The median problem for the reversal distance in circular bacterial genomes

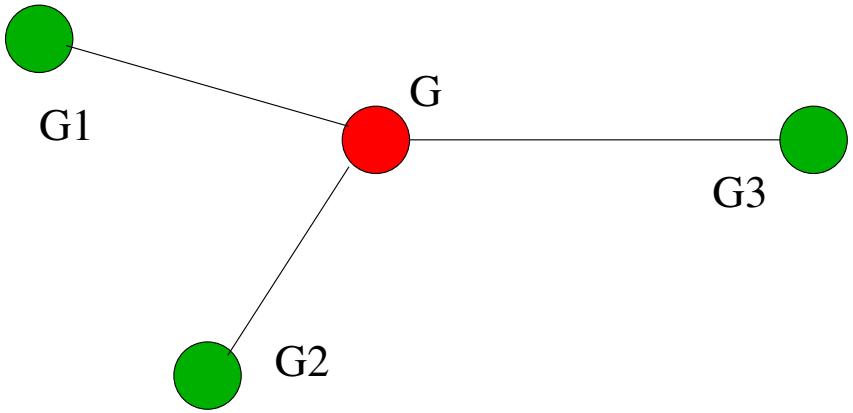
E. Ohlebusch, M.I. Abouelhoda, K. Hockel, J. Stallkamp

University of Ulm, Germany

CPM 2005

Median Problem

Given 3 genomes G_1, G_2, and G_3, find a genome G such that $d_m = \sum_{i=1}^{3} d(G, G_i)$ is minimized for a distance measure d.

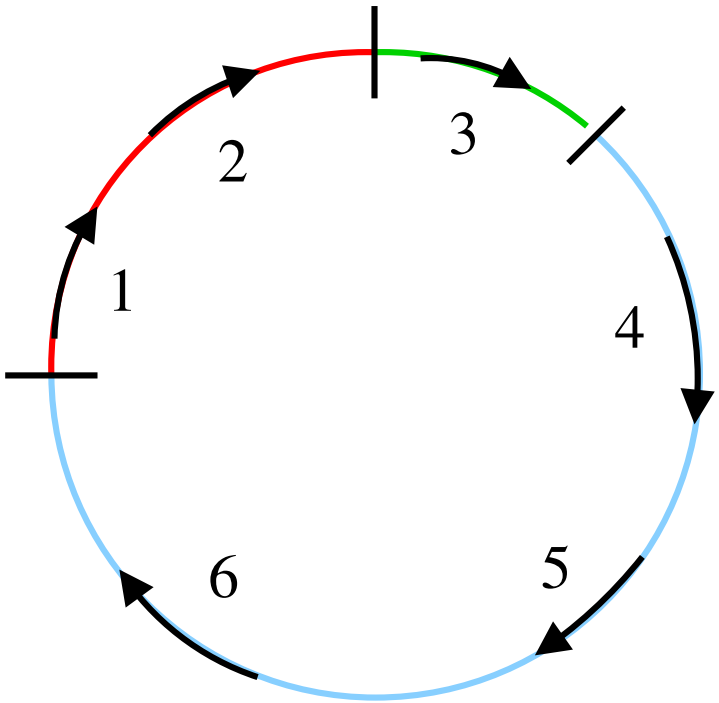


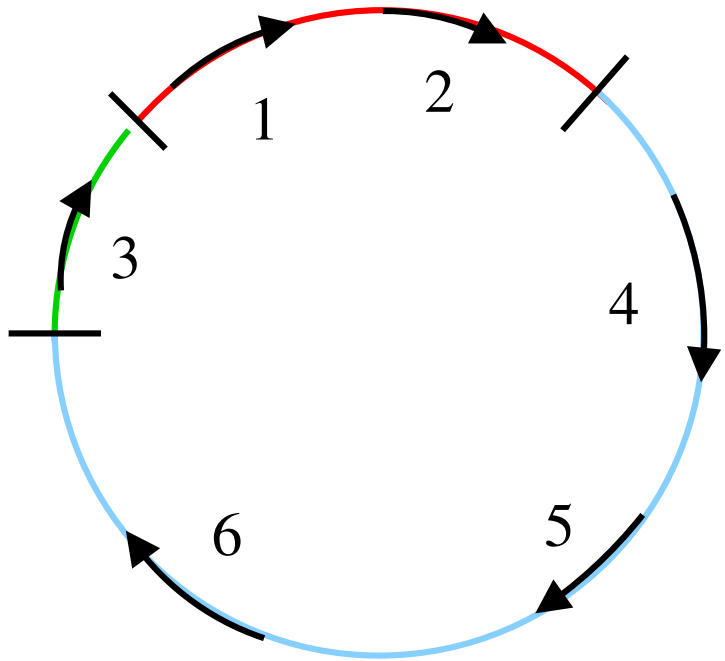
Needed: distance between two genomes $G = (\pi_1, \ldots, \pi_n)$ and $G' = (\rho_1, \ldots, \rho_n)$ on the same set of genes $\{1, \ldots, n\}$

Rearrangements

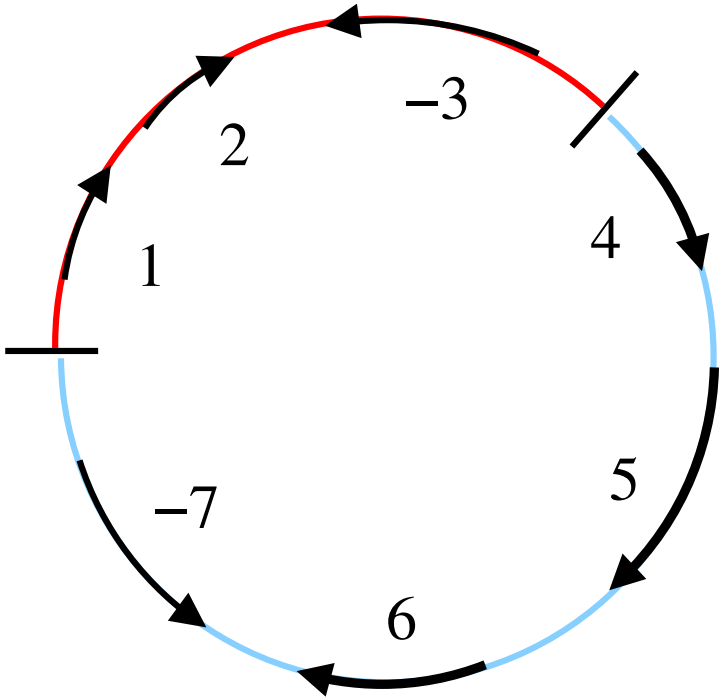
- genomes are subject to rearrangements
- less frequent than local changes
- information about the evolutionary distance between genomes
- affect large parts of the DNA
- change the order / orientation of involved genes

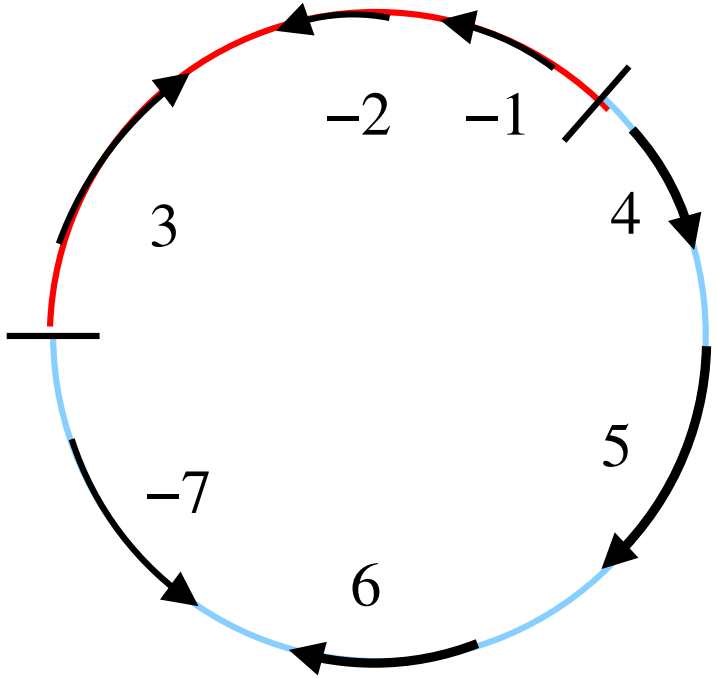
Example: Transposition





Example: Reversal





Rearrangement Distance

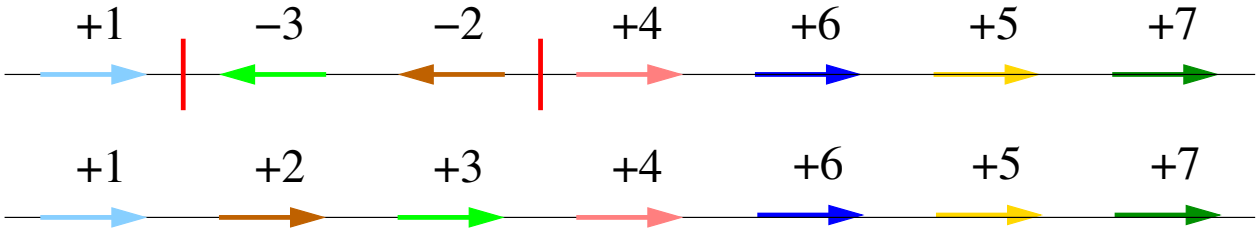
- minimum number of rearrangements needed to transform genome G into genome G'
- advantage: good estimation of evolutionary distance
- drawback: complexity not known; we can't compute it efficiently [*Hartman2003*]

Reversal Distance

- minimum number of reversals needed to transform G into G'
- advantage: can be computed in $\mathbf{0}(n)$ [*Bader*, *Moret*, *Yan2001*; *Bergeron*, *Mixtacki*, *Stoye2004*]
- drawback: other operations are not considered (e.g. transpositions)

Breakpoints

- $G = (\pi_1, \ldots, \pi_n)$, $G' = (\gamma_1, \ldots, \gamma_n)$ on the same set of genes $\{1, \ldots, n\}$
- two genes π_i π_{i+1} determine a breakpoint in G w.r.t $G' \Leftrightarrow$ neither π_i precedes π_{i+1} nor $-\pi_{i+1}$ precedes $-\pi_i$ in G'
- example:



Breakpoint Distance

- number of breakpoints between two genomes/permutations
- advantage: easy to compute
- draw back: only rough estimation of number of rearrangements
  [*Moret*, *Siepel*, *Tang*, *Liu2002*]

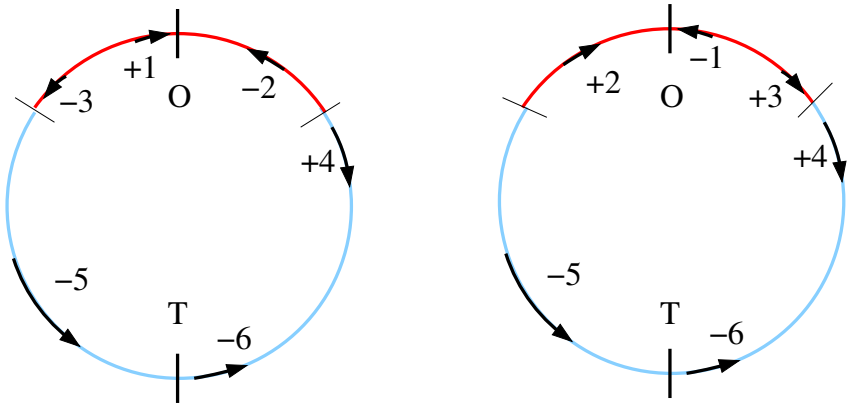
Bad and Good News

The median problem is NP-hard for both the breakpoint and the reversal distance!
[*Caprara1999*; *Pe′er*, *Shamir1998*]

Using biological constraints can simplify the problem significantly.

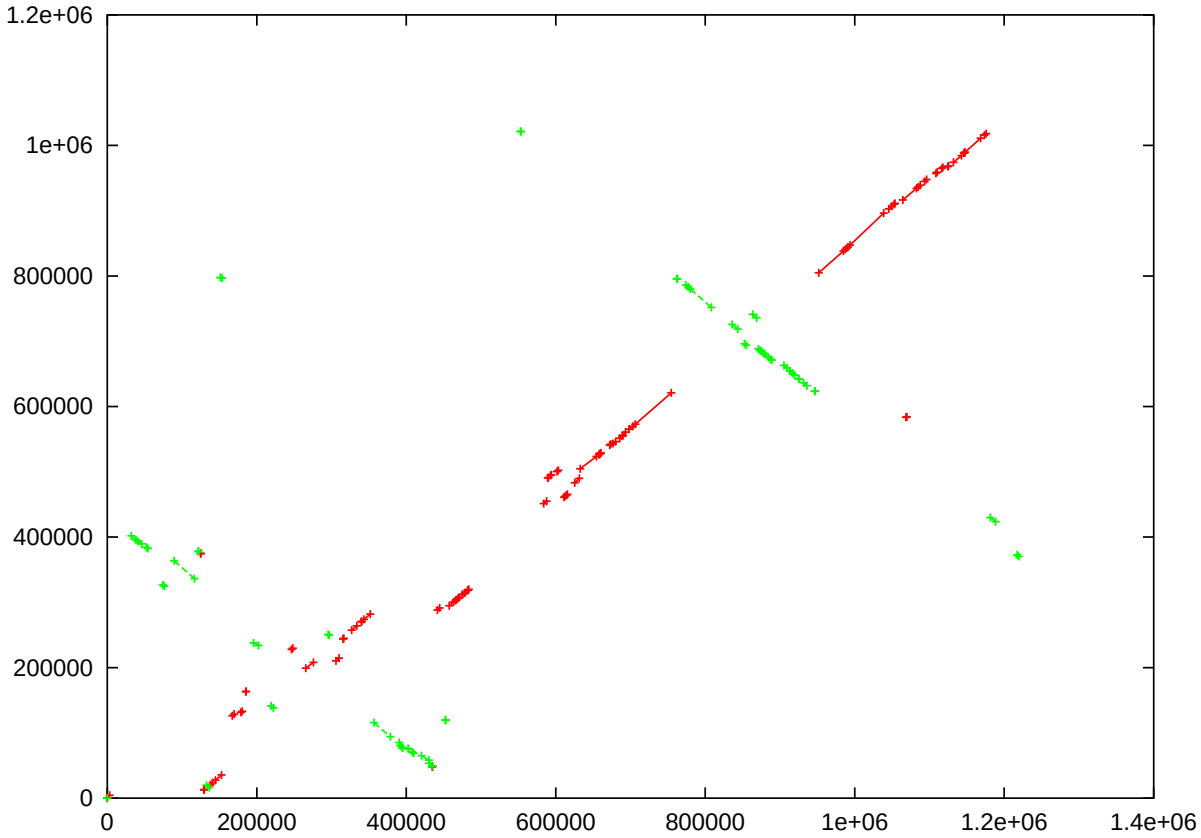
Circular Bacterial Genomes

Predominant: reversals around the origin/terminus of replication [*Eisenetal.2000*; *Tiller*, *Collins2000*]



- $\overline{\rho}(3)$:reversal centered around origin (analogous: $\underline{\rho}(i)$)
- genes keep their distance to origin/terminus
- genes change their orientation

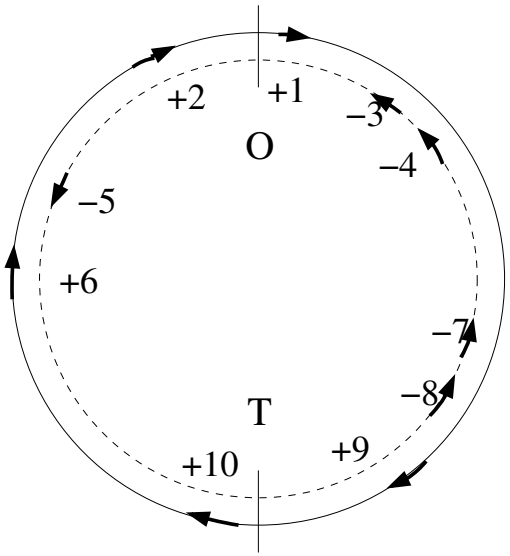
Example: Chlamydiae (pneumoniae, trachomatis)



Genome Representation

- bit vector:
  - 1: right side
  - 0: left side
- orientation vector:
  - $+$: forward, if right hand side; reverse, if left hand side
  - $-$: reverse, if right hand side; forward, if left hand side
- representation of genome by bit vector

Genome as Bit Vector



$$(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)$$

- bit vector: $(1, 0, 1, 1, 0, 0, 1, 1, 1, 0)$
- orientation vector: $(+, -, -, -, +, -, -, -, +, -)$

Only around Origin

procedure $rd\_O(G, G')$
determine the breakpoints $(i_1, i_1 + 1), \ldots, (i_k, i_k + 1)$ between G and G'
if $G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k) = G'$ **then** return k **else** return $k + 1$

Correctness

- reversal $\rho(i)$ doesn't change any existing breakpoints except at position $(i, i+1)$
- $(i, i+1)$ breakpoint $\Rightarrow$ $\rho(i)$ removes this breakpoint
- $(i, i+1)$ NO breakpoint $\Rightarrow$ $\rho(i)$ creates a new breakpoint

Some Definitions

Definition

Let $G = (b_1, b_2, b_3, \ldots, b_n)$ and $G' = (b'_1, b'_2, b'_3, \ldots, b'_n)$ be two circular genomes.
An interval $[i..j]$ of indices (where $1 \leq i \leq j \leq n$) is called a *strip* if $b_k = b'_k$ for all $i \leq k \leq j$, $b_{i-1} \neq b'_{i-1}$ if $i \neq 1$, and $b_{j+1} \neq b'_{j+1}$ if $j \neq n$.

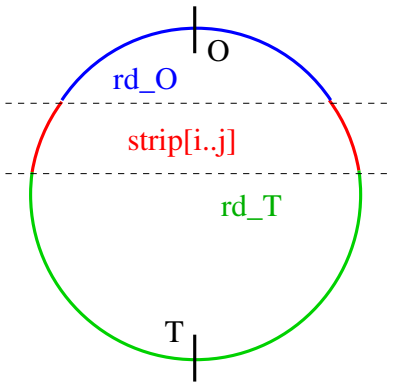
Definition

Let $b^1,\ b^2,\ b^3 \in \{0, 1\}$.

$$\text{majority}\left(b^1, b^2, b^3\right) = \begin{cases} 1 & \text{if } \sum_{j=1}^{3} b^j \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

Reversal Distance

procedure $rd(G, G')$
  if G and G' do not have a breakpoint **then**
    if $G = G'$ **then** return 0 **else** return 1
  else
    choose a strip $[i..j]$
    $k_l := rd\_O(G[1..i-1], G'[1..i-1])$
    $k_r := rd\_T(G[j+1..n], G'[j+1..n])$
    return $(k_l + k_r)$



The Problem

- Input: 3 genomes G_1, G_2 and G_3, represented by their bitvectors
- Output: median G, which minimizes $d_m = \sum_{i=1}^{3} rd(G, G_i)$
- Restrictions:
  - same set of genes in all 3 genomes
  - only reversals around origin / terminus of replication
- can be computed in $\mathbf{O}(n)$

Around Origin Only

procedure $median\_O(G_1, G_2, G_3)$ $\quad / \star\ G_j = (b_1^j, b_2^j, b_3^j, \ldots, b_n^j)\ \star /$

$d := 0$

for $i := n$ **downto** 1 **do**

$b := majority(b_i^1, b_i^2, b_i^3)$

if there is a j, $1 \leq j \leq 3$, such that $b_i^j \neq b$ **then**

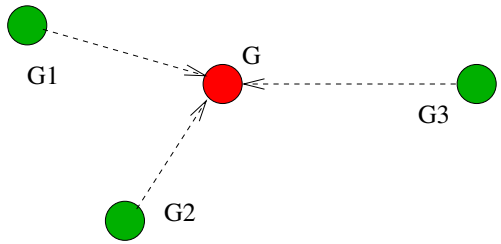
$G_j := G_j\overline{\rho}(i)$

$d := d + 1$

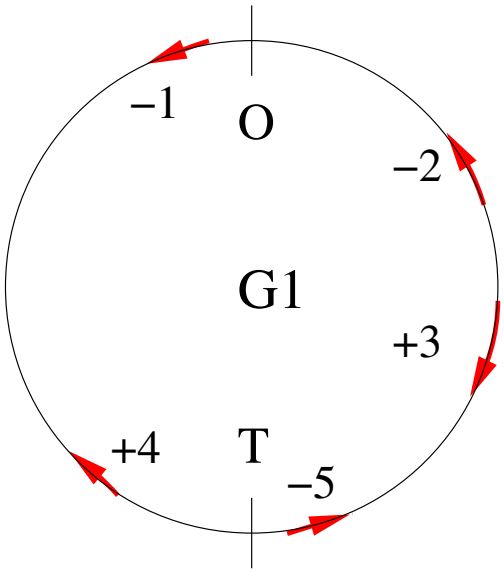
return (G_1, d)

Definition

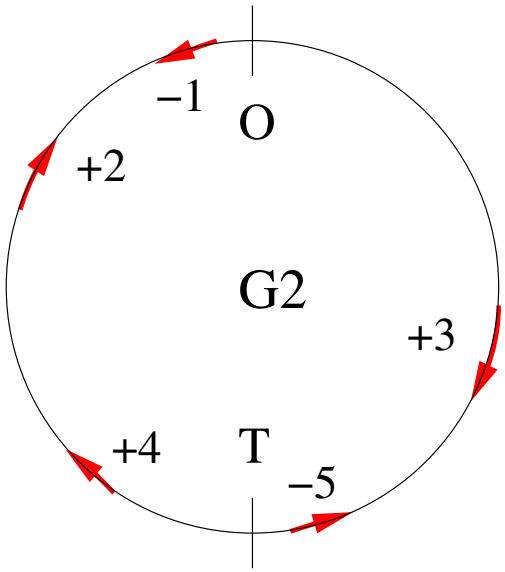
$$\text{majority}\left(b^1, b^2, b^3\right) = \begin{cases} 1 & \text{if } \sum_{j=1}^{3} b^j \geq 2 \\ 0 & \text{otherwise} \end{cases}$$



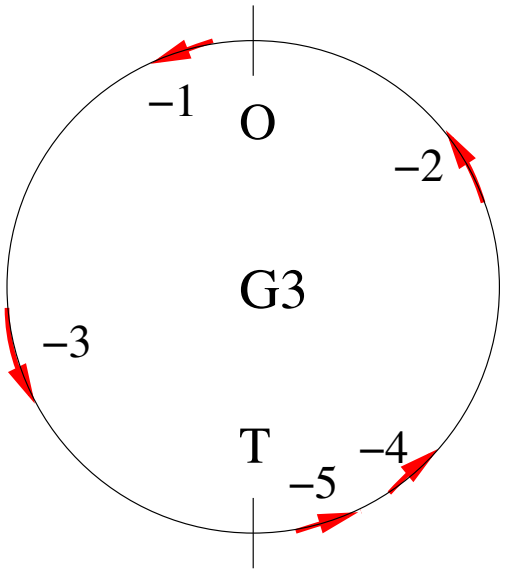
Around Origin Only: Example



0 1 1 0 1

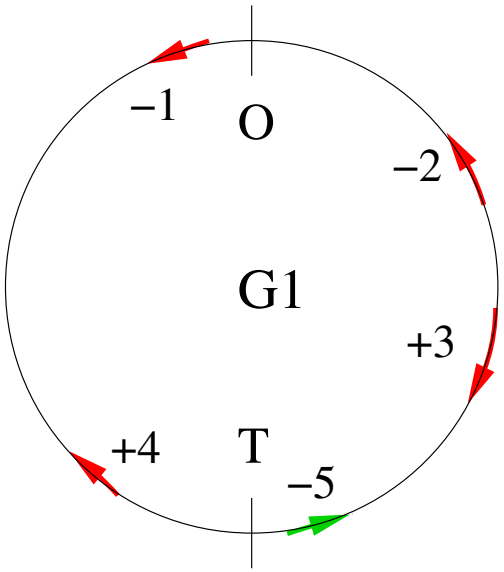


0 0 1 0 1

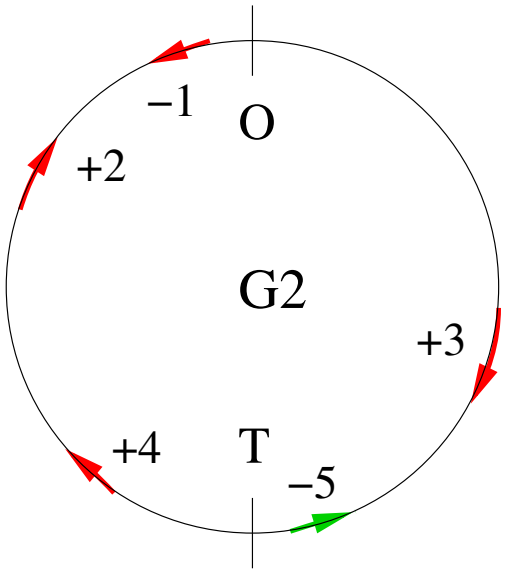


0 1 0 1 1

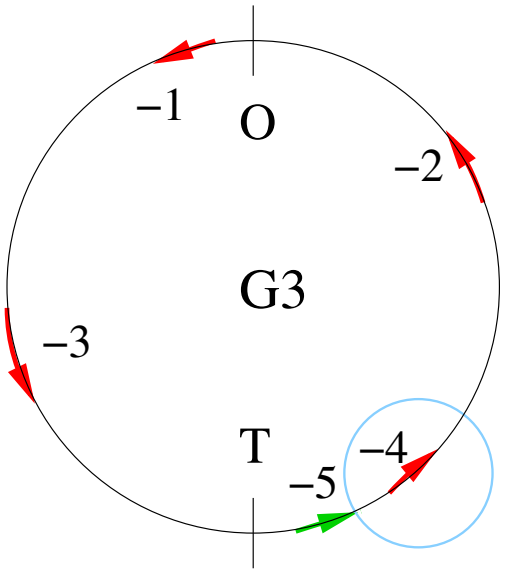
Around Origin Only: Example



0 1 1 0 1

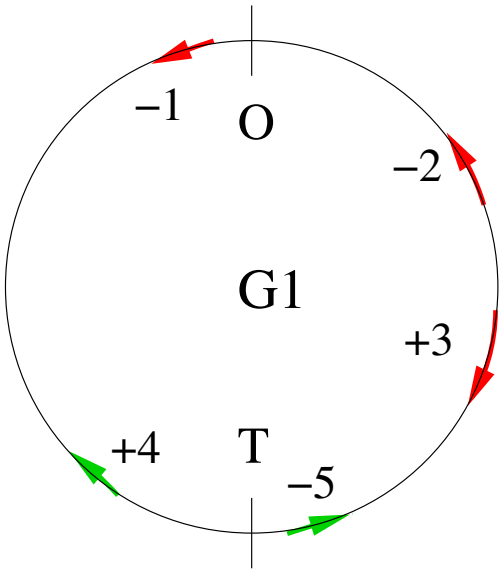


0 0 1 0 1

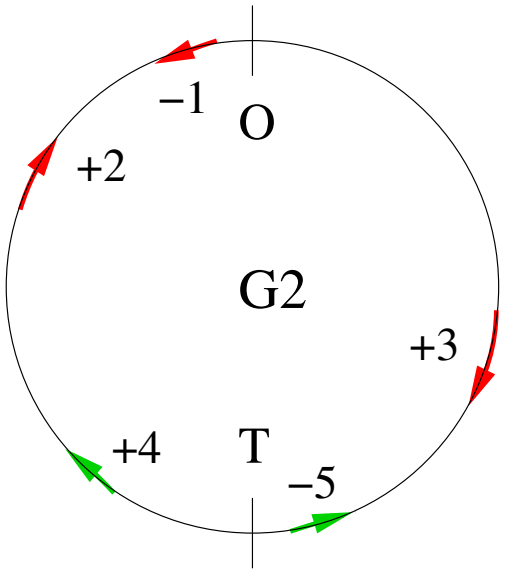


0 1 0 1 1

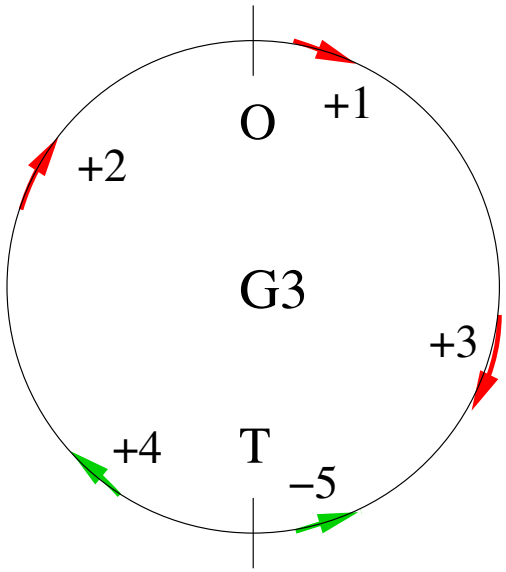
Around Origin Only: Example



0 1 1 0 1

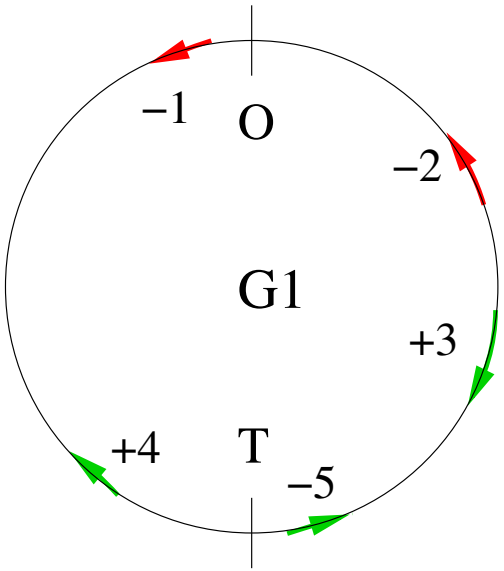


0 0 1 0 1

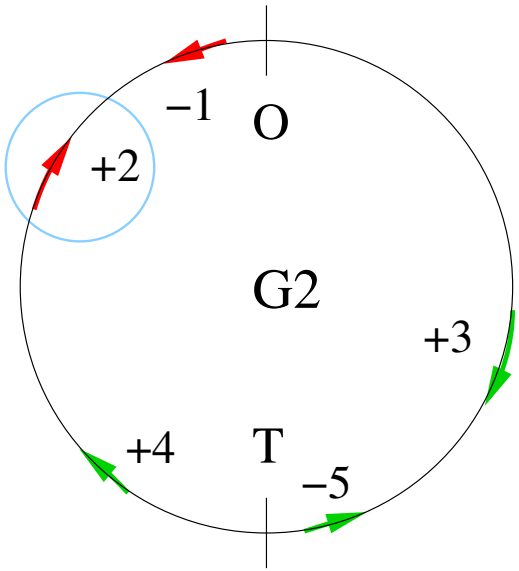


1 0 1 0 1

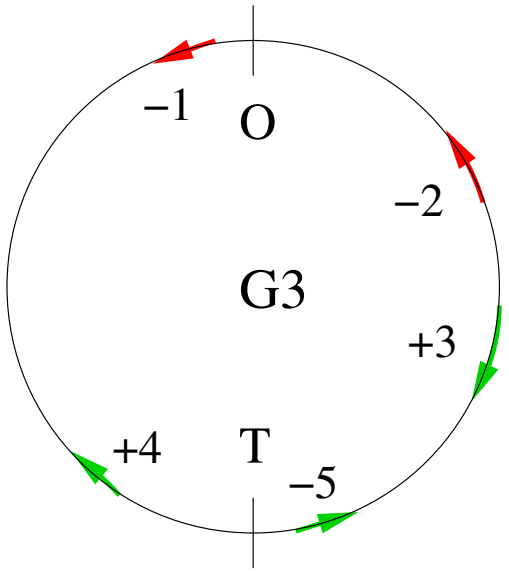
Around Origin Only: Example



0 1 1 0 1

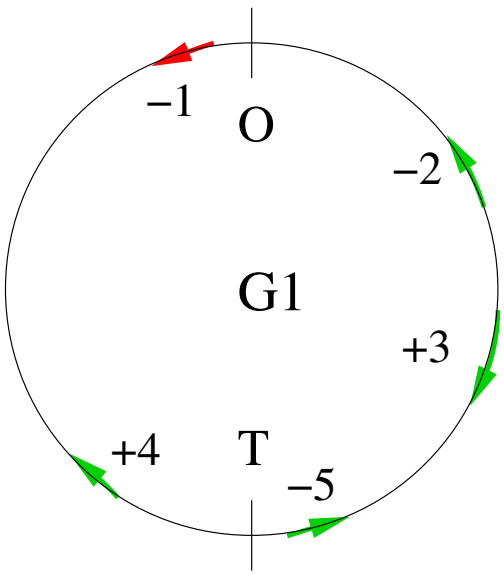


0 0 1 0 1

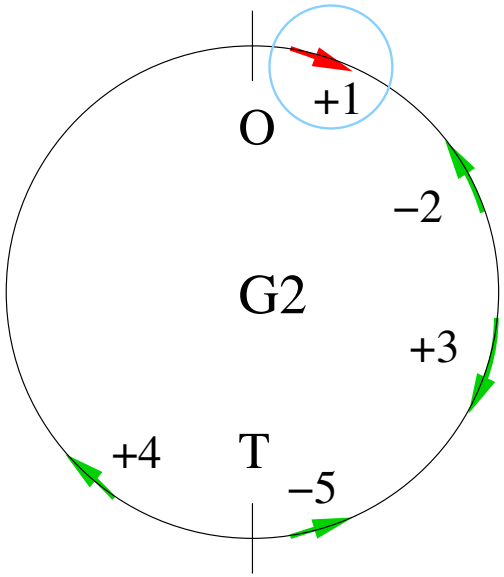


0 1 1 0 1

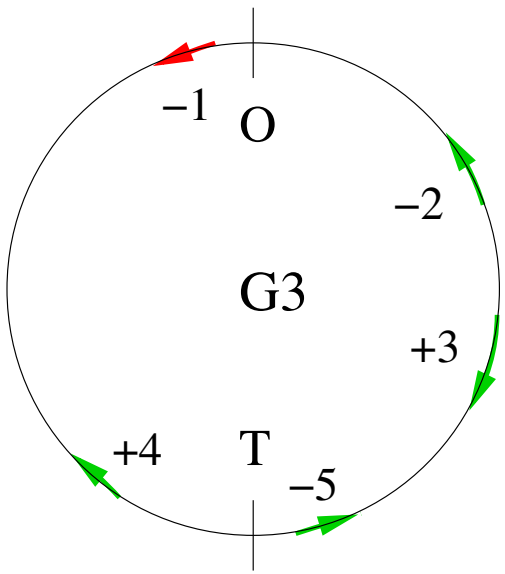
Around Origin Only: Example



0 1 1 0 1

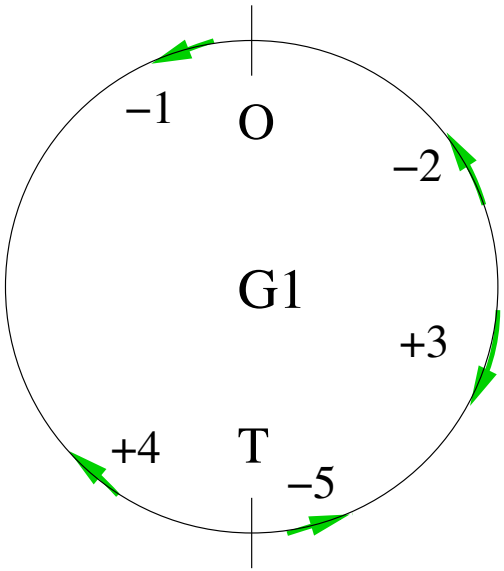


1 1 1 0 1

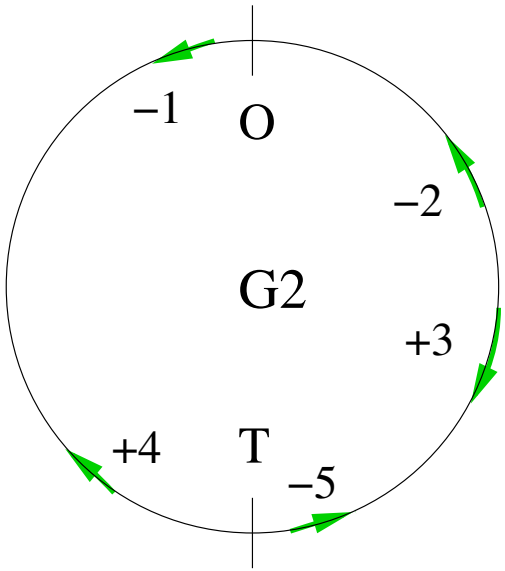


0 1 1 0 1

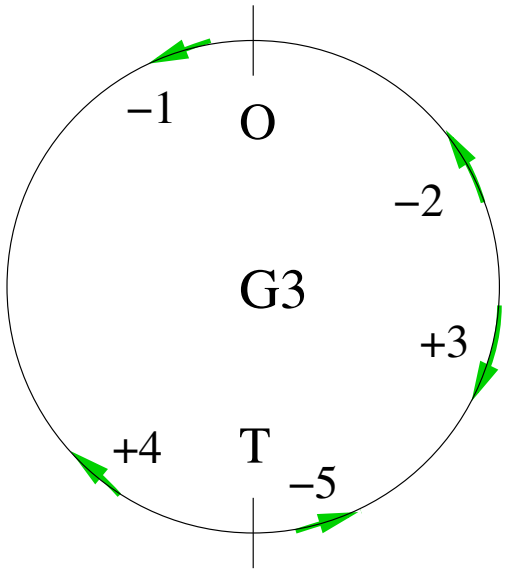
Around Origin Only: Example



0 1 1 0 1



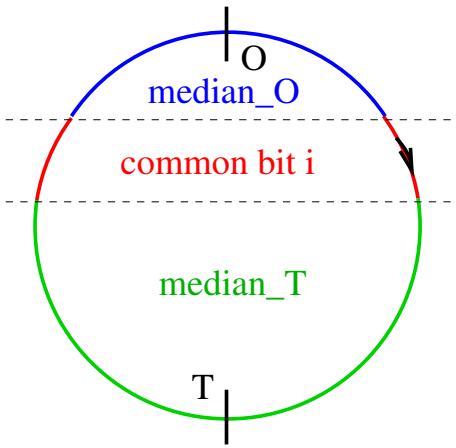
0 1 1 0 1



0 1 1 0 1

With Common Bit

procedure $median\_cb(G_1, G_2, G_3)$
determine a common bit i of G_1, G_2, and G_3
$(G_l, d_l) := median\_O(G_1[1..i-1], G_2[1..i-1], G_3[1..i-1])$
$(G_r, d_r) := median\_T(G_1[i+1..n], G_2[i+1..n], G_3[i+1..n])$
return $(G_l G_1[i] G_r, d_l + d_r)$



Without Common Bit

procedure $median\_ncb(G_1, G_2, G_3)$
  if two genomes coincide, say $G_i = G_j$ with $i \neq j$ **then** return $(G_i, 1)$
  else if one of the genomes is the inverse of another,
  say $G_i = inv(G_j)$ with $i \neq j$
    then return $(G_i, 1 + rd(G_i, G_k))$ where $k \in \{1,2,3\} \setminus \{i,j\}$
  else $/\star$ $G_i \neq G_j$ and $G_i \neq inv(G_j)$ for all $i \neq j$ $\star/$
    $(G', d') := median\_cb(inv(G_1), G_2, G_3)$
    $d'_1 := rd(inv(G_1), G_2) + rd(inv(G_1), G_3)$
    if $d'_1 = d'$ **then** return (G_1, d')
    else return (G', d')

Summary

- general median problem is NP-hard for both the reversal and the breakpoint distance
- circular bacterial genomes: reversals are centered around origin / terminus of replication
- using this biological constraint leads to an $\mathbf{O}(n)$ algorithm

Future Work

- shortcomings:
  - position of origin / terminus has to be known
  - restriction to reversals
- future work:
  - including tranpositions
  - including reversals affecting only one single gene (any position) [*Lefebvre*, *El – Mabrouk*, *Tillier*, *Sankoff2003*]

Thanks

- to my supervisor Prof. Ohlebusch
- to my colleagues Mohamed I. Abouelhoda and Jan Stallkamp
- to you for your attention