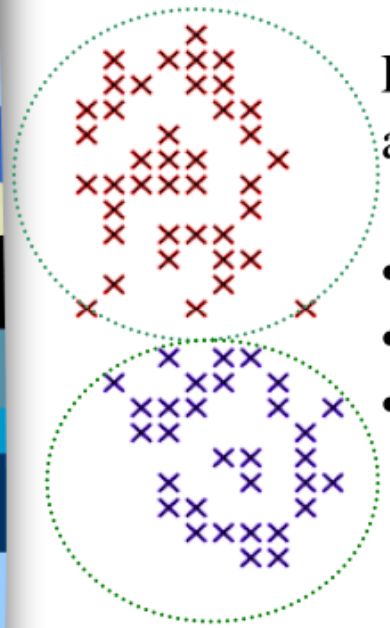
Single linkage	Complete linkage	Average linkage

Cluster refinement

- Markov clustering (https://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf)
  - Clusters whose members show high sequence variation but lack discontinuity remain fixed
  - Cluster whose show sequence variation with clear internal patitions are assigned to different OTUs even if their separation is less than 2.2 %

Graph Clustering

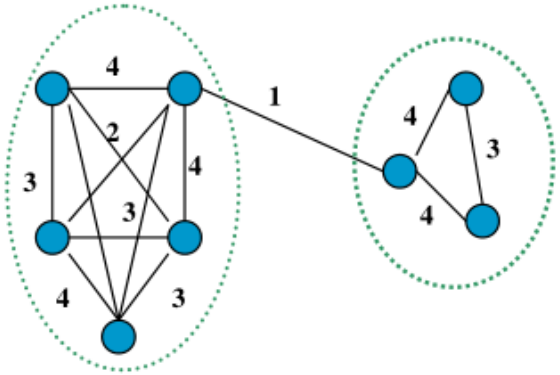
- Clustering – finding natural groupings of items.
- Vector Clustering



Each point has a vector, i.e.

- x coordinate
- y coordinate
- color

Graph Clustering



Each vertex is connected to others by (weighted or unweighted) edges.

A random walk in G that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited.

MCL

- Random walk from sequence to sequence (Graph Clustering)
  - **Expansion** increases traffic between nodes
  - **Inflation** raises the probability of walks within highly connected regions