Symplectic biology: the cell as a living computer
Authors

in silico
- Gang Fang
- Etienne Larsabal
- Géraldine Pascal
- Eduardo Rocha
- Ivan Moszer
- Claudine Médigüe

in vivo / in vitro
- Agnieszka Sekowska
- Anne Marie Gilles
- Octavian Barzu

Collective
- Stanislas Noria

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
Background

Physics: matter, energy, time

Biology: Physics + information, coding, control...

Arithmetics: strings of whole numbers, recursivity, coding...

Computing: Arithmetics + program + machine...
As is the case for building up a machine, one needs a book of recipe to build up a cell. This asks for changing the text of the recipe into something concrete: this transfers « information ».

In a cell, information transfer is managed by the genetic program.
What is Life?

Three processes are needed for Life:

- **Information transfer** (Living Computers?) => the goal of genomics is to decipher the blueprint of the “read-only” memory of the machine.

Driving force for a coupling between the genome structure and the structure of the cell:

- **Metabolism** (Internal organisation)
- **Compartmentalisation** (General structure)
What is computing?

Two processes are needed for computing:

- A read/write machine

- A program on a physical support (typically, a tape illustrates the sequential string of symbols that makes up the program), split (in practice) into two entities:
  - Program (providing the goal)
  - Data (providing the context)

The machine is distinct from the program.
Cells as computers

Genomics rest on an alphabetic metaphor, that of a text written with a four-letter alphabet, acting as a program

Conjecture: do cells behave as computers?

Genetic engineering
Viruses
Horizontal gene transfer
Cloning animal cells

all point to separation between

Machine
Data + Program
Is there a map of the cell in the chromosome?

If the machine has not only to behave as a computer but has also to construct the machine itself, one must find an image of the machine somewhere in the machine (J. von Neumann)


© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr
Is the gene order random in the chromosomes?

At first sight, despite different DNA management processes not much is conserved, and genes transferred from other organisms are distributed throughout genomes.

However, groups of genes such as operons or pathogenicity islands tend to cluster in specific places, and they code for proteins with common functions.

**First question:** how are generated and where are located repeats in the genome sequence?
Caveat: Repeats are meaningful

What does the smaller cube the round support supports support?
A ball.

Remember also:
This clock has a minute minute hand
Repeats in bacteria

- Abcissa: first occurrence of the repeat
- Ordinate: second position of the repeat
- Diagonal: repeats are located near to each other
DNA management: Repeats in genomes

Genome organisation

The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, which is full of « flexible patterns of class A »
The genome of *Helicobacter pylori* displays a period of 11 over regions spanning 60 nucleotides.
The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns.
Class A flexible patterns are ubiquitous

The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns
The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns
A universal rule: class
A flexible patterns

The flexible nature of the patterns permits DNA to accommodate superturns or local bending
The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, where the constraints of replication are visible in the leading and the lagging strand.
Different “Operating Systems”?

- **Escherichia coli**: 55% leading
- **Treponema pallidum**: 65% leading
- **Bacillus subtilis**: 75% leading
- **Thermoanaerobacter tengcongensis**: 87% leading

**CDS density**

**Leading CDS density**

http://www.pasteur.fr/recherche/unites/REG

(updated from Kunst et al, Nature, 97)
Chosing arbitrarily an origin of replication and a property of the strand (base composition, codon composition, codon usage, amino acid composition of the coded protein...) one can use discriminant analysis to see whether the hypothesis holds.

To lag or to lead, that is the question.
Visible even in proteins…

**B. burgdorferi**

![Graph 1: Threonine vs. Valine](image1)

![Graph 2: Isoleucine vs. Valine](image2)
Essentiality in *B. subtilis*

- **Essential genes**
  - Non-highly expressed: 100%
  - Highly expressed: 75%

- **Non-essential genes**
  - Non-highly expressed: 50%
  - Highly expressed: 25%

**Legend**
- Lagging
- Leading

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr
Gene persistence

B. subtilis

Number of B. subtilis genes
- All genes
- Essential genes

Number of genomes with an orthologue

E. coli

Number of E. coli genes
- All genes
- Essential genes

Number of genomes with an orthologue

Functional Categories of Nearly Ubiquitous Genes
- Identified as essential in both
  B. subtilis & E. coli
- Essential in B. subtilis only
- Essential in E. coli only
- Not identified as essential

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr
Gene persistence

Some of the genes missing from the list of persistent genes have diverged considerably. To assess the contribution of this effect we measured for each pair of genomes the correlation between the similarity of orthologous pairs and that of the 16S rRNA. As expected, the correlations were high. For example (Figure A), 38% (resp. 48%) of B. subtilis (resp. E. coli) persistent genes showed a correlation coefficient >0.9 between the sequence similarity of the pair of orthologs and the 16S. In contrast, some genes (Figure B) evolve in an erratic way. This may be due to horizontal gene transfer, local adaptations leading to faster or slower evolutionary pace, or simply wrong assignments of orthology. The latter can be a significant problem, especially in large protein families. However, the genes presenting such an erratic pattern are rare in the persistent set.
Replication transcription conflicts

- Transcription may proceed opposite to the movement of the replication fork movement
- This will abort transcription, leading to truncated mRNA
- If translated truncated mRNA may lead to truncated proteins, this will become negative dominant if in complexes…


© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr
When polymerases collide

Co-oriented
- DNA polymerase (DNAP) deceleration
- End of transcription

Consequences:
1. Replication slow-down
2. Loss of transcripts

Head-on
- Arrest of RNA polymerase (RNAP) & DNA polymerase (DNAP)
- Transcription abortion

Consequences:
1. Aborted transcripts
2. Truncated essential proteins

E. Rocha

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr
Distribution of highly expressed genes

Fast growers | Slow growers

Highly expressed genes cluster near the origin in fast-growing bacteria

B. subtilis | E. coli | C. crescentus | M. tuberculosis

0% 10% 20% 30% 40% 50% 60% 70%

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr
Genes do not operate in isolation
Proteins are part of complexes, as are parts in an engine
It is important to understand their relationships, as those in the planks which make a boat

*The Delphic Boat*: Harvard University Press, February 2003
Gene vicinity: synteny
Multivariate Analyses

In contrast to standard genetics, genomics analyses large collections of genes and gene products.

Multivariate analyses try to extract information by simplifying the number of relevant descriptors in the objects of interest.

Principal Component Analysis uses the centered average and a simple distance (identity); it is the reference method.

Correspondence Analysis belongs to the same family, but it uses the $\chi^2$ measure as a distance. This allows the user not only to work with highly heterogeneous objects but also to work simultaneously on the space of objects and on the space of descriptors.

Independent Component Analysis uses the non gaussian character of the values associated to descriptors.
Bias in amino acid distribution

Neighbourhood: distribution of aminoacids in the proteome
Universal biases in protein amino acid composition

- **First axis**: separates Integral Inner Membrane Proteins (IIMP) from the rest; driven by opposition between charged and large hydrophobic residues.

- **Second axis**: separates proteins according to an opposition driven by the G+C content of the *first* codon base.

- **Third axis**: separates proteins by their content in aromatic amino acids; enriched in orphan proteins.
The “gluons”

- There is an aromatic residues-oriented bias in all genomes

- With proteins of the same size this opposes ribosomal proteins to orphan proteins

- Hypothesis: orphans are “self”-specific proteins that stabilise complexes, they act as “gluons”
Temperature-dependent biases in protein amino acid composition

- The amino acid composition of proteins depends heavily on the phylogeny => need to compare organisms related to each other.
- The general trend of amino acid composition bias is to avoid some amino acids at higher temperatures.
- Mesophilic bacteria belong to at least two different classes (in a 5-clusters analysis).
- Biases are always dominated by the IIMP clustering.
Temperature-dependent amino acid biases
Codon usage biases

- 20 amino acids 61 codons
- Study of the genes in the codon space, using Correspondence Analysis ($\chi^2$ measure)
- At least three classes of genes, including one corresponding to horizontal transfer


© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr
Gene exchange

Class I: core metabolism

Class II: high expression in exponential growth

Class III: horizontal transfer

Genes expressed at a high level under exponential growth conditions
Codon usage, organisation and evolution of the *B. subtilis* genome

<table>
<thead>
<tr>
<th>gene1</th>
<th>AAA AAC AAG ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_{11} n_{21} n_{31}</td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td></td>
</tr>
<tr>
<td>n_{12} n_{22} n_{32}</td>
<td></td>
</tr>
<tr>
<td>gene3</td>
<td></td>
</tr>
<tr>
<td>n_{13} n_{23} n_{33}</td>
<td></td>
</tr>
<tr>
<td>....</td>
<td></td>
</tr>
<tr>
<td>n_{1n} n_{2n} n_{3n}</td>
<td></td>
</tr>
<tr>
<td>geneN</td>
<td></td>
</tr>
</tbody>
</table>

Correspondence analysis

Classification

- **Highly expressed**
- **Atypical / HGT**
- **Others**

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr
The cell organizers

It is too early to understand the selection pressures that organize the cell architecture. However, at least in bacteria, the role of gasses and chemical highly reactive radicals play probably a major role. Most of the corresponding genes are still unknown….
Sulfur undergoes oxido-reduction reactions from -2 to +6

Incorporation of sulfur into metabolism usually requires reduction to the gaseous form $\text{H}_2\text{S}$

$\text{H}_2\text{S}$ is highly reactive, in particular towards dioxygen

$\Rightarrow$ These two gasses, despite their diffusion properties, must be kept separate as much as possible

Sulfur scavenging is energy-costly

$\Rightarrow$ Sulfur containing molecules have to be recycled

Sulfur metabolism: an unexpected organiser of the cell’s architecture

• Sulfur metabolism-related proteins are more acidic (average pI 6.5) than bulk proteins (richer in asp and glu), they are poor in serine residues

• They are significantly poor in sulfur-containing amino-acids

• Their genes are very poor in codons ATA, AGA and TCA

• There are no class III (horizontal transfer) genes in the class (only 2 in 150 genes)

• => sulfur-metabolism genes are ancestral and may for a core structure for the E. coli genome
Proximity in the chromosome
Sulphur islands

E.P.C. Rocha, A. Sekowska & A. Danchin Sulfur islands in the *Escherichia coli* genome: markers of the cell’s architecture?

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
The error catastrophe

- Similarity in sequence leads to functional inference
- Because of recruitment of pre-existing structures, there is often no obvious link between a structure and a function (the book-paperweight)
- Hence a propagation of annotation errors
- \textit{ykrS} (\textit{mtnA}) annotated as « translation factor » is a component of sulfur metabolism!

A new metabolic pathway

A. Sekowska
Just so story: proximity in the genome

Escherichia coli

Bacillus subtilis no rpsA !!!

Haemophilus influenzae

Sinorhizobium meliloti
The pyrimidine diphosphate paradox

In order to make deoxyribonucleotides the cell uses ribonucleosides diphosphates, not triphosphates

\[ \text{NDP} \rightarrow \text{dNDP} \rightarrow \text{dNTP} \]

\[ \text{NDR} \quad \quad \text{NDK} \]

And here is the paradox:

\[ \text{OMP} \rightarrow \text{UMP} \rightarrow \text{UDP} \rightarrow \text{UTP} \rightarrow \text{CTP} \]

no CDP !!!
How is the paradox resolved?

OMP → UMP → UDP → UTP → CTP

RNases → CMP → Cmk → CDP → dCDP → DNA

mRNA

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr
 Phylogenetic neighbours: the S1 box

• *rpsA* codes for ribosomal protein S1. It contains the S1 box (PROSITE PS50126). Many other proteins contain a similar box: polynucleotide phosphorylase, RNases E, G and R, RNA helicases etc.
• protein RegB of bacteriophage T4, associated to S1, cuts mRNA at GAGG motifs.
• S1 is a subunit of bacteriophage Qβ replicase…

=> All this points to a function for S1 in RNA metabolism
<table>
<thead>
<tr>
<th>Gene</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>bla</td>
<td>long mRNA turnover</td>
</tr>
<tr>
<td>cat</td>
<td></td>
</tr>
<tr>
<td>dicB</td>
<td></td>
</tr>
<tr>
<td>lpp</td>
<td></td>
</tr>
<tr>
<td>ompA</td>
<td></td>
</tr>
<tr>
<td>pyrF</td>
<td>pyrimidine metabolism</td>
</tr>
<tr>
<td>hflB</td>
<td>cell architecture</td>
</tr>
<tr>
<td>ftsH</td>
<td></td>
</tr>
<tr>
<td>mrsACF</td>
<td></td>
</tr>
<tr>
<td>lpp</td>
<td></td>
</tr>
<tr>
<td>nusA</td>
<td>RNA maturation and turnover</td>
</tr>
<tr>
<td>pcnB</td>
<td></td>
</tr>
<tr>
<td>metY</td>
<td></td>
</tr>
<tr>
<td>pnp</td>
<td></td>
</tr>
<tr>
<td>rna</td>
<td></td>
</tr>
<tr>
<td>rnb</td>
<td></td>
</tr>
<tr>
<td>rnc</td>
<td></td>
</tr>
<tr>
<td>rne/ams</td>
<td></td>
</tr>
<tr>
<td>rng</td>
<td></td>
</tr>
<tr>
<td>rph</td>
<td></td>
</tr>
<tr>
<td>trxA</td>
<td>oxido-reduction, subunit of T7 replicase, needed for synthesis of deoxyribonucleotides</td>
</tr>
</tbody>
</table>

Codon usage bias neighbours
Protein complexes: the Degradosome

- PNPase
- PolyA polymerase
- RNAse E
- S1
- Polyyphosphate kinase
- Enolase

mRNA degradation

CDP for de novo DNA synthesis

GDP recycling of GTP for carbohydrate secretion

GDP + PEP $\rightarrow$ GTP

NDK + PYK

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr
Just so story: the \textit{cmk rpsA} operon

\\[ \text{cmk (mssA)} \rightarrow \text{rpsA} \rightarrow \text{Escherichia coli} \]

\textit{mssA} was discovered as a suppressor of \textit{smbA (pyrH)}, itself a suppressor of MukB, a myosin-like protein involved in chromosome segregation.

=> DNA synthesis is involved in the function.

**Conclusion:**

The function of the \textit{cmk rpsA} operon is to make CDP for DNA synthesis.
Selection pressure for compartmentalisation: a dangerous intermediate

Uridylate kinase (UMK) pyrH (smbA)

OMP → UMP → UDP → UTP → CTP

No CDP: no DNA…

dTTP

dTDP

dTMP

DNA

DNA

dUTP → dUMP + PPi

In conclusion:

UMK must be compartmentalised

A prediction: ribosome recycling and UTP

This organisation is conserved in most Gram+ and Gram- bacteria. Why?
Ribosome recycling and UTP

*frr* codes for the ribosome recycling factor, that allows 70S ribosomes to split into 30S and 50S subunits. In polycistronic operons, the 70S ribosome can go on from one gene to the next one without recycling (*this requires formylation of the first methionine*). At the end of the message, the ribosomes must recycle. This happens in a context where transcripts make stem and loops, ending with a polyU sequence.

**Conjecture:** is UTP controlling the activity of Frr? Remember that one cannot speak of « concentrations » of molecules in a cell. 1 micromolar would mean 600 molecules. There are 20,000 ribosomes, therefore 1 mM means only **30 individual molecules** in the immediate vicinity of each ribosome...
At Rho-independent sites for termination of transcription the messenger RNA ends with rows of U. This must lower the local availability of UTP….

This suggests Frr as a drug target, with analogs of UTP as leads….
A preconceived ideology

Mathematics

Physics

Chemistry

Biology

Sociology

Molecular Biology

Structural Biology

Genetics and Genomics

© Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr