



Mutual information (2)

$$\begin{aligned} \sum_{i \in I, j \in J} |p_{ij} - p_i \cdot p_j|^2 &= \sum \{(p_{ij} - p_i \cdot p_j)^2 / p_i \cdot p_j \mid i \in I, j \in J\} \\ &= \sum \{p_i \cdot p_j [(p_{ij} / p_i \cdot p_j)^2 - 1] \mid i \in I, j \in J\} \end{aligned}$$

which can be written as a function of the type

$$= \sum \{p_i \cdot p_j f(p_{ij} / p_i \cdot p_j) \mid i \in I, j \in J\}$$

knowing by definition of the laws p_I (normalisation)

that the sum of the product $p_i p_j$ as well as that of the mutual law p_{ij} is equal to 1.



Correlation and independence

Both ways to evaluate mutual information are similar: they correspond to two functions with the same value for $x = 1$, the same first and second derivatives:

in the χ^2 paradigm, it is x^2-1 , while in Shannon's law it is $x\log_2 x$



Conclusion

It is of major importance to adapt the appropriate measure of information, correlation etc., to the phenomenon of interest. In particular, in genome studies, this is important for evaluating codon usage biases and neighborhoods



A genetic Turing machine

The machine performs the following operations, all easily carried out (the operations performed by computers are the same conceptually speaking, although they are carried out in a more condensed and therefore more complicated fashion):

- changing a symbol in a finite number of places, after reading the symbols found there (note that changing more than one symbol at a time can be reduced to a finite number of successive basic changes).
- changing from the point which is being read to other points, at a given maximum distance away in the message
- changing the state of the machine

All this can be summarized a *Turing machine*. It can be specified by a finite set of 'quintuples', each of which has one of three possible forms:

$$p\alpha\beta Lq \text{ or } p\alpha\beta Rq \text{ or } p\alpha\beta Nq$$

A quintuple indicates that the machine is in configuration p , where it reads the symbol α , and replaces it by β , before changing to the configuration q , moving to the left (L), to the right (R), or not moving (N).



A genetic Turing machine

- ① In a Turing machine the machine is separated from the data, and the program
- ② Data and program play the same role (*i.e.* they can be thought of as 'declarations')



Cells as Turing Machines

Genomics rest on an alphabetic metaphore, that of a text written with a four-letter alphabet

Genetic engineering

Viruses

Horizontal gene transfer

Cloning animal cells

all point to separation between

Machine

Data

Programme



Paradox

If the machine has not only to behave as a Turing machine but also to make the machine, one must find a geometrical programme somewhere in the machine (J. von Neumann)



There is a map of the cell in the chromosome

- ✧ The gene order is not random in bacteria. Hot spots allow insertion of foreign DNA (in particular near the terminus of replication)
- ✧ In insect homeogenes follow the plan of the segmented organism
- ✧ In vertebrates four sets of homologous homeogenes follow the plan of the organism
- ✧ Coliculus, drosophiloculus, homunculus.... The algorithm is preformed, not the animal



Genome organisation

Is the gene order random?

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

However, pathogenicity islands tend to cluster in specific places, and they code for proteins with common functions



Genome organisation

The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, which differentiates the leading strand from the lagging strand

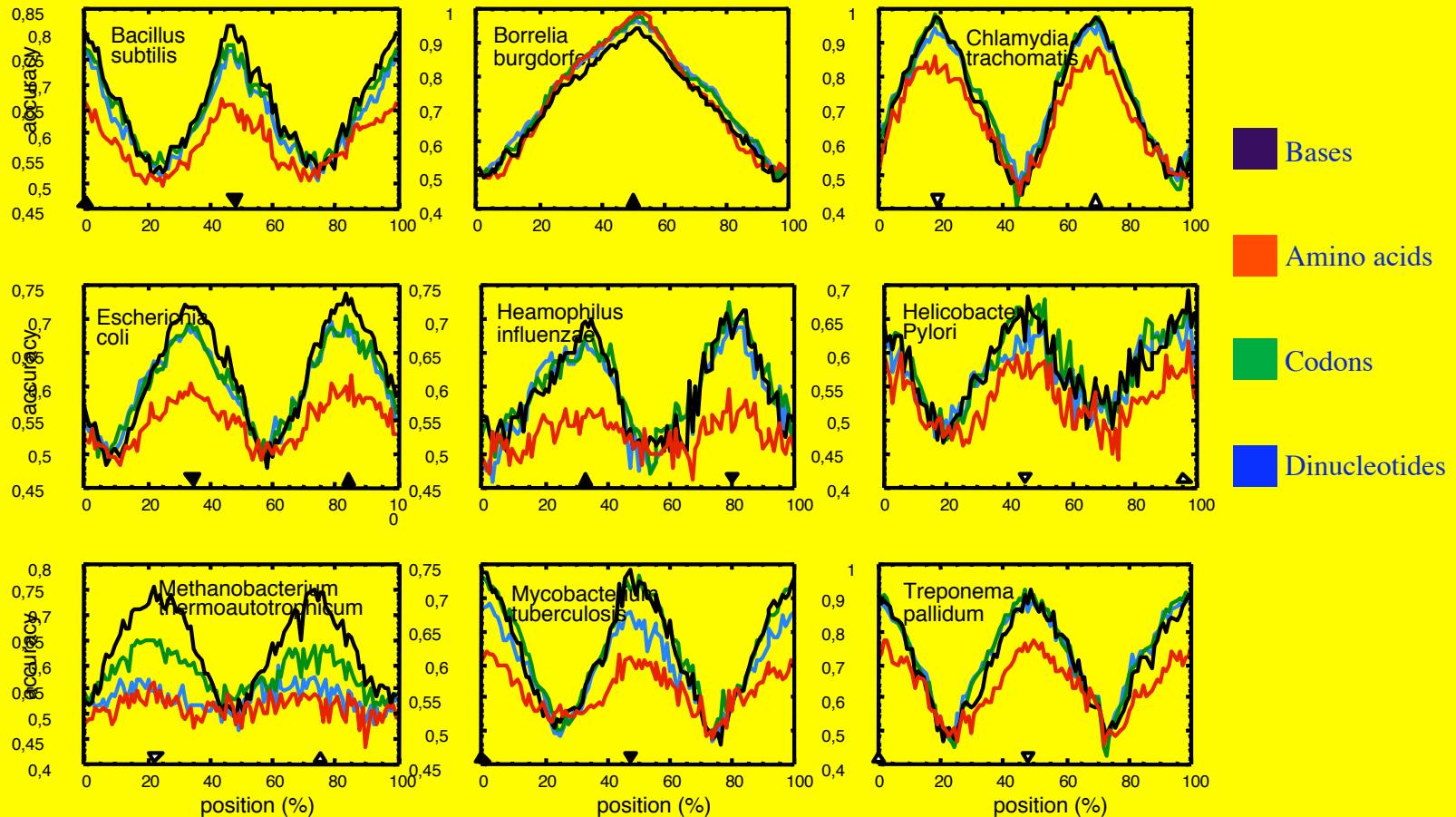


To lead or to lag...

Is it possible to see whether the position of genes in the chromosome is randomly distributed on the leading and lagging strand?

Choosing 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





Oxido-reduction

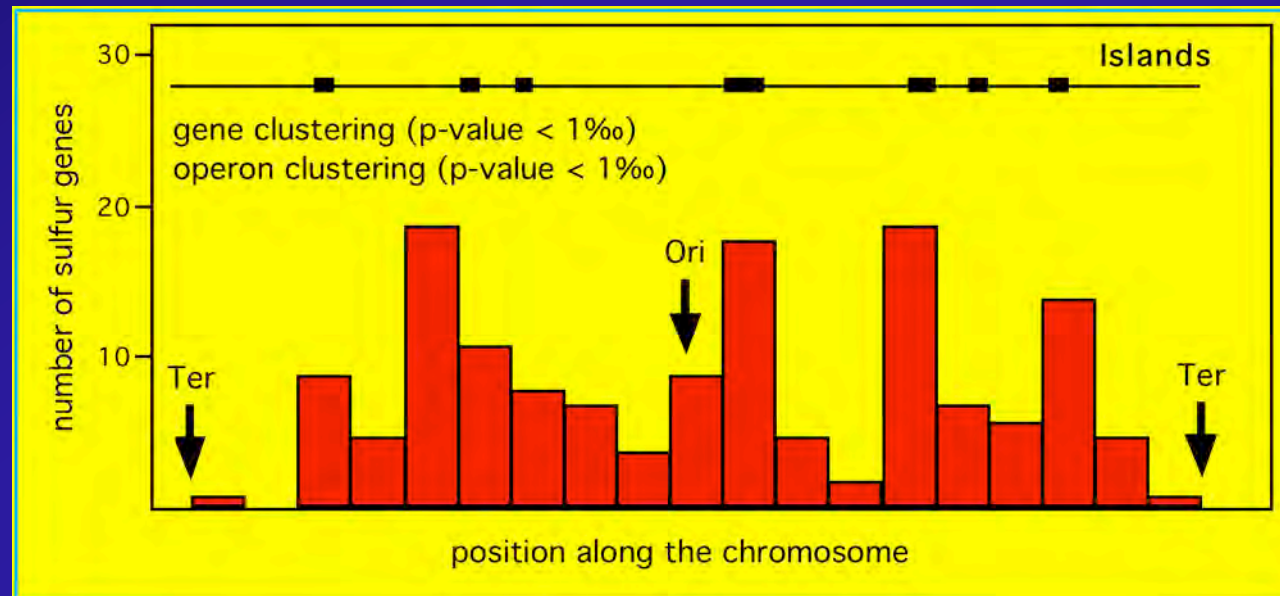
- ⊖ Sulfur undergoes oxido-reduction reactions from -2 to +6.
- ⊖ Incorporation of sulfur into metabolism usually requires reduction to the gaseous form H_2S .
- ⊖ H_2S is highly reactive, in particular towards dioxygen.
- ⊖ \Rightarrow These two gasses, despite their diffusion properties, must be kept 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 form a core structure for the *E. coli* genome

Sulphur islands

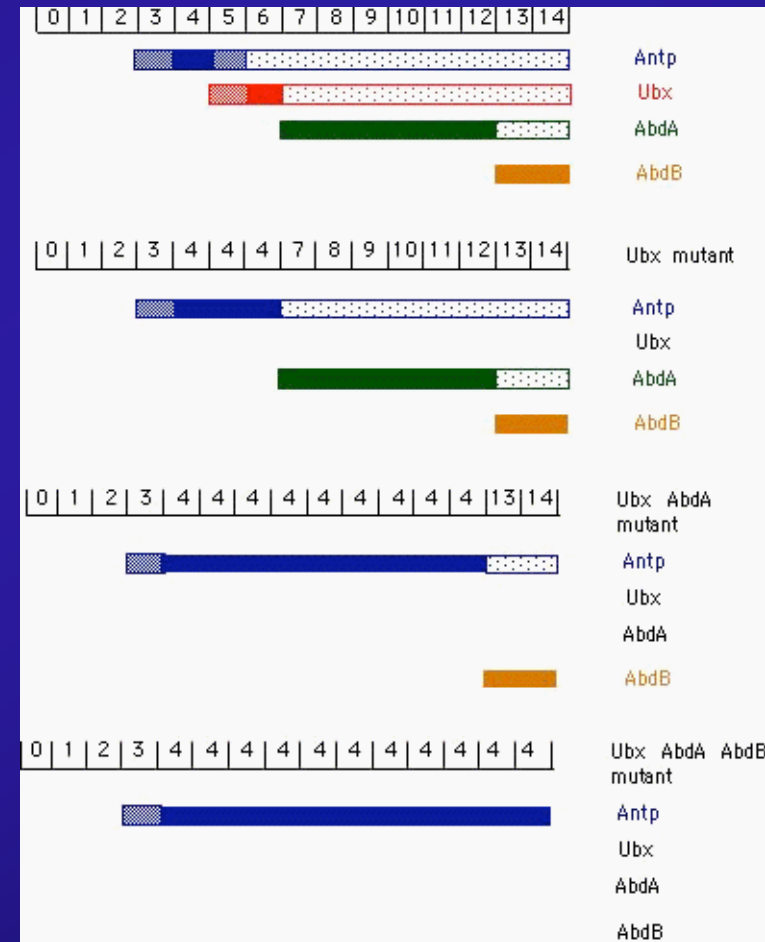
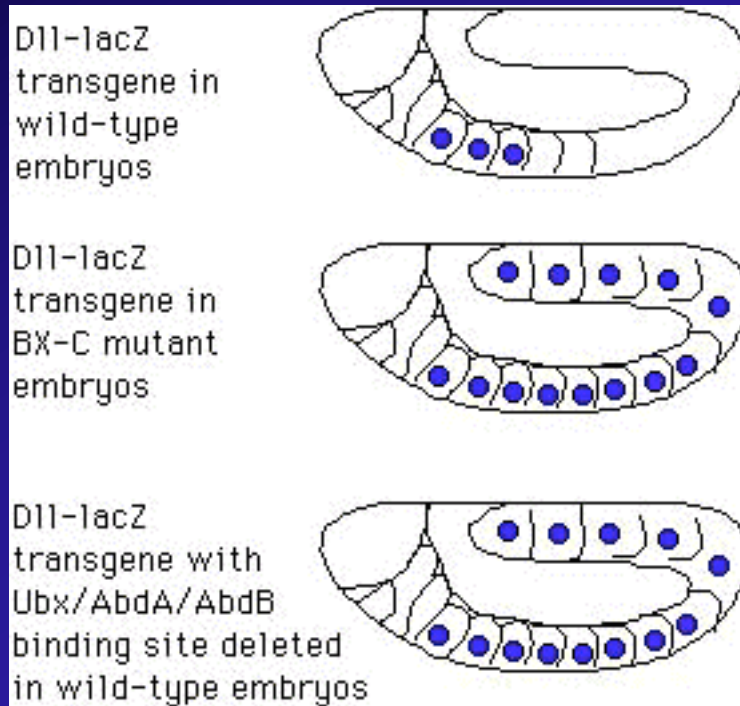




The cell organizers

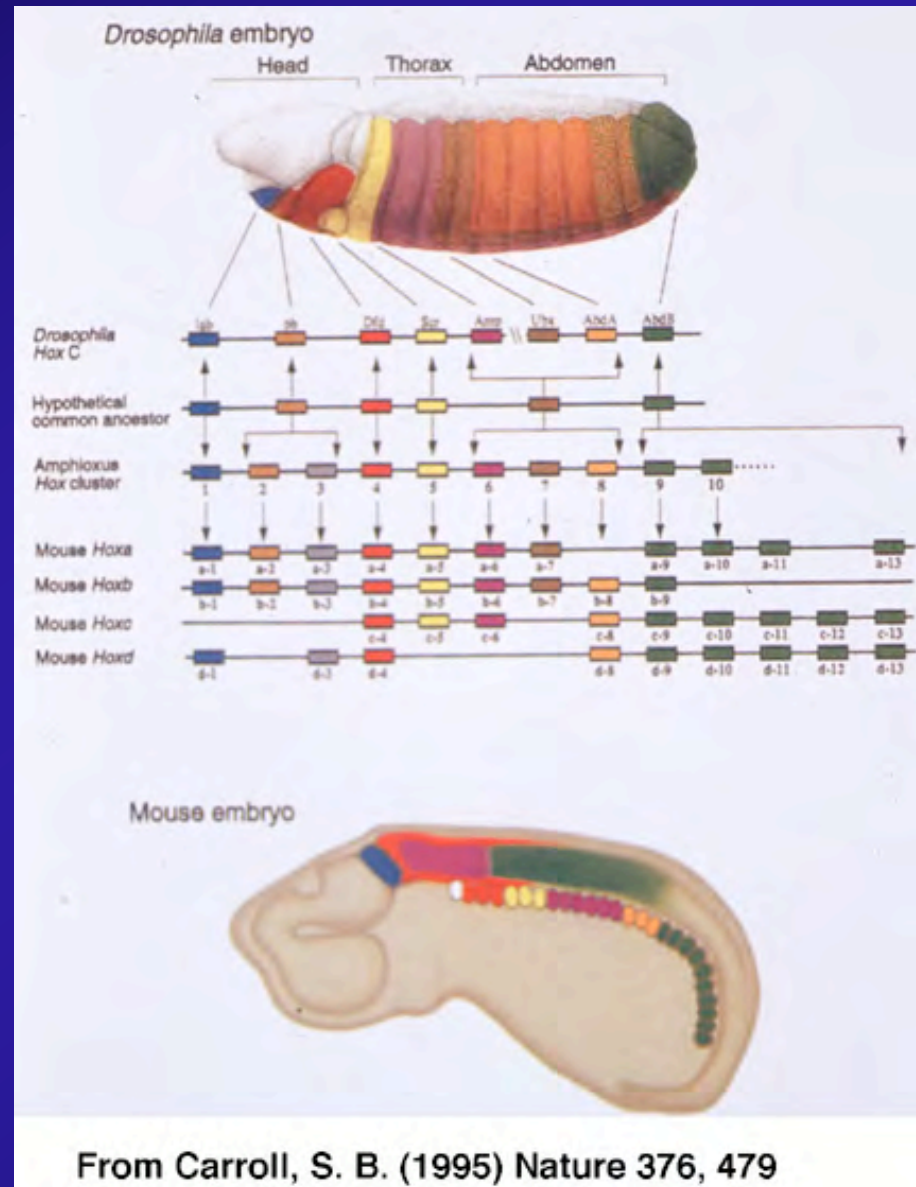
It is too early to understand the selection pressures that organize the cell architecture. However, at least in bacteria, the role of gasses and chemical highly reactive radicals play probably a major role. Most of the corresponding genes are still unknown....

Drosophiloculus, Homunculus ?

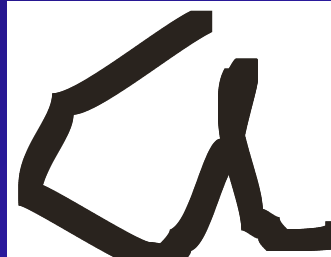


Drosophiloculus,

Homunculus ?



Gestalt and Algorithms



« Start from the top, middle

Go down from right to left

Accelerate

Turn right

Etc



Basic References

- ⊖ Hofstadter *Gödel, Escher, Bach: an Eternal Golden Braid*
- ⊖ Yockey *Information Theory and Molecular Biology*
- ⊖ Delahaye *Information, complexité et hasard*
- ⊖ Segal *Théorie de l'information: sciences, techniques et société de la seconde guerre mondiale à l'aube du XXIe siècle*
- ⊖ AD *Une Aurore de Pierres*, and *La Barque de Delphes (The Delphic Boat Harvard U Press)*