Highways and byways in group-theoretic genome space

Attila Egri-Nagy, joint work with Andrew Francis and Volker Gebhardt

Centre for Mathematics Research,
School of Computing, Engineering and Mathematics
University of Western Sydney

Phylomania 2013

Questions

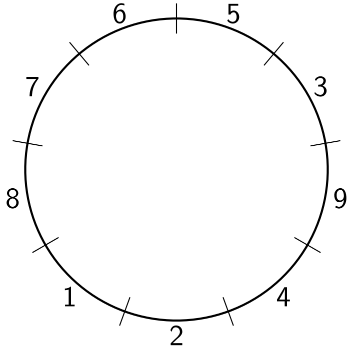
Is the distance a good enough measure?

Can we use the number of shortest evolutionary paths?

Maybe the 'shape' how these paths are put together...

Biology→Math

Genome→ permutations



Genomic distance→ Length of geodesic words

Groups, generator sets

Let G be a group with generators $S = \{s_1, \ldots, s_n\}$.

S^* is the set of all finite sequences, *words* of the elements of S. The group element realized by the word w is denoted by $\overline{w}$, thus $w \in S^*$ and $\overline{w} \in G$.

Example

$S = \left\{s_1 = (1,2), s_2 = (2,3)\right\}$

$s_1 s_2 s_1 s_2 = (1,2)(2,3)(1,2)(2,3) = (1,2,3)$

So $\overline{s_1 s_2 s_1 s_2} = (1,2,3)$.

sequences of generators $\Longleftrightarrow$ sequences of events

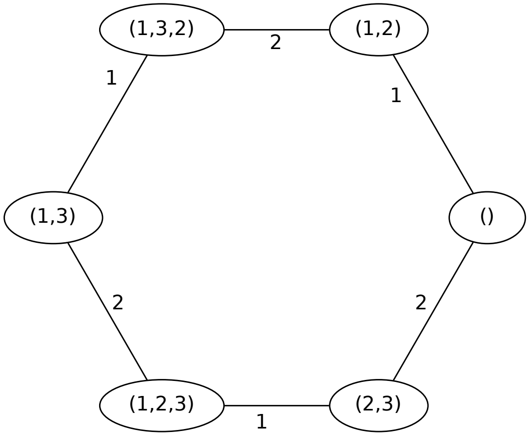
Cayley graph

The *Cayley graph* $\Gamma(G, S)$ of G with respect to the generating set S is the directed graph with group elements as nodes and the labeled edges encoding the action of G on itself. Thus $g \xrightarrow{s} gs$ is an edge.

Cayley graph of S_3

Example

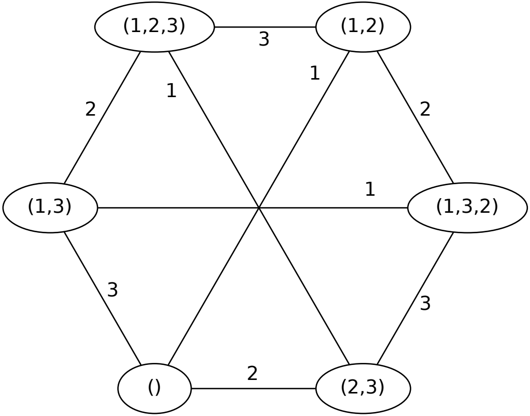
$S = \{s_1 = (1,2), s_2 = (2,3)\}$



Cayley graph of S_3 – different generators

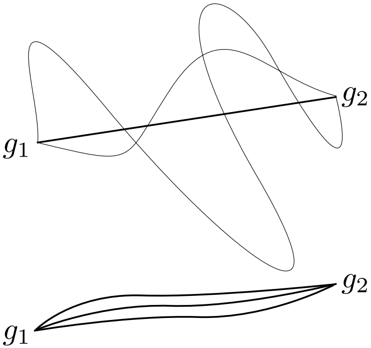
Example

$S = \{s_1 = (1,2), s_2 = (2,3), s_3 = (3,1)\}$



Geodesic distance, shortest path

The *geodesic distance* defined by $d_S(g_1, g_2) = |u|$, where u is a minimal length word in S^* with the property that $g_1\overline{u} = g_2$ also denoted by $g_1 \xrightarrow{u} g_2$, and u is called a *geodesic* word. $\mathrm{Geo}_S(g_1, g_2)$ is the set of all geodesic words from g_1 to g_2.



What is $Geo_S(g_1, g_2)$?

A partial order defined by the geodesics

Due to a translation principle we can simpy write $\ell(g)$ instead of $d(1, g)$. Similarly, we use $\mathrm{Geo}(g)$ instead of $\mathrm{Geo}(1, g)$.

Definition

For group elements $g_1, g_2 \in G = \langle S \rangle$ we write $g_1 \leq g_2$ if $\exists w = uv \in S^*$ such that $\overline{w} = g_2, \overline{u} = g_1, w \in \mathrm{Geo}(g_2)$, i.e. there is a geodesic from the identity to g_2 and g_1 is on it.

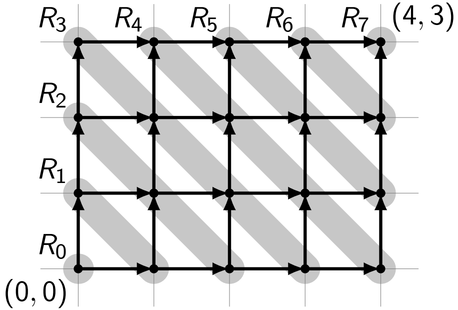
Also called the *prefix* order, or *weak* order for Coxeter groups.

Intervals

With the partial order closed intervals are defined in the obvious way

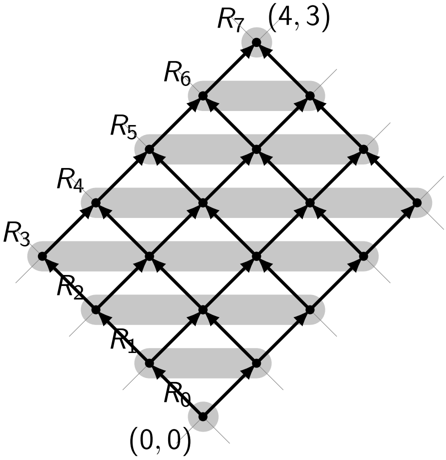
$$[1, h] := \{g \in G \mid 1 \leq g \leq h\}$$

Ranked poset



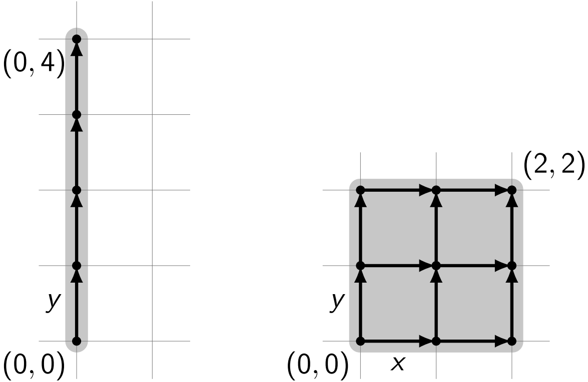
The rank-sets of the interval $\big[(0,0),(3,4)\big]$ in $\mathbb{Z}\times\mathbb{Z}$.

Ranked poset



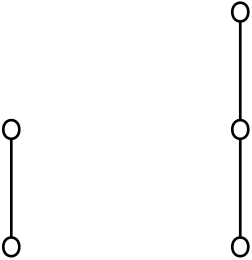
Length and size

In general there is no connection.

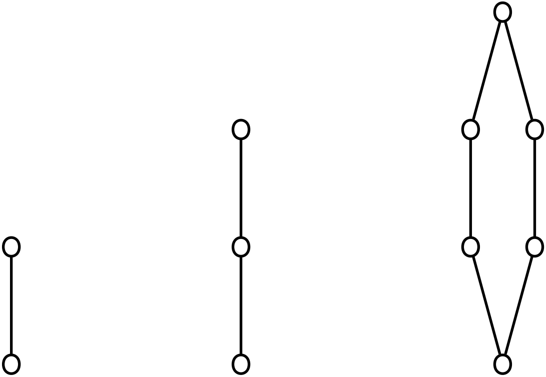


In $\mathbb{Z}^2$ two group elements with same length can have intervals of different size. $\left|[(0,0),(0,4)]\right| = 5$, $\left|[(0,0),(2,2)]\right| = 9$.

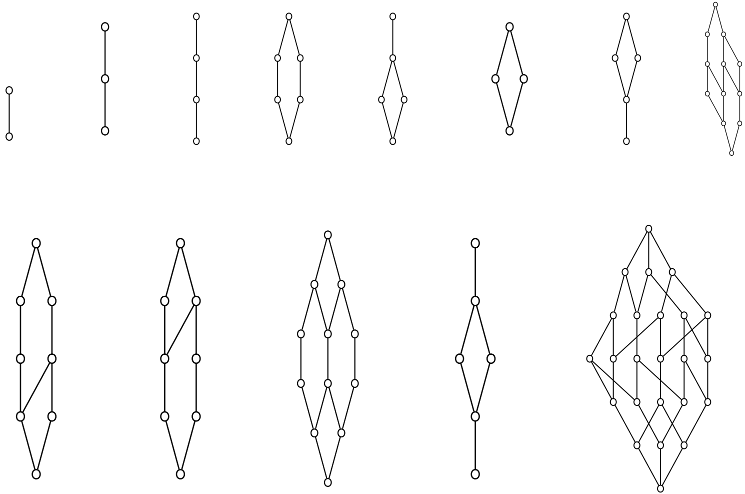
Interval lattices in $S_3 = \langle (1,2,3), (1,2) \rangle$



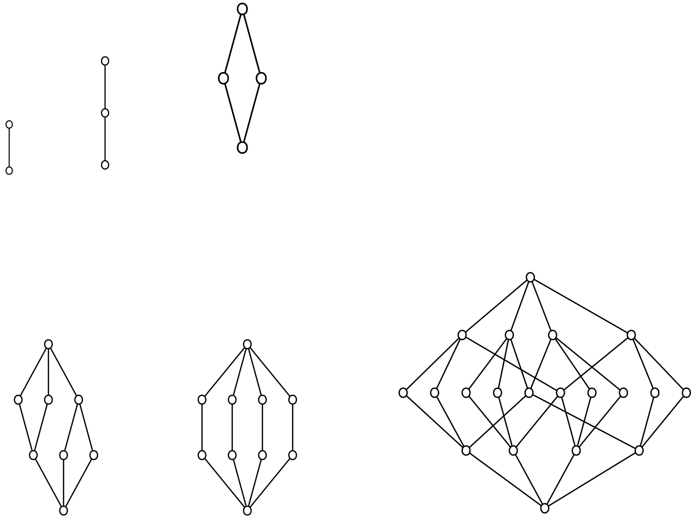
$S_3 = \langle (1,2),(2,3) \rangle$



$S_4 = \langle (1,2), (2,3), (3,4) \rangle$

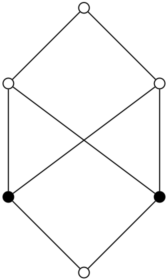


$S_4 = \langle (1,2), (2,3), (3,4), (1,4) \rangle$

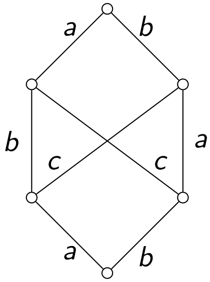


Is it a lattice?

An obvious mathematical but biologically not so relevant question.
A minimal counterexample would be:



Trying with involutions



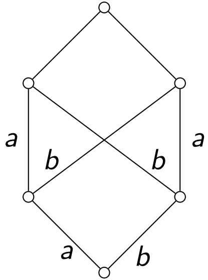
$$ab = bc = ca,$$
$$ac = ba = cb.$$

But since they are involutions,

$$ba = cb \implies c = bab$$

Trying it with 2 generators

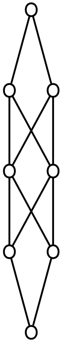
Minimal counterexamples



$a^2 = b^2$, $ab = ba$

For instance, $a = (3, 4, 5)$, $b = (1, 2)(3, 4, 5)$.

$C_4 \times C_2 = \langle (3,4,5,6), (1,2)(3,4,5,6) \rangle$



$[(), (1,2)]$

Sperner property?

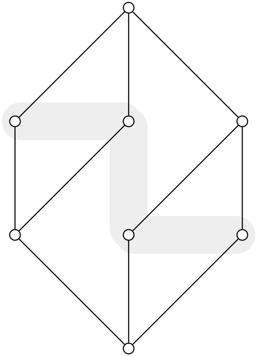
Sperner property: no antichain is bigger than the size of the maximal rank-set.

Do these intervals have the Sperner property?

Sperner property?

Sperner property: no antichain is bigger than the size of the maximal rank-set.

Do these intervals have the Sperner property? NO.



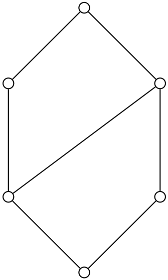
$$s_4 s_3 s_1 = s_4 s_1 s_3 = s_3 s_1 s_2 = s_1 s_3 s_2$$

Anti-chains

Do anti-chains give the number of paths?

Anti-chains

Do anti-chains give the number of paths? NO.



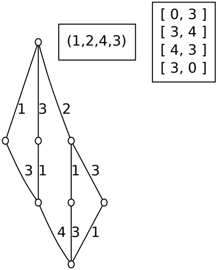
Possible equivalence relations

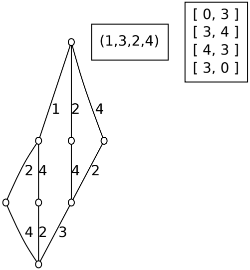
The ultimate goal is to find equivalence classes of group elements.

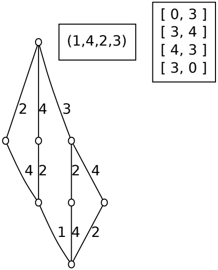
1. Same length: $\ell(g_1) = \ell(g_2)$.
2. Same 'width': $|\,\mathrm{Geo}(g_1)| = |\,\mathrm{Geo}(g_2)|$. Probably the most decisive property for the biological application.
3. Same profile.
4. Same interval.

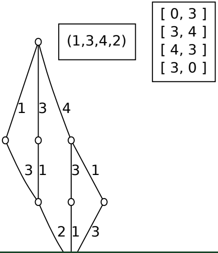

(1,4,2,3)
(1,3,4,2)
[0, 2] [2, 2] [2, 0]
(1,2,4,3)
[0, 2] [2, 2] [2, 0]
(2,3,4)
(1,2,3)
[0, 3] [3, 3] [3, 0]
(1,2)
[0, 1] [1, 0]
(1,3)
[0, 1] [1, 0]
(2,3)
[0, 1] [1, 0]
(1,2)(3,4)
[0, 1] [1, 0]
(2,4)
(1,4)
(3,4)
(1,4)(2,3)
[0, 1] [1, 0]
()
[0, 0]

$S_4 = \langle (1,2),(2,3),(3,4),(1,4) \rangle$

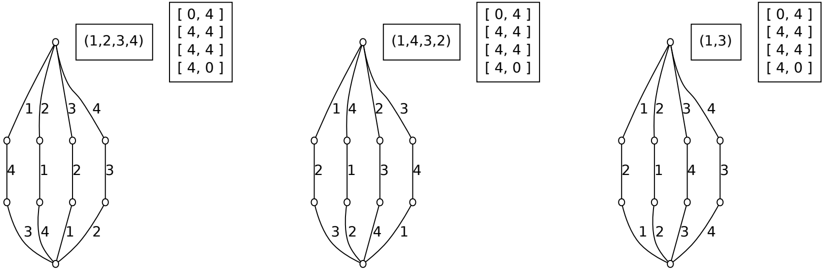








$S_4 = \langle (1,2), (2,3), (3,4), (1,4) \rangle$



$n = 5$	all inversions	circular	linear
length	4	7	11
[length,width]	7	14	30

Number of paths

Assuming that we have an efficient algorithm for calculating the distance, we can also calculate the interval.

For biological applications it is probably enough to estimate the interval by partially calculating it.

Algorithm 1: Constructing the graded interval $[g, h]$.

input : $g, h \in G$, S generator set, d distance function
output: $[g, h]$ interval, R_i rank-sets

```
GradedInterval (g, h, S, d):
n ← d(g, h);
R_0 ← {g};
foreach i ∈ {1, ..., n} do
    R_i ← ∅;
    foreach g' ∈ R_{i-1} do
        foreach s ∈ S do
            if d(g's, h) = n - i then
                R_i ← R_i ∪ g's;
```

TODO list

- Study individual generating sets. (since no grand theory is available)
- Find the right interpretation in order to modify the distance function.

Thank You!