Symplectic biology: Universals in microbial genomes

24 April 2006
Symplectic biology: The Delphic Boat

- Biology is a science of relationships between objects rather than from objects: from συν together, πλεκτειν, to weave
- Proteins are part of complexes, as are parts in an engine
- As for constructing a boat, failing to understand their relationships will result in ultimate failure of synthetic objects

*The Delphic Boat*: Harvard University Press, February 2003
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
- Compartmentalisation
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 programme), split (in practice) into two entities:
  - Programme (providing the goal)
  - Data (providing the context)

The machine is distinct from the programme
Cells as computers

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

**Conjecture:** do cells behave as computers?

- Genetic engineering
- Viruses
- Horizontal gene transfer
- Cloning animal cells

*all point to separation between*

- Machine
- Data + Programme

---

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

adanchin@pasteur.fr
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 (John von Neumann)
Is the gene order random in the chromosomes?

At first sight, consistent with 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. « Persistent » genes are clustered together.

Also, some motifs are ubiquitously present, suggesting general rules constraining genome organisation.

E Larsabal, A Danchin
Genomes are covered with ubiquitous 11bp periodic patterns, the "class A flexible patterns"
BMC Bioinformatics (2005) 6: 206

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

adanchin@pasteur.fr
A universal feature of the program: the period of 10-11.5

Helicobacter pylori

real

model

difference

Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG/
adanchin@pasteur.fr
Flexible motifs of type A

1- \(AxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAxTxAx\): All kingdoms

2- \(GxxxxTTxxxxCxxxxTTxxxxx\): Proteobacteria

4- \(xxxxxTxxxxxAGxxxxTTxxxxxTxxxxxx\): Archaea

5'-xxx-10xxxxxx0xxxxxx10xxxxxxbp-3'

The nucleotides composing this class A flexible pattern are fully accessible through this side and the dinucleotides are set in major grooves.

The nucleotides composing this class A flexible pattern are accessible through this side too but the dinucleotides are set in minor grooves.
Is it possible to see whether the position of genes in the chromosome is randomly distributed on the leading and lagging strand?
Genetics of Bacterial Genomes

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

adanchin@pasteur.fr
Genetics of Bacterial Genomes

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

<table>
<thead>
<tr>
<th>Organism</th>
<th>Ori</th>
<th>Ter</th>
<th>CDS Density</th>
<th>Leading CDS Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td></td>
<td>270</td>
<td>55%</td>
<td>90</td>
</tr>
<tr>
<td>Treponema pallidum</td>
<td></td>
<td>270</td>
<td>65%</td>
<td>90</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td></td>
<td>270</td>
<td>75%</td>
<td>90</td>
</tr>
<tr>
<td>Thermoanaerobacter tengcongensis</td>
<td>180</td>
<td>90</td>
<td>87%</td>
<td>90</td>
</tr>
</tbody>
</table>
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**

- % Threonine vs. % Valine
- % Isoleucine vs. % Valine
Essentiality in *B. subtilis*

Essentiality, not expressiveness, drives gene-strand bias in bacteria

EPC Rocha, A Danchin

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

adanchin@pasteur.fr
When polymerases collide

Co-oriented

Head-on

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

Consequences:
1. Aborted transcripts
2. Truncated essential proteins
Three examples of the role of the context

Microbial genes are of infinite diversity but there exists universals; only about 10% of their genes are of persistent and recognized function; we do not have yet a fair idea of the number of microbial species; the number of genes in a given species is highly variable (horizontal gene transfer).

Example 1: persistent genes
Example 2: orphan genes and universal amino acids
[Example 3: a new metabolic pathway]
....

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

**B. subtilis**

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

**E. coli**

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

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
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. The correlations were high. For example (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 RNA.

In contrast, some genes (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. The genes presenting such an erratic pattern are rare in the persistent set.

G Fang, EPC Rocha, A Danchin  
*How essential are non-essential genes?*  
Genomic islands

A clustering method based on the analysis of codon usage biases, using an information theory leads to group the genes into homogeneous clusters, which are not distributed randomly in the chromosome. One cluster corresponds to highly expressed genes. Other clusters are linked to specific functions or processes: horizontally transferred genes, motility or intermediary metabolism.
Genome islands

One cluster is related to expression levels. Other groups feature an overrepresentation of genes belonging to different functional groups: horizontally transferred genes, motility and intermediary metabolism. Genes with a similar bias are close on the chromosome and organized in coherent domains, more extended than operons, demonstrating a role of translation in structuring bacterial chromosomes. A sizeable contribution to this effect comes from the dynamic compartimentalization induced by the recycling of tRNAs, leading to gene expression rates dependent on their genomic and expression context.
Genome organization

P. haloplanktis
Pseudoalteromonas haloplanktis

P. haloplanktis: Correspondence Analysis

http://www.pasteur.fr/recherche/unites/REG/
adanchin@pasteur.fr
Universal biases in 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
Temperature-dependent biases in protein amino acid composition

- The general trend of amino acid composition bias is to avoid some amino acids at higher temperatures (associated to aging processes)
- Mesophilic bacteria belong to at least two different classes (in a 5-clusters analysis)
- Biases are always dominated by the IIMP clustering

Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125
Genome Research (2005) 15: 1325-1335
Comparative proteomics

A specific asparagine bias in psychrophiles

- Motility
- Cell wall, outer membrane
- Transport (TonB), secretion
- Adaptation to stress
- Metabolism of DNA and RNA

http://www.pasteur.fr/recherche/unites/REG/
adanchin@pasteur.fr
Asparagine deamidates: a major contribution to protein aging

- **Main post-translational modification**
- Reaction still poorly understood
- Spontaneous reaction (untargeted?)
- Affects the protein structure (and function?)
- Role in regulating protein folding
- Signal for degradation of intracellular proteins

Asparagine (N) → Degradation: succinimide

Intermediary: succinimide

Aspartate → Isoaspartate

http://www.pasteur.fr/recherche/unites/REG/
adanchin@pasteur.fr
In 1991, at the EU meeting on genome programs in Elounda, Greece, the presentation of the yeast chromosome III and the first 100 kb of the *Bacillus subtilis* genome revealed that, contrary to expectation (the only cases where this had been observed were phages, for obvious reasons), at least half of the genes uncovered were totally unknown, whether in structure or in function.
Orphans: the gluons

A remarkable role of aromatic amino acids creates a **universal bias**. Expressed orphan proteins are enriched in these residues, suggesting that they might participate in a process of gain of function during evolution. We postulate that the majority is made of proteins — **gluons** — involved in stabilising complexes, thus defining the "self" of the species.

G Pascal, C Médigue, A Danchin
Universal biases in protein composition of model prokaryotes

Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG/
adanchin@pasteur.fr
Why aromatic amino acids in orphan proteins?

From Orphans to « Gluons »

Thank you