Apprehending Life’s complexity and its evolution with networks

Eric Bapteste
A tree metaphor provides a genealogical framework to study evolution.
Yet, a genealogical tree cannot display all evolutionary relationships.
Consider the origins of an obvious merger: yourself.
Introgression is a general process, affecting biological objects at many levels of organisation.

Many adaptations have their origins outside rather than within vertical lineages.

Bapteste E. Frontiers in Micro. 2014

Bapteste E. Les gènes voyageurs. Belin 2013
Moreover, not all evolving objects are genealogically related.

One genealogical tree cannot represent all the evolutionary history.
Introgression results in phenomena, resisting classic phylogenomics

- Microbial social life
- Chimerism & major evolutionary transitions
- Holobionts
**Take-home message**

- With the increasing realization of the quantitative and qualitative importance of reticulate processes in evolution, network-based metaphors and methods are improving, inviting evolutionary biology to experience a network-thinking era.

- Sequence similarity networks, genome networks and bipartite graphs **now** provide tools of choice to tackle the evolution of complex phenomena.
1. Microbial social life raises numerous challenges

The genome a ‘read-write’ storage organelle rather than a ‘read-only’ memory
How to model genetic transmission in prokaryotic communities?

- There is **more than one channel** for gene transmission, i.e. many routes through which genes pass from one structure to another structure.

- Is gene transmission random in terms of targets?

- Is it random in terms of what gets transmitted?

- Can we find groups of co-transmitted genes?
Introducing shared genes networks

- Introducing **new metaphors** and **new tools** in evolutionary biology
- **Simultaneous** analysis of **mobile genetic elements and cells evolution**
Genome networks: connecting genomes from entities sharing genes

Sharing between virus & plasmids

Sharing between virus

Sharing between Bacteria & virus

plasmids, virus, bacteria

Halary et al. PNAS 2010
Genetic worlds suggest isolated transmission groups among microbes

plasmids, virus, bacterial chromosomes

$10^{-20}; 20 \%$ identité, 578 527 séquences, 111 génomes
Genome network representations encourage different questions

- **What barriers** to transmission?
- **What partnerships**?
  (clubs of genomes with common genetic goods)

---

While powerful, a genome network requires further information (for example on its edges) to address these questions.

Bapteste et al., PNAS 2012.
While powerful, a genome network requires further information (for example on its edges) to address these questions.

Bapteste et al., PNAS 2012.
A multiplex approach can be used to analyze the mobilome network.

*Complex interconnected mobilome of the giant viruses.*

Yutin et al. Virology Journal 2013
Yet other graphs are also fruitful for larger datasets

Our dataset (approximately 1.6 million sequences):
complete genomes for **152 archaea + 230 bacteria** (one genome per family, except for cyanobacteria, where 16 genomes were kept) + **4219 plasmids + 3613 viruses**
Application of bipartite graphs to study microbial social life

- Multilevel
Application of bipartite graphs to study microbial social life

• Accurate
Application of **bipartite graphs** to study microbial social life

- Accurate
From genome networks to bipartite graphs: a finer grained-view

- Inclusive

- (archaea, bacteria, virus, plasmids)

- Genomes

- Genes Families

- Inclusive

- Eases the identification of clubs of genomes sharing genes or co-transmitted genes
The partition of such graph is informative
Grooming the bipartite moustache...

- a gene family shared by many genomes otherwise with totally distinct gene contents

- Several gene families exclusively shared by the same sets of genomes
Informative patterns in bipartite graphs

• Articulation points

- May identify rather public genetic goods, e.g. similar genes found in host genomes from almost different sets

>= 80% mutual cover
>= 90% identity
Informative patterns in bipartite graphs

*Ruminococcus bromii*
Firmicutes
Gram +
Mesophile
Rumen
Obligate anaerobe

*Fibrobacter succinogenes*
Fibrobacter/Acidobacteria
Gram -
Mesophile
Rumen
Anaerobe

3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase

When articulation points are enriched in genes from specific functional categories, transmission is biased in terms of what gets transmitted.
Traits impacted by microbial symbionts with bipartite graphs?

Your favorite microbiomes

Hosts with remarkable phenotype A

Host with another phenotype

Twin

(Gene from) Bacteria (or MGE) i

(Gene from) Bacteria (or MGE) j
Joint traits in holobionts offer an alternative explanans for diversity

Brucker & Bordenstein, TREE 2012

- Species specific property?
- Individual specific property?
- Joint phenotypes?

See Eduardo Corel’s and Phil Lopez’s talks
2. The **composite nature** of things raises numerous challenges

How to better understand:

- Traits with multiple origins / joint phenotypes?
- Rules of (genetic) combination?
- Saltatory evolution?
What happened after distinct genetic materials associated into a new host?

- Competition between genes?
- Cooperation between genes?

Do different components evolve alike in their composite host?
Introducing sequence similarity networks (SSN)

- complex data: a plurality of processes and evolving objects

Weak divergence

High divergence

Introgression

Fast, flexible => exponential increase of data
SSN: a simple representation of evolutionary processes and patterns

<table>
<thead>
<tr>
<th>Nodes (individual sequences)</th>
<th>Edges (weighted) (connect 2 nodes with significant similarity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
<tr>
<td>1  2  3  4</td>
<td>9  10 11 12 13 14</td>
</tr>
</tbody>
</table>

- **Nodes**: Individual sequences
- **Edges**: Weighted edges that connect 2 nodes with significant similarity
Graph theory allows automated analyses of such networks.

Connection rule: minBLAST $10^{-20}$, >30% ID

- Bacteria
- Archaea
- Mobile Element

- Connected components:
  - Diameter
  - Clustering coefficient
  - Minimal spanning tree

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

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

Translation initiation factor *SUI1*  
Type 1 Restriction-Modification nuclease
Application to a large dataset: 38 protists + 382 prokaryotes

- *Cyanidioschyzon merolae strain 10D*
- *Galdieria sulphuraria*
- *Porphyridium purpureum*

**Protists**:
- *Micromonas pusilla*
- *Ostreococcus tauri*
- *Chlamydomonas reinhardtii*
- *Chlorella variabilis*
- *Coccomyxa subellipsoidea*
- *Volvox carteri f. nagariensis*
- *Cyanophora paradoxa*
- *Cyanidioschyzon merolae strain 10D*
- *Galdieria sulphuraria*
- *Porphyridium purpureum*

**Prokaryotes**:
- *Bodo saltans*
- *Naegleria gruberi*
- *Trypanosoma brucei*
- *Allomyces macrogynus*
- *Spizellomyces punctatus*
- *Mortierella verticillata*
- *Saccharomyces cerevisiae*
- *Sphaeroforma arctica*
- *Salpingoeca rosetta*
- *Capsaspora owczarzaki*
- *Monosiga brevicollis*
- *Thecamonas trahens*

**Species**:
- *Symbiodinium minutum*
- *Tetrahymena thermophila*
- *Paramecium tetraurelia*
- *Ichthyophthirius multifiliis*

**Other**:
- *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, *Nannochloropsis gadiata*, *Pseudo-nitzschia multiseries*, *Aureococcus anophagefferens*

**Phylum**:
- *Polysphondylium pallidum*
- *Dictyostelium discoideum*
- *Acanthamoeba castellanii*
Prokaryotic components within eukaryotic genomes evolve differently

In sequence similarity networks, **eukaryotic genes** with similarity to prokaryotic genes are either connected to archaeal genes or to bacterial genes.

Gene families with bacterial origin

**Variable #** in eukaryotic genomes: these are more evolvable components

Gene families with archaeal origins

Relatively constant # in eukaryotic genomes: these are more stable components
In investigating eukaryotic evolution, the following rules of connection apply:

- > 80% cover
- >= 30% ID
- E-value < 1e^{-5}

The diagram illustrates a connection tree with labeled nodes indicating bacterial and archaeal genes.
Searching for Endosymbiotic Gene Transfer in eukaryotes

Rules of connection:
> 80% cover, >= 30% ID, E-value < 1e-5
Seeing EGT in SSN

Distant eukaryotic homologs

Bacterial genes
Archaeal genes

Shortest path

EUK 1  BAC  ARC  EUK 2

Alvarez-Ponce et al. PNAS 2013
Searching for undetected EGT and (endo)symbioses?

Your favorite molecular dataset
(sequences from hosts, putative symbionts, relatives...)

Shortest path
HOST copy 1 \(\rightarrow\) SYMBIONT \(\rightarrow\) X \(\rightarrow\) HOST copy 2
Seeing composite genes in SSN

Composite

Component

Intransitive triplets

Jachiet et al. Bioinformatics
Many processes can give birth to composite genes.

Gene fusion

Exon-shuffling, domain-shuffling

Illegitimate recombination

Patthy, 1999. *Gene*
Yanai, I et al. 2001 P. N. A. S.

Systematic analysis of composite genes in composite organisms

EUK

CYANO.

α-prot

ARC
Systematic analysis of composite genes in composite organisms

‘The light challenge (phototropism in Chlamydomonas)’
‘Composites in composite?’: a cutting-edge question

- Archaeplastida
- Paulinella chromatophora
- Planococcus citri
- Desulfurococcales
- Methanosarcinales
- Haloarchaea
- Sulfolobales
- Thermoproteales

Gavelis et al. ‘Eye-like ocelloids are built from different endosymbiotically acquired components’ Nature - July 2015

W. Martin’s work

Méheust et al. TREE 2015
Saltatory molecular evolution is also at play in mobile elements.

3 008 genomes, 122 392 genes, (BLAST E-value < 1e-10, masking of low complexity regions)

A sequence similarity network

See Jananan Pathmanathan’s talk...
Proof of concept: re-discovery of known composite genes

A. Excision bases repair

B. Host specificity


-log (E-value)  

-  < 40  
  40 - 50  
  50-80  
  80-200  
  > 200
The bigger picture: genetic diversity in viral genomes

9,872 composite genes
8% of genes

Component
Composite 8% data
Composite of composite 3%
Genetic fragments from very different viruses are combined

Evolutionary classes (Koonin)

- +RNA
- Retroid + elements
- Small DNA, plasmids...
- Tailed bacteriophages
- NCLDV (large DNA)

Jachiet et al., 2014. GBE

Genetic fragments from very different viruses are combined.

Jachiet et al., 2014. GBE


Diemer et al. 2012. Biology Direct
4. High divergence raises major challenges

- High divergence

see Phil Lopez’s and FJ Lapointe’s talk on genetic diversity and networks.
Why evolutionists need a broader picture

Some processes in cultured organisms

All processes in nature

An important American

All American citizens
CONCLUSION: BIOLOGISTS FACE A DIVERSITY OF PROCESSES AND OBJECTS

Consortia with emerging properties

Fusions with emerging properties

Clubs sharing goods

Entities high-jacking other's bodies
REPRESENTING AND ANALYSING ALL THESE PHENOMENA IS COMPLEX
EVOLUTIONISTS CAN (AND SHOULD) BE INCREASINGLY PLURALISTS

Tree models
Network models

Evolutionary phenomena
AND THERE IS EVEN MORE TO LIFE!

Evolution of synthetic micro-organisms and their communities...
Evolutionary biologists often assume that they study the evolution of simple objects (e.g., chunks of trees), while in fact they are studying the evolution of networks. When networks have causal properties on their own or influence the evolution of their components, evolutionists have additional causes of phenotypes worth exploring with networks methods.
THANKS A LOT FOR INVITING US

Please register for free classes on networks in Roscoff (3-9 july 2016)!

Thanks to all my great colleagues: Dr. Lopez, Pr. Lapointe, Pr. Burian, Dr. Bouchard, Dr. O’Malley, Dr. Halary, Dr. Jachiet, Dr. Bittner, Pr. Bhattacharya, Dr. Cheng, Dr. Karkar, Dr. Alvarez Ponce, Pr. Habib, Pr. McInerney

Ed. Belin

Ed. Matériologiques

Philippe Lopez

S. Halary

R. Méheust

J. Pathmanathan

P-A Jachiet

E. Corel
Introgression for example results in ‘open’ pangenomes

Core genome of *E. coli* = 6%

In the lab, only 61 genes over 246,065 cannot be transferred to *E. coli*.

Sorek et al., Science. 2007
Informative patterns in bipartite graphs

- Partitions: **connected components**

  - correspond to sets of genomes using an exclusive pool of genes with > 90% identity

---

When connected components are **homogeneous** (in host’s taxonomy, lifestyle or vehicle type), transmission is biased in terms of targets.
Transmission groups homogeneity may also be related to original molecular properties of the genomes.

Spiroplasma kunkelii virus SkV1_CR2-3x

Spiroplasma phage 1-C74

Spiroplasma phage 1-R8A2B

Common variant genetic code (*Mycoplasma/Spiroplasma* code)

A variant genetic code may introduce a **barrier**, preventing succesfull transmission of genes, privatized by the club of genomes.

*McInerney et al. Biol. Direct 2011*
Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Shijulal Nelson-Sathi, Tal Dagan, Giddy Landan, Arnold Janssen, Mike Steel, James O. McInerney, Uwe Deppenmeier, and William F. Martin

*Institute of Molecular Evolution, *Institute of Genomic Microbiology, *Mathematisches Institut, Heinrich Heine University, 40225 Düsseldorf, Germany; *Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand; *Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; and *Institute of Microbiology and Biotechnology, University of Bonn, 53115 Bonn, Germany

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved October 25, 2012 (received for review May 29, 2012)

www.pnas.org/cgi/doi/10.1073/pnas.1209119109 PNAS

« This sounds like a job for point mutation... » (W. Martin)
Le réseau du mobilome fait partie d’un réseau de partage plus vaste

<table>
<thead>
<tr>
<th>% identité minimal</th>
<th>#noeuds</th>
<th>%noeuds dans des triangles multi-niveaux</th>
</tr>
</thead>
<tbody>
<tr>
<td>336,402 cellular proteins (54 Archaea 70 bacteria, 7 Eukaryotes) + 228,042 sequences from mobile genetic element .</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>295,606</td>
<td>40.4 %</td>
</tr>
<tr>
<td>99</td>
<td>44,592</td>
<td>50.9 %</td>
</tr>
</tbody>
</table>

Informative patterns in bipartite graphs

- **Twins:**
  
  Several gene families exclusively shared by the same sets of genomes
  
  **Bacteria**  **Plasmids**

- May identify common genetic goods common of a club of genomes
Practical implementation: searching for patterns in bipartite graphs

**Twins:**
Several gene families exclusively shared by the same sets of genomes

**Ruminococcus bromii**
*Firmicutes*

**Eubacterium rectale**
*Firmicutes*

**Fibrobacter succinogenes**
*Fibrobacter/ Acidobacteria*

---

Predicted **membrane protein** + **Signal transduction** histidine kinase + Na+-driven **multidrug efflux pump**

---

UDP-N-acetylglucosamine 2-epimerase + nucleoside-diphosphate sugar epimerases
Seing composite genes in networks

Intransitive triplets

Informations
• Network topology
• Aligned regions
• Alignment scores (logEvalue)
La problématique des transitions évolutives a de grandes chances de revenir sur le devant de la scène

LETTER

Origins of major archaeal clades correspond to gene acquisitions from bacteria

Shijulal Nelson-Sathi¹, Filipa L. Sousa¹, Mayo Roettger¹, Nabor Lozada-Chávez¹, Thorsten Thiergart¹, Arnold Janssen², David Bryant³, Giddy Landan⁴, Peter Schönheit⁵, Bettina Siebers⁶, James O. McInerney⁷ & William F. Martin¹,⁸
Composite genes have key functions for viruses

Functional distribution

- **83%** Cellular
- **15%** Viral

COG, KOG categories

Annotation: RPSBLAST

Jachiet et al., 2014. GBE
Composite genes have key functions for viruses

Jachiet et al., 2014. GBE
Composite genes have key functions for viruses

- Highjacking cellular machinery
- Attachment and entrance

Fisher test, p-value <0.05
Screening for patterns in large datasets
3. Holobionts raise numerous challenges

- Multiplicity of interacting transmission systems and channels

- which complicates analyses of the causes of phenotypes
  (e.g. species incompatibility, health conditions)

- Are processes and channels at the origin of traits in holobionts fully described?

Brucker & Bordenstein, TREE 2012
Consider the tripartite ‘Human+gut microbes+ MGE’ system

Microbiota’s components:
- cells
- mobile genetic elements
- 3 types of genes families:
  - No similarity with MGE
  - Similarity with 1 type of MGE
  - Similarity with >1 type of MGE
  - More ubiquitous in humans

• 31 humans

Why are our gut contents partly similar?

- Genes families
- Host

>=80% mutual cover
>= 90% identity

• can model **different transmission channels & systems**
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
Why are our gut contents partly similar?

- Suggests **host phenotype** caused by multiple transmission channels (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
The mobilome provides genetic resilience and public genetic goods to gut microbes.

Hypergeometric test, Bonferroni correction, p-value < 0.01

Resident
Mobilized by 1 type of vector
Mobilized by >=2 types of vectors

(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.
Gene families mobilized by an increasing number of types of vectors have increasingly broader host ranges distribution.

- **Resident**
- Mobilized by 1 type of vector
- Mobilized by $\geq 2$ types of vectors

![Image of bacteria and people](image)

- **# of host genera**
- **# of host phyla**
- **# of human hosts**

Mann Whitney Wilcoxon test, $\alpha = 0.01$

Bicep et al. in prep.
About holobionts

• Increasingly mobilized genes are increasingly shared in gut microbes and in humans

**Multiple transmission channels** introduce a common structure within the human gut microbiome, at 2 distinct hosts levels

- Horizontal gene transfer partly shapes the human gut functional core microbiome
- Integrative dynamic studies of tripartite ‘MGE-microbes-human’ holobiont are needed
Evidence for significant amounts of gene externalization.

Gene externalization is another form of paralogy, e.g., an extragenomic paralogy, where paralogs sit on different genomes.

Within communities, extragenomic paralogy:
- introduces some genetic heterogeneity
- results in gene redundancy
- decreases the probability of the externalized gene expression.
Functional categories have distinct externalization profiles. Most recent events show an identity between cellular and MGE genes >= x % ID.
• Is it random in terms of targets? 
  (no, taxonomy and lifestyle constrain horizontal transmission)

• Is it random in terms of what gets transmitted? 
  (no, some functions and genetic goods are much more frequently transmitted)

Transmission and gene externalization is an integral part of microbial evolution
Les génomes eucaryotes ont 2 types de composants

Dans les réseaux, les séquences eucaryotes ressemblant à des séquences procaryotes sont ou bien connectées à des gènes archaea ou bien à des gènes bactériens.
Les génomes eucaryotes ont 2 types de composants, qui n’évoluent pas pareil.

Dans les réseaux, les séquences eucaryotes ressemblant à des séquences procaryotes sont ou bien connectées à des gènes archaea ou bien à des gènes bactériens.

<table>
<thead>
<tr>
<th>Famille de gènes d’origine bactérienne</th>
<th>Famille de gènes d’origine archaea</th>
</tr>
</thead>
<tbody>
<tr>
<td># variable dans les génomes eucaryotes: ce sont des composants plus évoluables</td>
<td># assez constant chez tous les génomes eucaryotes: ce sont des composants plus stables</td>
</tr>
</tbody>
</table>
Preliminary works with bipartite graphs

- can model different transmission channels & systems
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
Interest for this formalism using bipartite graphs

- can model different transmission channels & systems
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
Interest for this formalism using bipartite graphs

- can model different transmission channels & systems
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
Interest for this formalism using bipartite graphs

- Analyze the **transmission processes** associated with distinct genes distributions in hosts
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
Interest for this formalism using bipartite graphs

- Suggests **host phenotype** caused by multiple transmission channels
  (resident, mobilized by virus, mobilized by plasmid, mobilized by both vectors)
From genome networks to bipartite graphs: a finer grained-view

- Eases the identification of **clubs of genomes** sharing genes or co-transmitted genes

Bapteste et al., PNAS 2012.