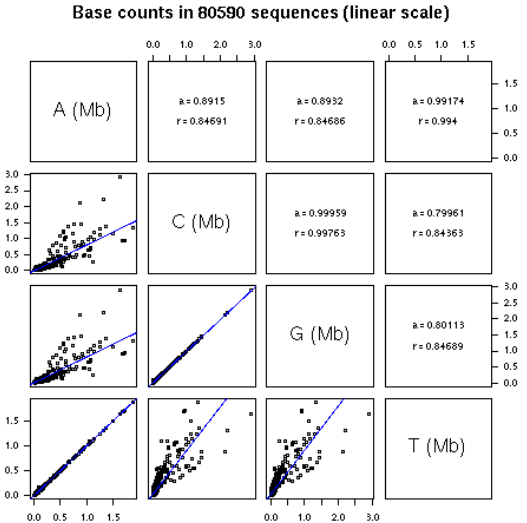
Chirochores: base composition biases

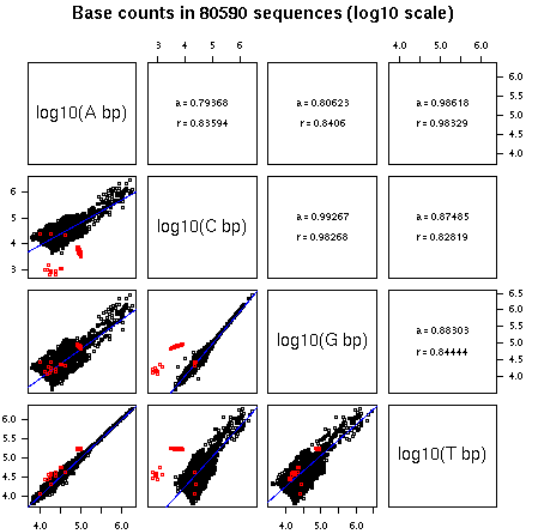
PR2 parity rule number 2

- In double-stranded DNA we have exactly $\mathrm{A} = \mathrm{T}$ and $\mathrm{C} = \mathrm{G}$ as a direct consequence of Watson-Crick base pairing rules.
- More surprisingly, in non artificial single-stranded DNA we have approximately $\mathrm{A} \approx \mathrm{T}$ and $\mathrm{C} \approx \mathrm{G}$.
- Parity rule number 2, or PR2 state, refers to the second assertion.
- The following examples are from the base counts in all ssDNA sequences ($>$ 50 Kb and $<$ 1 % of ambiguous bases) from GenBank (24-NOV-2004).

PR2 illustration (linear scale)



PR2 illustration (log scale, synthetic sequences in red)



PR1 parity rule number 1

- PR1 parity rule number 1 is an hypothesis about the process of evolution of the DNA sequences.
- PR1 hypothesis is that substitution rates are symmetric with respect to the two DNA strands.
- PR1 hypothesis doesn't mean that the substitution matrix itself is symmetric.

PR1 derivation

In the general case, let

$$\mathrm{r}(\mathrm{X} \to \mathrm{Y})$$

be the substitution rate from basis X to Y on one strand, and

$$\overline{\mathrm{r}}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}})$$

the substitution rate for the complementary event on the other strand. The apparent substitution rate on one strand is equal to the sum of these two substitution rates:

$$\mathrm{R}(\mathrm{X} \to \mathrm{Y}) = \mathrm{r}(\mathrm{X} \to \mathrm{Y}) + \overline{\mathrm{r}}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}})$$

PR1 derivation

Still in the general case, consider the complementary event:

$$\mathrm{R}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) = \mathrm{r}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) + \bar{\mathrm{r}}(\overline{\overline{\mathrm{X}}} \to \overline{\overline{\mathrm{Y}}})$$

Since

$$\overline{\overline{\mathrm{N}}} = \mathrm{N}$$

this can be rewritten as

$$\mathrm{R}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) = \mathrm{r}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) + \bar{\mathrm{r}}(\mathrm{X} \to \mathrm{Y})$$

PR1 derivation

We introduce now PR1 hypothesis:

PR1 hypothesis:

$$\forall X, Y \in N : r(X \to Y) = \overline{r}(X \to Y)$$

In general we had:

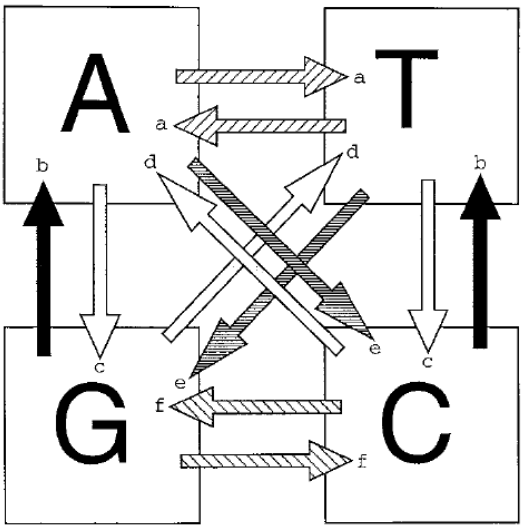
$$R(X \to Y) = r(X \to Y) + \overline{r}(\overline{X} \to \overline{Y})$$

$$R(\overline{X} \to \overline{Y}) = r(\overline{X} \to \overline{Y}) + \overline{r}(X \to Y)$$

So that under PR1 hypothesis we have:

$$R(X \to Y) = R(\overline{X} \to \overline{Y})$$

PR1 graphically



PR1 in matrix notations

$$\mathbf{X} = \begin{pmatrix} A(t) \\ T(t) \\ G(t) \\ C(t) \end{pmatrix}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{R}\mathbf{X}$$

$$\mathbf{R} = \begin{pmatrix} -a-e-c & a & b & d \\ a & -a-e-c & d & b \\ c & e & -b-d-f & f \\ e & c & f & -b-d-f \end{pmatrix}$$

Relationship between PR1 and PR2

- PR2 state is an asymptotic property of systems evolving under PR1 hypothesis.
- This true even for non-autonomous systems $\frac{d\mathbf{X}}{dt} = \mathbf{R}(t)\mathbf{X}$ (*Mol. Biol. Evol.* **16**:719-723).
- If PR2 is not observed for natural ssDNA sequences, PR1 can be rejected safely.

AT and GC skews

The AT skew is the deviation from A = T:

$$\mathrm{AT}_{\mathrm{skew}} = \frac{\mathrm{A} - \mathrm{T}}{\mathrm{A} + \mathrm{T}}$$

The GC skew is the deviation from C = G:

$$\mathrm{GC}_{\mathrm{skew}} = \frac{\mathrm{C} - \mathrm{G}}{\mathrm{C} + \mathrm{G}}$$

Skews are not the same for both strands

Let

$$\mathrm{ACGT}$$

the primary formula on one strand. Then, its complementary strand composition is given by :

$$\mathrm{A_tC_gG_cT_a}$$

The AT and GC skews are affected :

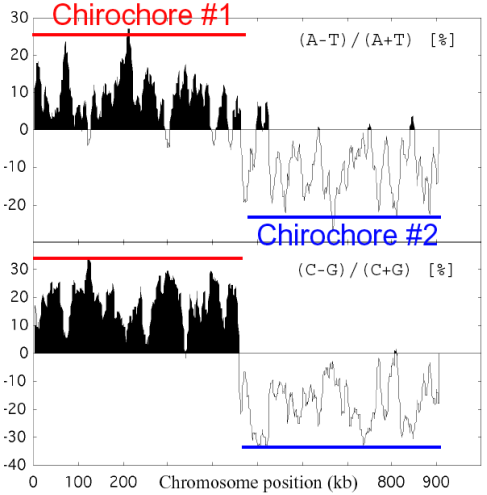
$$\frac{\mathrm{a}-\mathrm{t}}{\mathrm{a}+\mathrm{t}} = -\frac{\mathrm{t}-\mathrm{a}}{\mathrm{t}+\mathrm{a}}$$

$$\frac{\mathrm{c}-\mathrm{g}}{\mathrm{c}+\mathrm{g}} = -\frac{\mathrm{g}-\mathrm{c}}{\mathrm{g}+\mathrm{c}}$$

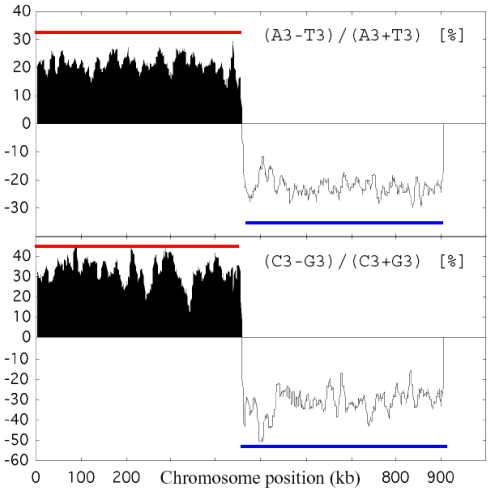
Chirochore: definition

- A chirochore is a segment of ssDNA homogeneous for its deviation from PR2 state.
- A chirochore is therefore characterized by constant AT and GC skews.
- Note the difference with isochores that are characterized by a constant G+C content.
- The study of these skews can be local or global: different mechanisms will explain skews at different scales.

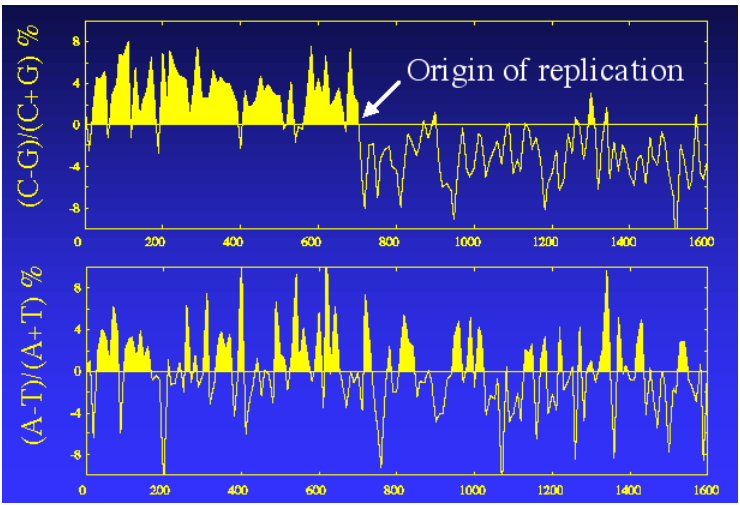
Chirochores in *B. burgdorferi*



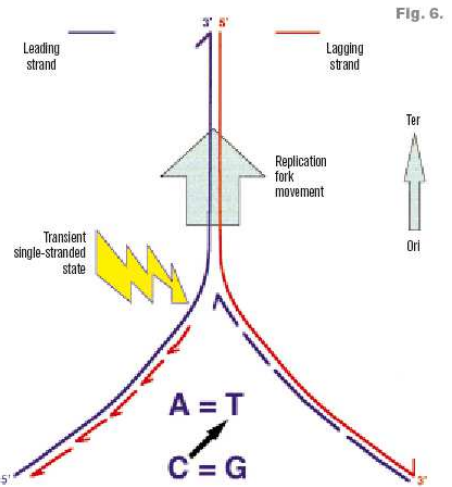
Chirochores in *B. burgdorferi* (third codon positions)



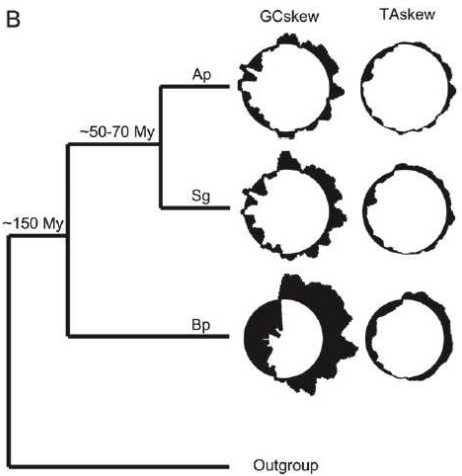
Usually GC skew > AT skew *e.g. E. coli*



Cytosine deamination theory



Testing the hypothesis: *Buchnera aphidicola*



Klasson, L. & Anderson, S.G.E. (2006) *Mol. Biol. Evol*, **23**:1031-1039.

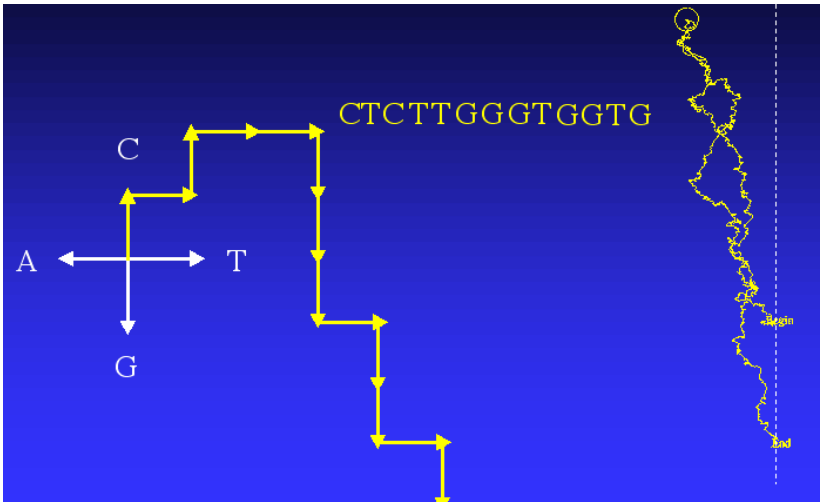
Chirochore practical: global skew

Study this yourself with the complete genome from *Chlamydia trachomatis*. To do this, use the package `seqinR` in  .
The zipped fasta file for the genome is accessible at this address:
`http://pbil.univ-lyon1.fr/members/mbailly/AMIG/data/ct.zip`

Start by measure its genome G-C and A-T *global* skew. How close is this genome from PR2? Make a graph representing it.

Chirochore practical : one example graph

What is a DNA walk?



Chirochore practical: local skew

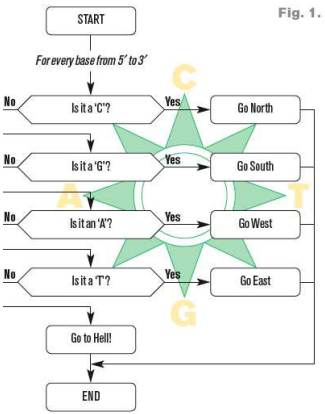


Fig. 1.

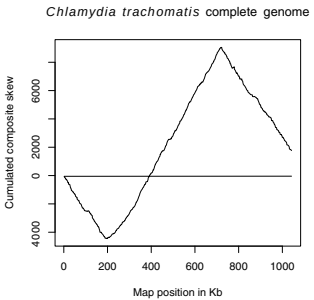
Make a simple DNA walk on this genome. By example use the functions `ifelse()`, `cumsum()` and plot a point *every Kb* with the same scale (also in Kb) for both axes.

Chirochore practical: the DNA walk

Chirochore practical

Chirochores practical

Read the documentation of the `oriloc()` function and apply it to your data[1].



Could you make the same type of figure using your previous results?

[1] Try applying it to the downloaded files, not the library ones!