Exploring environmental genetic diversity with similarity networks

Philippe Lopez
UMR CNRS 7138 Evolution Paris Seine
Université Pierre et Marie Curie – Paris, France
Why look at environmental data?

Great Plate Anomaly

Cultivable organisms
- Genomics
  - Complete genomes
  - $10^3$ to $10^4$ genes
  - $<1\%$ of diversity

Communities
- Metagenomics
  - Short fragments
  - $10^6$ to $10^8$ reads
  - $>99\%$ of diversity

Is environmental diversity just ‘more of the same’?
Environment is a reservoir of genetic novelty

Molecular evolutionists have to **describe** and **structure** this massive genetic novelty:
- New gene variants
- New organisms
Why alternative approaches are needed for such data

- Typically millions of sequences for one sample

- Assessment of homology can be difficult:
  - Extreme divergence
  - Introgressive processes

Can the environmental diversity be described and structured by an alternative approach that is faster and more inclusive than phylogenetic trees/networks?
Exploring a new datatype: sequence similarity networks

- Node: individual sequences
- Edge: connects two nodes that show some significant shared property

Sequence similarity networks are a mapping of sequences resemblances for given rules.

- Efficient way of displaying and analyzing diversity
- Usually fast assessment of similarity
- Quadratic complexity

Can easily accommodate millions of objects.
Transitivity of sequence similarity

Full length similarity

• Allows the detection of distant homologies

Partial similarity

• Allows the detection of composite objects

Slow detectable divergence

• Expected pattern for phylogenetic markers

Extreme divergence

Gene fusion, partial transfer, etc.
Sequence similarity networks...

Sequence $i$ | AAATTCGTAGG
Sequence $j$ | CAAATTCATA

$(\text{minBLAST } 1^{-20}, >30\% \text{ identity})$

**Connected components** naturally define extended **gene families**

Translation Initiation Factor I

Restriction endonuclease subunit S
... are a much more inclusive datatype

**Sequences amenable to phylogeny**

- **Vizualisation** of diversity
- Includes **all the data**
- Complex, thus **information** rich, topologies

*Beauregard et al., 2011, Biol Dir*
Networks can be mathematically described by graph theory

**New datatype** in which to look for **regularities** and **singularities**

**Connected components:**
- Diameter
- Clustering coefficient
- Minimum spanning tree

**Nodes:**
- Degree
- Closeness, betweenness
- Articulation points

**Groups of nodes (colored by function, lifestyle, taxonomy,...):**
- Modularity
- Conductance, assortativity

These measures allow a fine **description** of the networks and thus potentially **rich comparisons**
Similarity networks are not phylogenetics, but...

Sequence similarity networks are a mapping of sequences resemblances for given rules. If sequences evolved along a tree, node clusters correspond to clades.
Similarity networks and environmental data

1) Visualizing and structuring diversity
   Ecological study of marine ciliates

2) Screening for gene families of interest
   Mobilization in the human gut microbiome

3) Screening for highly divergent sequences
   Environmental variants of Domain signature sequences
Structuration of marine ciliates populations

Marine ciliate diversity was assessed by comparing the V4 region of SSU-rDNA

- **16,911** unique sequences from BioMarks environmental data
- **928** unique sequences from previous environmental studies
- **308** sequences from cultivable ciliates

Is the corresponding sequence similarity network **structured** and how so?

Forster et al., 2015, BMC Bio
Building and structuring a sequence similarity network

Identifying and pooling densely connected groups of nodes help simplify the network.
Sequence similarity networks as an efficient way of visualizing diversity

Nodes (sequences), and clusters of nodes, can be colored according to various attributes. Measures from graph theory help analyze the resulting topology.
Environmental ciliate diversity is largely underestimated

Closeness of a node $x$ is given by $C(x) = \frac{1}{\sum_y d(x,y)}$, i.e. lower for peripheral nodes.

Recent environmental sequences show significantly lower closeness centralities.

Environmental studies have revealed a substantial amount of novel diversity.
**Assortativity measures**

High assortativity \( r \sim 1 \)

Low assortativity \( r \sim 0 \)

\[
E = \begin{bmatrix}
GG & RG \\
GR & RR
\end{bmatrix}
\]

GG = proportion of green-green edges
RR = proportion of red-red edges
GR = RG = proportion of green-red edges
GG + RR + GR + RG = 1

\[
r = \frac{Tr E - \|E^2\|}{1 - \|E^2\|}
\]

\( \|E\| = \text{sum of } E \text{ components} \)

r = 1 for perfect assortativity
r = 0 for random assortativity
r < 0 for dissortativity

**Significance** can be assessed by shuffling colors on the same network

Newman *et al.*, 2002
Ciliate diversity is strongly structured

Each of the 3 habitats show significant assortativity
About half of the 8 investigated locations show significant assortativity

Ciliates are thus not globally dispersed but structured by habitat and geographical location

SSN are an efficient and extremely scalable alternative for rRNA diversity surveys

Forster et al., 2015, BMC Bio
Mobilization in the human gut microbiome

Large microbial communities associated to human body:

- **Large** gene pools
- Potentially **extensive** gene flow

Can networks help us study the mobilization of gene families in microbiomes?

- **311,265** ORFs from **13 Japanese** individuals gut microbiomes (Kurokawa et al., 2007)
- **195,521** ORFs from **18 North-Amercan** individuals gut microbiome (Turnbaugh et al., 2009)
- All NCBI **mobile genetic elements** sequences (viruses, plasmids, integrons)

Building of a sequence similarity network with **748,688** sequences

_Bicep et al. (in prep)_
The network of 31 human gut microbiomes

The human microbiome gene network
499,869 seqs, BLAST score < 1e-5;
False BBH; > 20% identity

Bicep et al. (in prep)
Assessing the potential mobility of gene families

Working hypothesis:
Gut microbiome sequences similar to those found in mobile genetic elements are potentially mobilizable

>1 type of MGEs

VIRUS only

PLASMID only

21,525 gene families

- 13,259 non mobilized
- 7,468 potentially mobilized
- 798 potentially very mobilized

>33% of large gene families from the gut microbiome could be mobilized

21,525 large enough connected components (#ORFS>4)

Bicep et al. (in prep)
Gut microbiome mobile genes are functionally biased

Hypergeometric test, Bonferroni correction, p-value < 0.01

(J) Translation; (K) Transcription; (L) Replication and repair; (V) Defense mechanisms; (C) Energy production and conversion; (E) Amino Acid metabolism and transport; (F) Nucleotide metabolism and transport; (G) Carbohydrate metabolism and transport; (H) Coenzyme metabolism; (P) Inorganic ion transport and metabolism; (R) General Functional Prediction only; (S) Function Unknown.
Mobile genes are widespread both in microbes and humans

# of microbial host genera (assessed by MGRAST)

# of microbial host phyla (assessed by MGRAST)

Mann Whitney Wilcoxon test, $\alpha = 0.01$

*Bicep et al. (in prep)*
Exploring the functioning of large microbial communities

Screening of a very large sequence dataset:

- ‘useful’ functions are preferentially mobilized
- gene flow within but also between microbiomes

Further developments:

- **Comparison** of the various individuals microbiomes: coloring the nodes by individuals and exploiting their topological relationships
- Very **scalable** protocol: investigating more recent (and thus much larger) datasets
Screening environmental data for genetic variants of interest

• Sequence similarity networks can accommodate **extreme divergence**

Reference → Variant

High match cover

• **Selecting a reference**: connected components with a strong **Archaea/Bacteria signal**

Genomic dataset:
- 560,000 sequences
- 54 Archaea
- 70 Bacteria
- 8 Eukaryotes

Extended phylogenetic coverage
Selecting relevant reference gene families

- **Conductance** of a group of nodes $X$ (Leskovec, 2008):

$$C(X) = \frac{s}{s+2e} \quad \text{with} \quad s = \text{number of edges between } X \text{ and non}(X) \text{ nodes}$$

$$e = \text{number of edges between } X \text{ nodes}$$

- **86** components with low bacterial AND archaeal conductance
  - NUCLEI

- 0.2% of gene families
- Wide functional diversity
- 61% average inter Domain identity

Where would **environmental** sequences fall in that network?
Looking for homologs in environment

Metagenomic dataset:
- 236 microbial samples
- > 9,400,000 non redundant predicted ORFs (>50AA)

Nuclei sequences
- 1st BLASTP round
- 10,822 NUCLEI (cultivable) sequences

Environmental sequences
- 1st BLASTP round
- 131,162 environmental sequences:
  - 85% human microbiome
  - 15% aquatic environments
Environmental homologs and NUCLEI

Edge creation rules:

- BLAST score > $10^{-5}$
- Identity > 30%
- Match > 80% of shortest sequence (sequences of comparable size)

Two unconnected nodes (sequences) are unalignable
Proof of concept

At least two types of divergent and highly divergent Maltose ABC Transporters in the environment

What’s the divergence between these sequences and those of cultured organisms?
To whom do these divergent sequences belong?

1st and 2nd sequences are compared against the NCBI nr database using BLASTP.

Max identity to nr > 60%
Max identity to nr < 60%

Empirical average identity is 61%.
Environmental data indeed contain a large part of variants

A large part of environmental potential homologs of NUCLEI sequences show a very low similarity to what’s already known.

Human gut microbiome seem to be better known than other environments.
Back to the Maltose ABC Transporters

- Env. sequences distance 1
- Env. sequences distance 2
- Archaea
- Bacteria
- Max identity to Reference > 60%
- Max identity to Reference < 60%
Back to the Maltose ABC Transporters

- Highly divergent
- Related to genomic sequences
- Clustering outside Domains

New Domain(s) ?

Env. sequences distance 1
Env. sequences distance 2
Archaea
Bacteria
Max identity to Reference > 60%
Max identity to Reference < 60%
Some compelling examples

DUF167

Cobalamine Phosphate Synthase

- **Archaea**
- **Bacteria**
- **Eukaryotes**

- Grey: Max identity to Reference > 60%
- Orange: Max identity to Reference < 60%
Some compelling examples

Metalloendoprotease

Ribosomal Protein RPL 23/25

- **Archaea**
- **Bacteria**
- **Eukaryotes**

- **Orange** Max identity to Reference > 60%
- **Gray** Max identity to Reference < 60%
In-depth analysis of variants

- Most cliques of environmental ‘unknown’ sequences seem to be under selection
- When they can be aligned, some of these show very interesting phylogenetic positions
What are these sequences?

- **Metagenomic sequencing errors / frameshifts?**
  - 80% covering constraint
  - Clusters in networks

- **Viral sequences?**
  - Only ‘microbial’ metagenomes
  - NCBI nr database already contains viruses
  - If so, where do they come from?
  - Explanation for similarity AND isolation?

- **Sequences from genuine cellular organisms?**
  - Let’s look for them!
Environmental studies with similarity networks

Similarity networks are powerful exploratory tools:
• they help visualize the diversity of large datasets
• they are amenable to various mathematical treatments

Two main advantages:
• Fast: especially suited for large datasets (like metagenomics)
• Inclusive: making use of all the data
• New datatype, providing new type of evidence
• Complementary to other approaches (phylogenetic trees and networks)

Plenty of other applications in evolutionary studies
Try similarity networks!
Many thanks to my excellent coworkers

- Eric Bapteste
- Eduardo Corel
- Cédric Bicep
- Pierre-Alain Jachiet
- Raphaël Meheust
- Jananan Pathmanathan
- Chloé Vigliotti
- Sukithar Rajan

Many thanks to the organizers!
Thank you!
Similarity networks built on metagenomics reads

Some connected components exhibit the standard ‘laminar’ topology
Similarity networks built on metagenomics reads

Others exhibit much more complex topology, especially large cycles (repetitions, insertions ?)
Mobile genes show more diversity...

**Clustering coefficient** of a connected component:
Number of edges / max number of edges

Low clustering coefficient = more diversified component?

---

- **Clustering coefficient**
  - Number of edges / max number of edges

- **Mobile genes**
  - Non mobilized
  - Potentially mobilized
  - Potentially very mobilized

---

*Bicep et al. (in prep)*
Ces processus non divergents sont dits introgressifs

“In vertical descent, the genetic material of a particular evolutionary unit is propagated by replication inside its own lineage. In introgressive descent, the genetic material of a particular evolutionary unit propagates into different host structures and is replicated within these host structures. “

Bapteste et al., 2012, PNAS

Les processus introgressifs, en combinant du matériel génétique provenant de sources distinctes, affectent les objets biologiques à tous les niveaux d’organisation

Recombinaison

Transfert

(Endo)symbiose

Séquence A

Séquence composite

Génome A

Superorganisme

Séquence B

Séquence composite

Génome A

Génome B

Génome A

Génome A

Séquence B

Séquence B

Apparition d’objets composites à plusieurs niveaux
Des topologies complexes en cas de données complexes

Nœuds : séquences
Ressemblance : score BLAST
Composante connexe du réseau de reads

> 90 % ID
> 80% cover
Composante connexe du réseau de reads

Présentation du réseau :
- Nb de CC : 238,243
- Nb de noeuds du réseau : 5,605,140
- Nb d'arêtes du réseau : 707,399,890
- Nb de noeuds dans cette CC : 650
- Nb d'arêtes dans cette CC : 4720

- DNA-directed RNA polymerase
- Unknown function
- Phospholipid Binding Protein
- Cation Transport ATPase
- NADH : ubiquinone oxidoreductase subunit 4
- 3-oxoacyl
A fast and inclusive description of diversity

Can these new datatypes be useful for evolutionary studies?
L’histoire évolutive du vivant est donc sans doute bien plus complexe qu’un arbre

Multitude d’objets, de types d’objets et de processus évolutifs

Comment décrire, structurer et expliquer une telle diversité ?
Mobile genes are under the same selection

Sequences aligned and treed with PhyML; dN/dS ratio estimated with PAML

Bicep et al. (in prep)