#10
Genome-wide Comparison

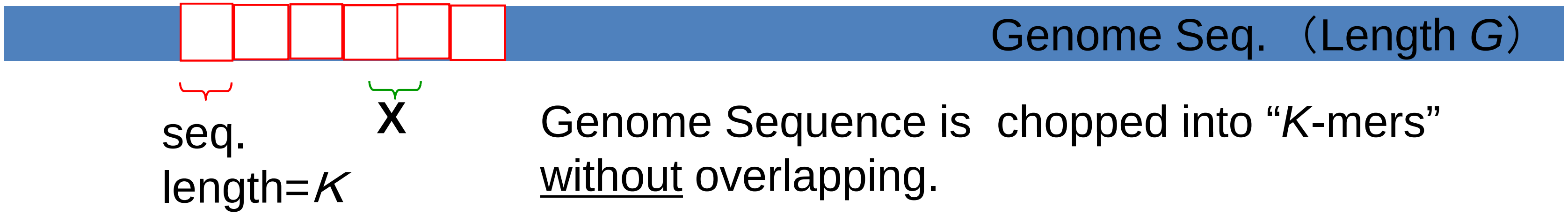
Topics:

- BLAT
- 2-D Dot Plot
- Edit Distance between Genome Sequences
  - Inversion, Edit Distance,
  - Comparing X chromosome of human and mouse
  - Graph representation (Reality and Desired graph)
  - Independent Alternative Cycles

BLAT

Fast comparison of DNA sequences versus a genomic DNA.
Developed by James Kent (UCSC).
Target genome DNA sequence is pre-processed and a huge index table is prepared.
In some cases, about 500-times faster than BLAST.

James Kent



Genome Sequence is chopped into “*K*-mers” <u>without</u> overlapping.

AAAAA →	1, 1012, 2245, 4560, …
AAAAC →	2, 2246, 3135, 5235, …
AAAAG →	
AAAAT →	
·	
·	
·	
TTTTT →	

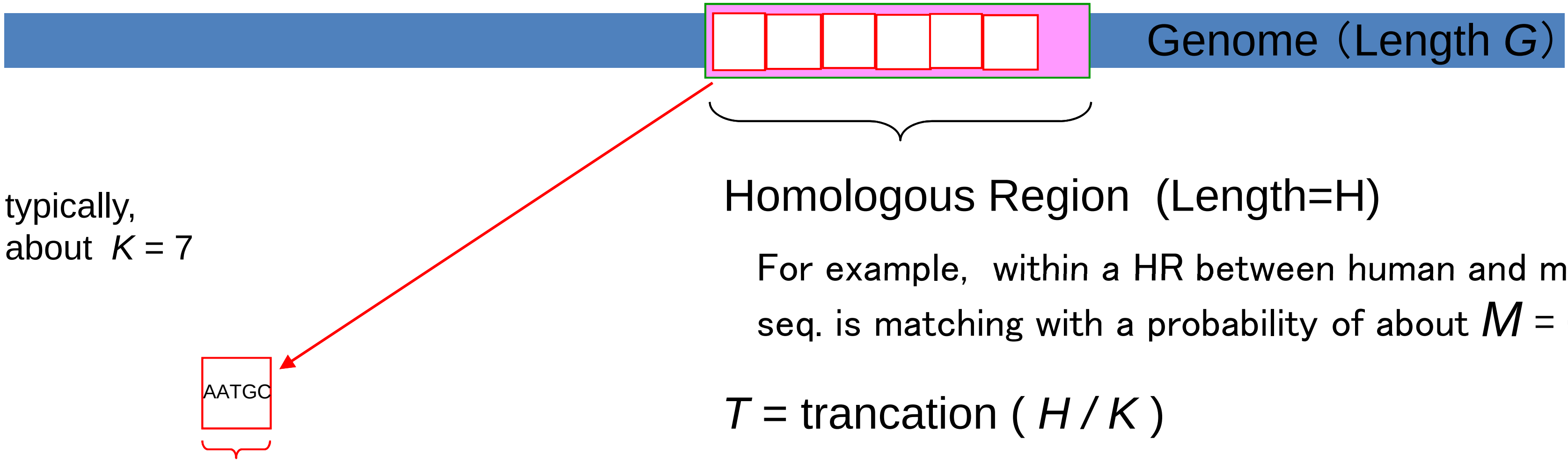
(*G* / *K*) - subsequences are stored in an index table (like left fig.).

Query input sequence is searched against this table.

Approximation:

1) search “exact match” only
2) *K*-mer with another boundary (like subseq. X) is not subject to search

BLAT（2）



Homologous Region (Length=H)

typically,
about $K = 7$

For example, within a HR between human and mouse, seq. is matching with a probability of about $M = 0.98$.

T = trancation (H / K)

T pieces of K−mer exist in a HR

K-mer

$$P_1 = M^K$$

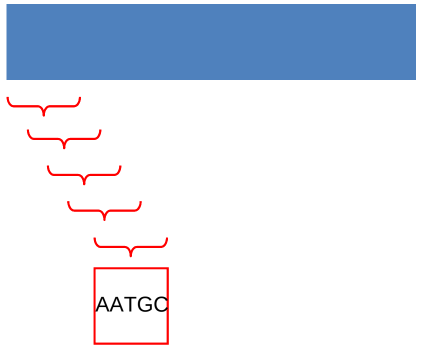
Probability of having at least one exact match of *K*-mer in HR

$$P = 1 - (1 - P_1)^T = 1 - (1 - M^K)^T$$

**If any one exact match with K-mer is discovered,
BLAT assumes the hit is within a homologous region and start detailed search around the hit block.**

BLAT（3）

Query seq. (length=Q)



K-mer

From a query sequence, all K-mers with overlapping are examined. Then frequency of random hit is about

$$F = (Q - K + 1) \times (G / K) \times (1 / 4)^{K}$$

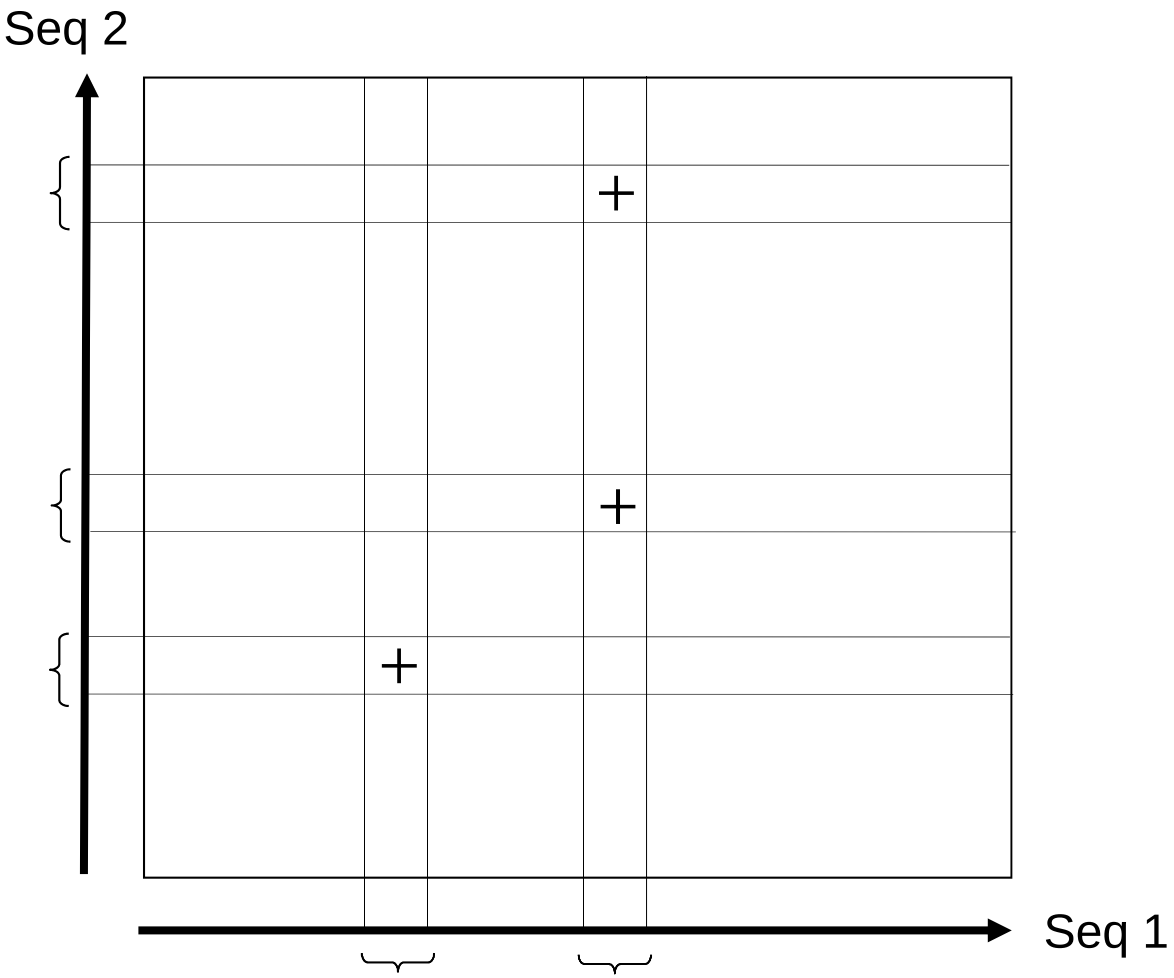
Too small *K* value brings many noisy hits.
Too large *K* value leads to miss important HR.

Alternative 1: Allow 1-miss match in K-mer (not exact K-mer match)

Alternative 2: Request to have N (for example, N=2) K-mer exact matches in HR. Use relatively small K value, but use N > 1 for balancing.

Memo: P_1 is the probability of observing one **random** hit within a HR.
The probability of observing multiple N hits within HR (T blocks) is
binomial distribution $P_n = {}_{T}C_n \times P_1^{\,n} \times (1 - P_1)^{T-n}$
The P-value of having N (or more) hits is
$P(x >= N) = P_N + P_{N+1} + \dots + P_T$
Choose appropriate K and N values to have small enough P-value.

2-D Dot Plot



Compare two sequences with a **fixed-length window** (for example K=7, K=29)
Put a mark (+) or dot (･) with a place of exact match between two sequences.
For DNA sequences, “**reverse complimentary strand**” is simultaneously examined.

2-D Dot Plot

Genome-wide comparison

horizontal axis: MED4

(prochlorophytes 原核緑藻, surface type)

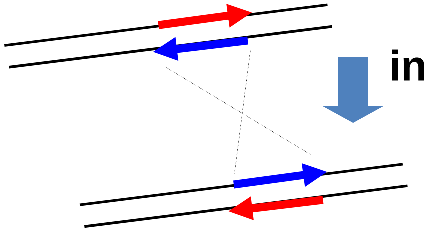
vertical axis: MIT9313

(prochlorophytes 原核緑藻, deep sea)

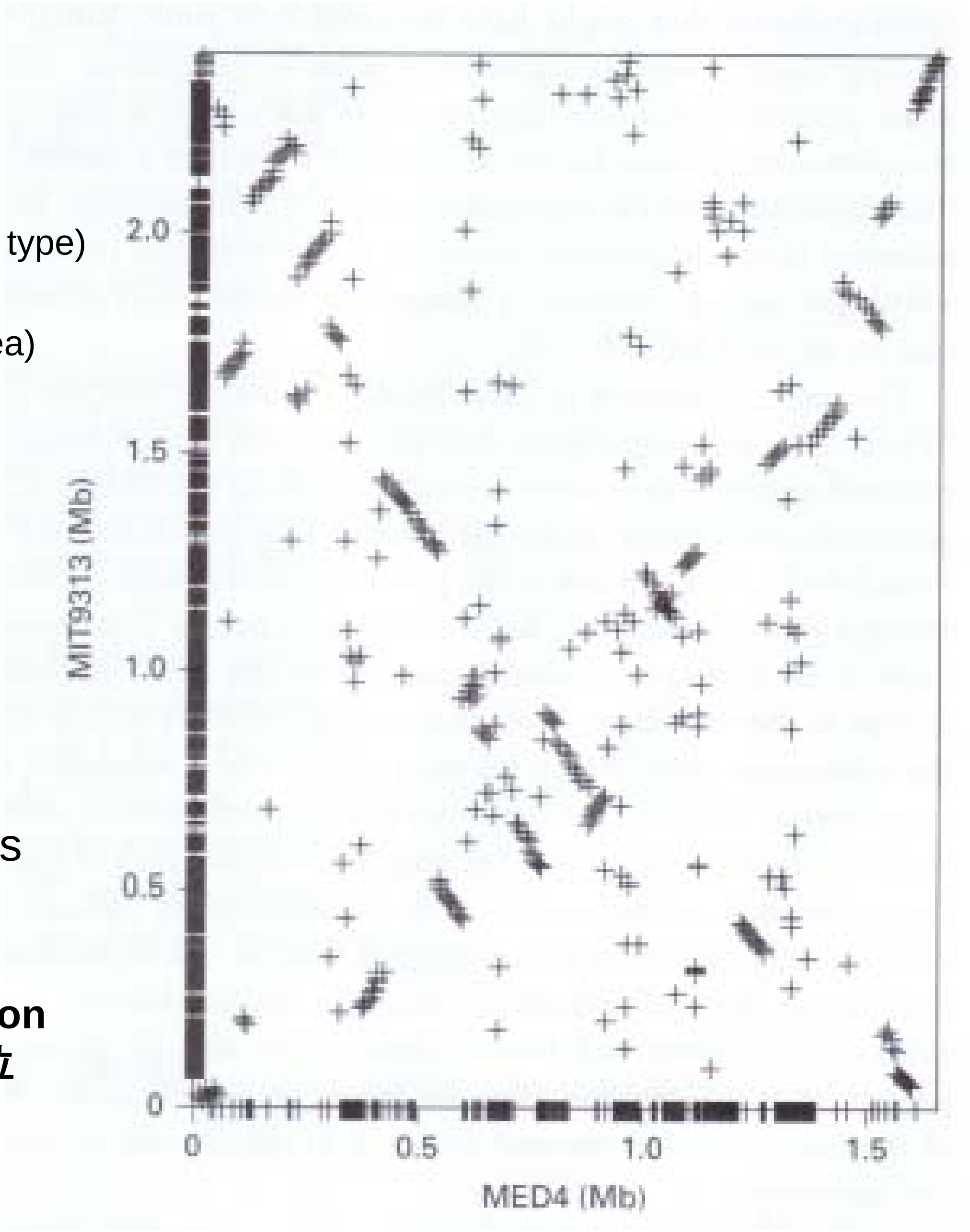
Discovering Genomics,
Proteomics, and Bioinformatics
(CSHL press)

/ a series of homologous regions

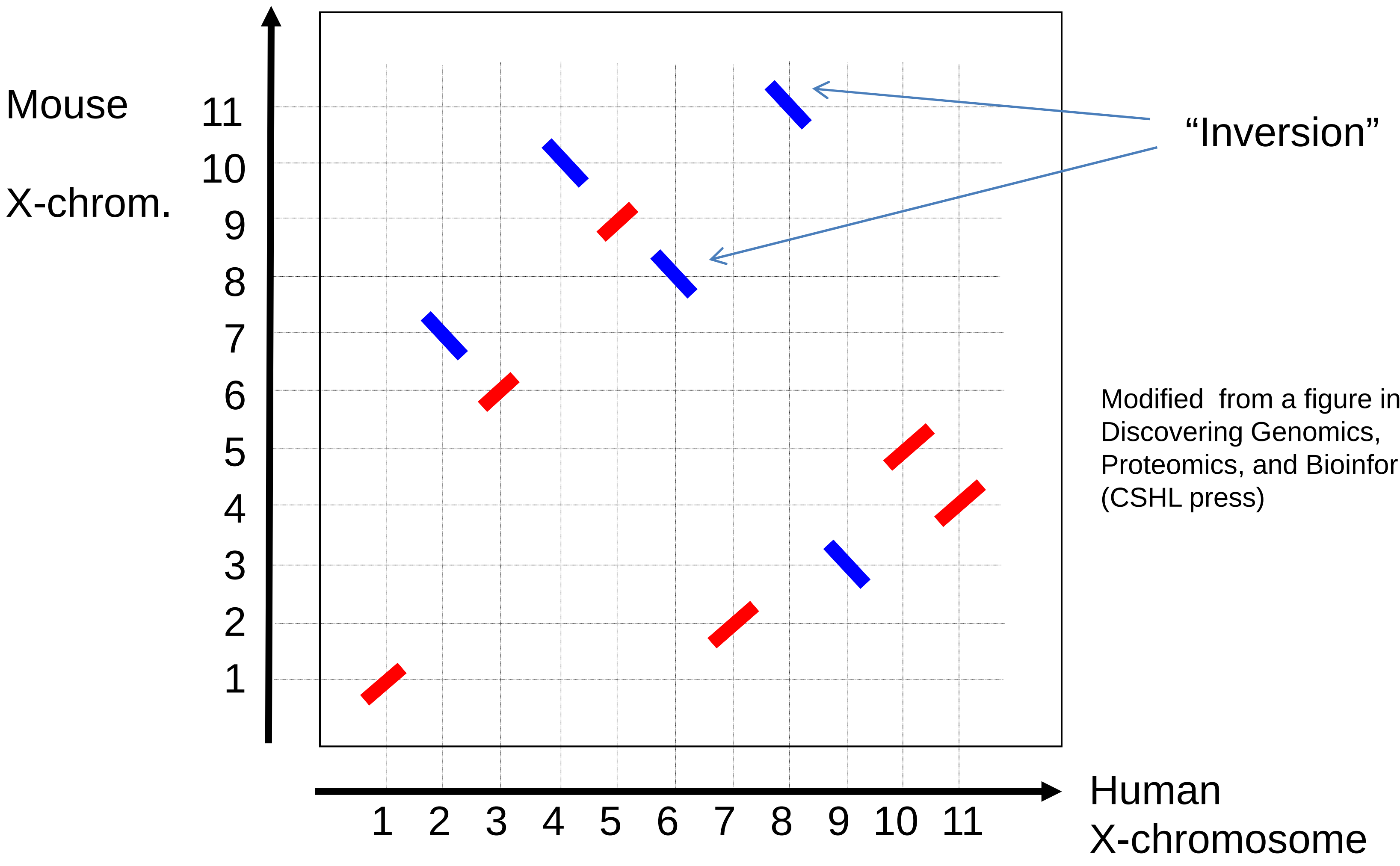
\ a series of **inverted** homologous regions



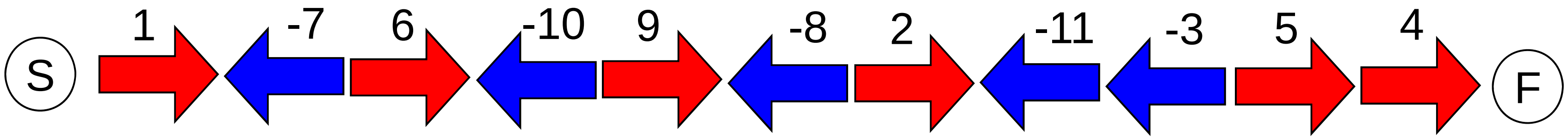
inversion
逆位



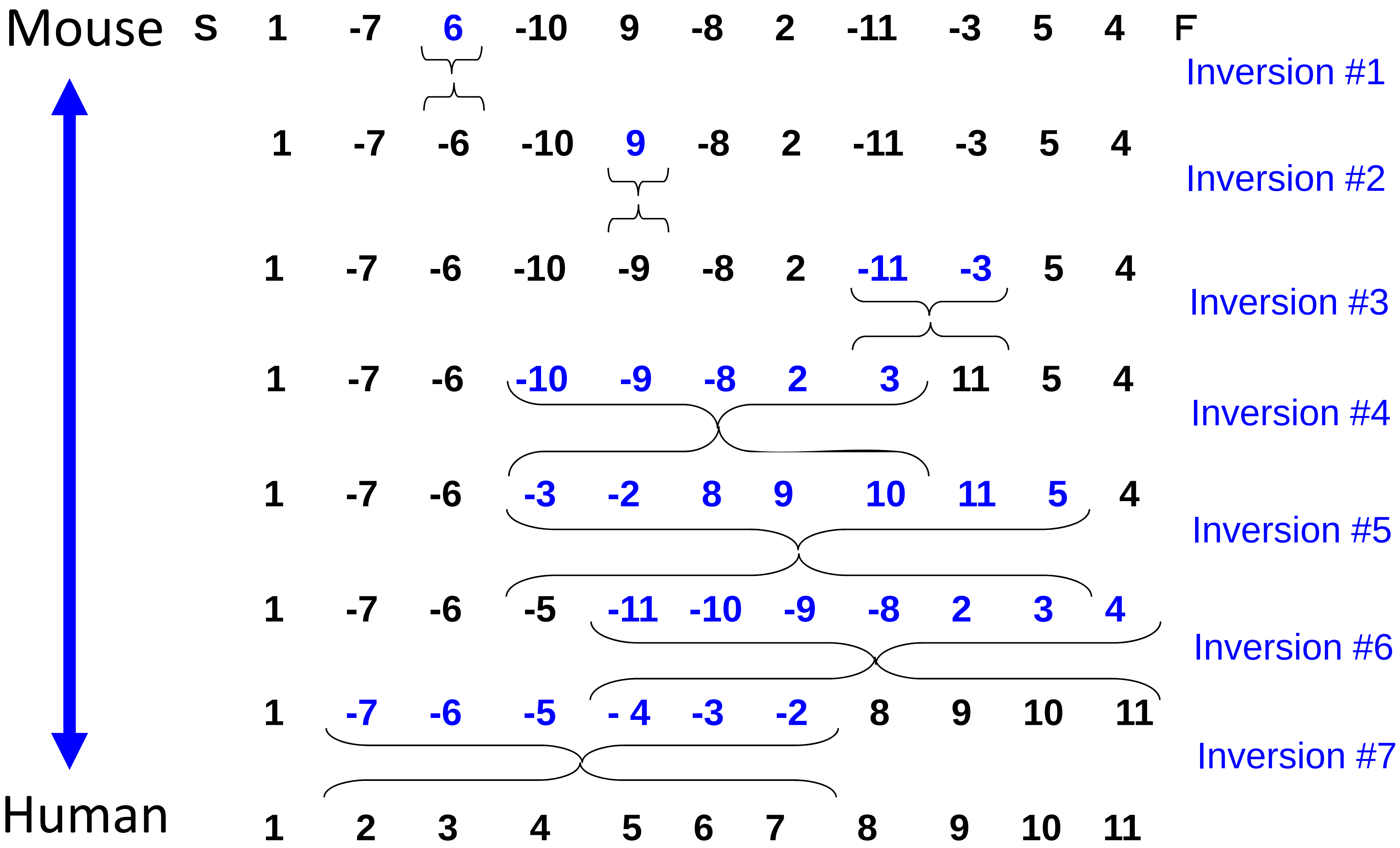
X chromosome (human and mouse)



Modified from a figure in Discovering Genomics, Proteomics, and Bioinformatics (CSHL press)



X chromosome (Mouse and Human)



Mouse S 1 -7 6 -10 9 -8 2 -11 -3 5 4 F

Inversion #1

1 -7 -6 -10 9 -8 2 -11 -3 5 4

Inversion #2

1 -7 -6 -10 -9 -8 2 -11 -3 5 4

Inversion #3

1 -7 -6 -10 -9 -8 2 3 11 5 4

Inversion #4

1 -7 -6 -3 -2 8 9 10 11 5 4

Inversion #5

1 -7 -6 -5 -11 -10 -9 -8 2 3 4

Inversion #6

1 -7 -6 -5 - 4 -3 -2 8 9 10 11

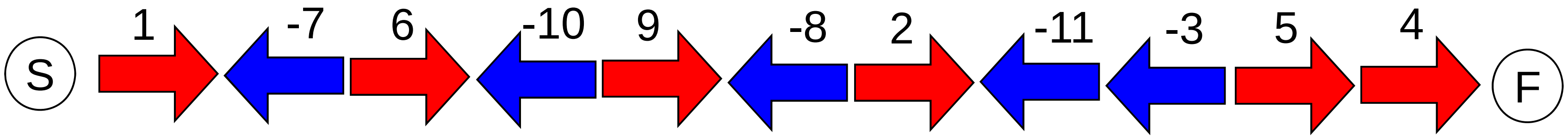
Inversion #7

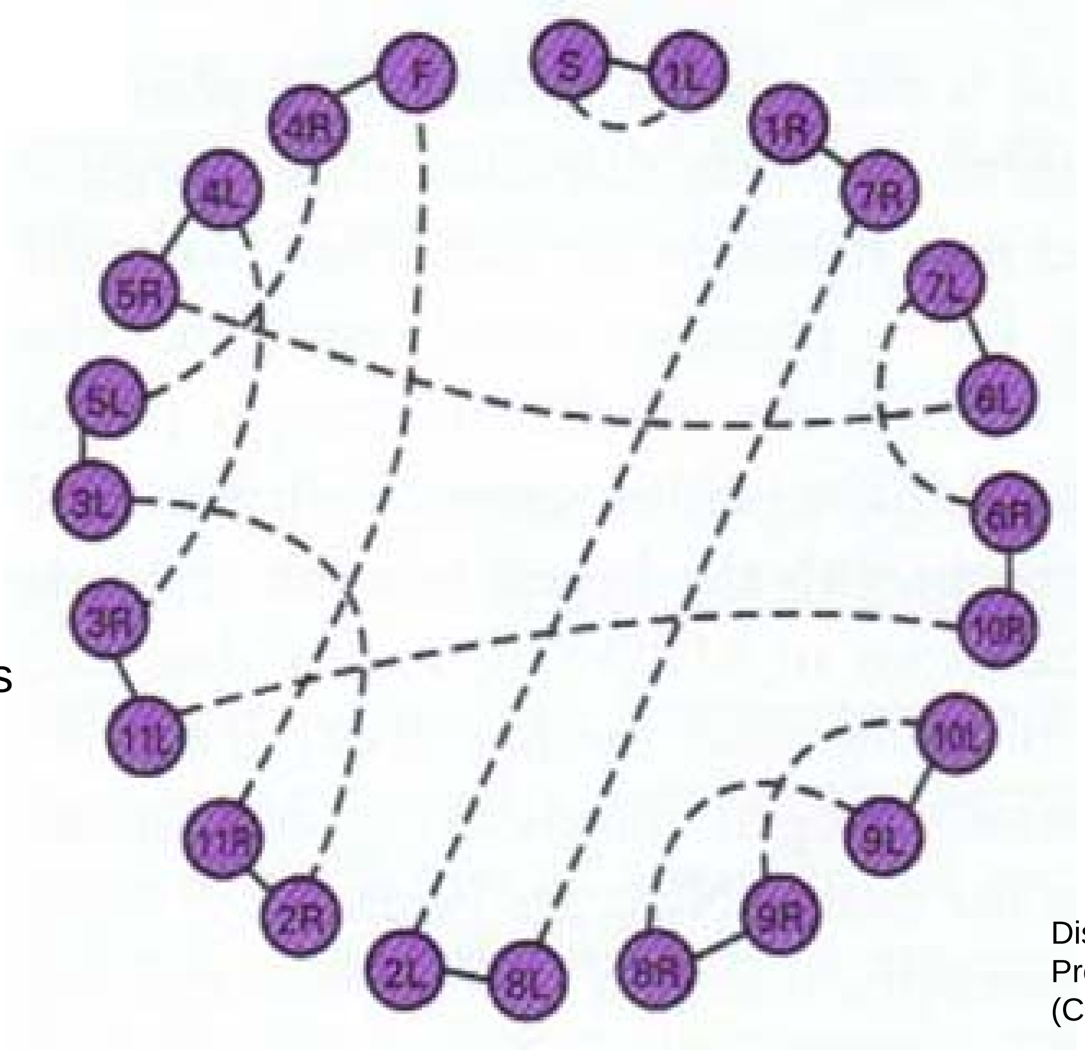
Human 1 2 3 4 5 6 7 8 9 10 11

“**Edit distance**” between Mouse and Human genome is “7” inversion operation.
However, note that mouse is not a direct ancestor of human, and *vice versa.*

Graph Representation

Mouse X chromosome





Outer solid lines
Reality graph
(order in mouse)

Inner dotted lines
Desired graph
(order in human)

Discovering Genomics,
Proteomics, and Bioinformatics
(CSHL press)

Independent Alternative Cycles

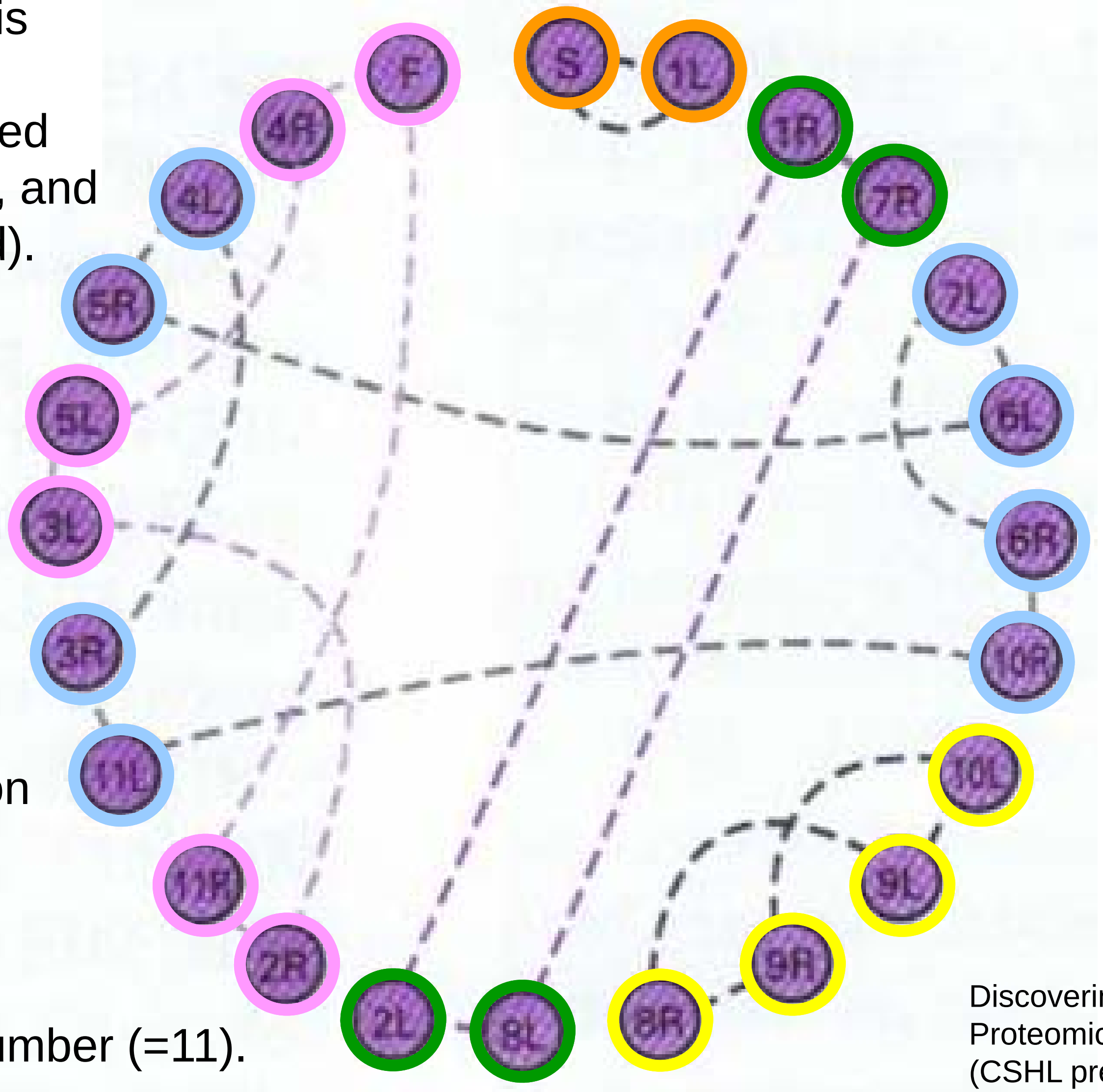
Alternative Cycle:
A closed loop which is composed of alternatively connected Reality edges (solid), and Desired edge (dotted).

C = number of independent (non overlapping) alternative cycles.

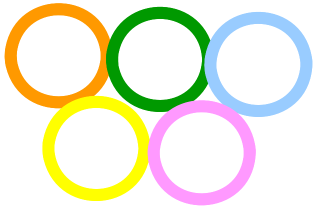
Required number of “inversion” operation is (almost always) given by

$$N+1-C$$

where *N* is gene number (=11).

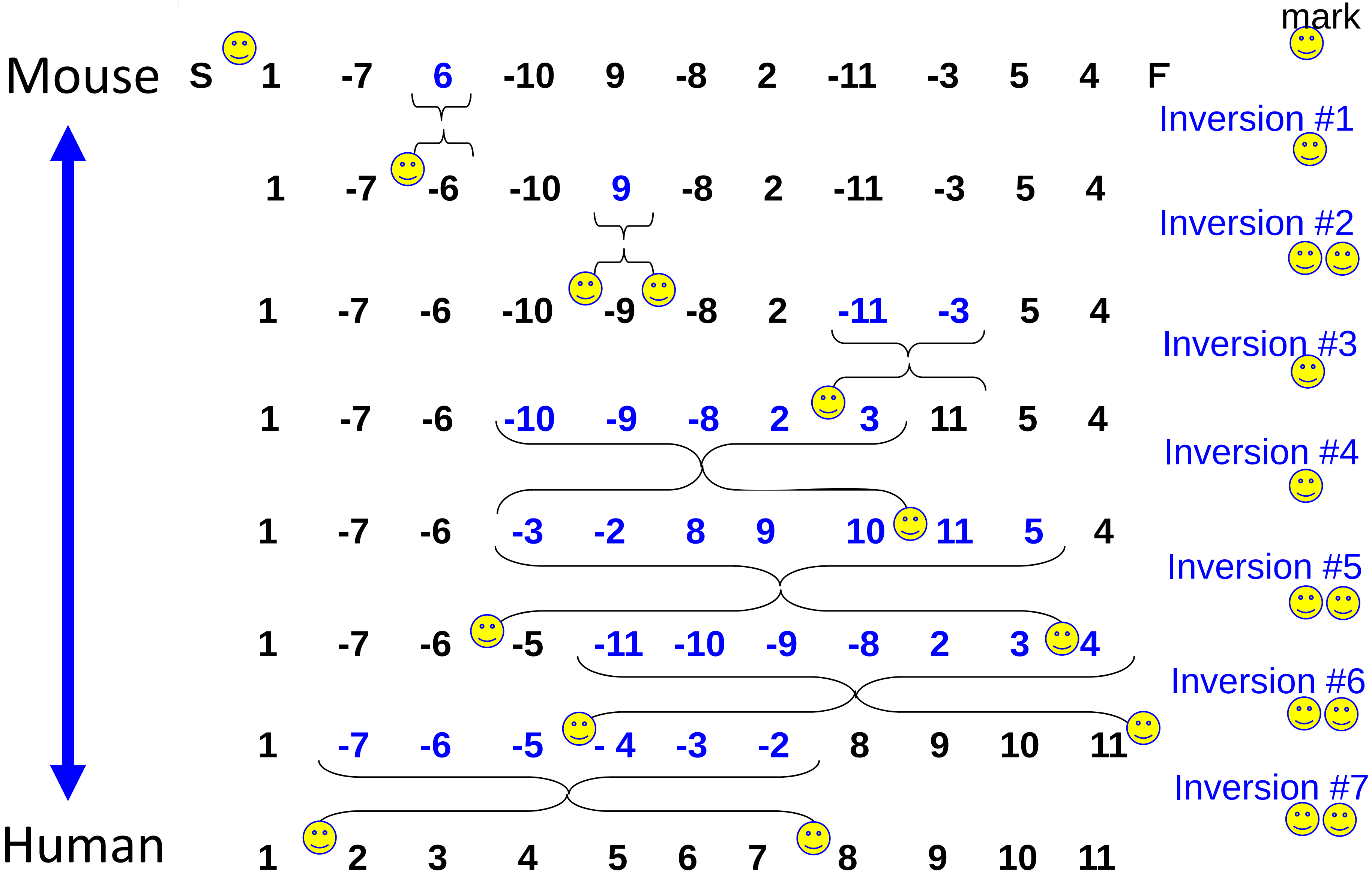


C=5



Discovering Genomics, Proteomics, and Bioinformatics (CSHL press)

X chromosome (Mouse and Human)



satisfaction mark ☺

Mouse S ☺ 1 -7 6 -10 9 -8 2 -11 -3 5 4 F

Inversion #1 ☺

1 -7 ☺ -6 -10 9 -8 2 -11 -3 5 4

Inversion #2 ☺☺

1 -7 -6 -10 ☺ -9 ☺ -8 2 -11 -3 5 4

Inversion #3 ☺

1 -7 -6 -10 -9 -8 2 ☺ 3 11 5 4

Inversion #4 ☺

1 -7 -6 -3 -2 8 9 10 ☺ 11 5 4

Inversion #5 ☺☺

1 -7 -6 ☺ -5 -11 -10 -9 -8 2 3 ☺ 4

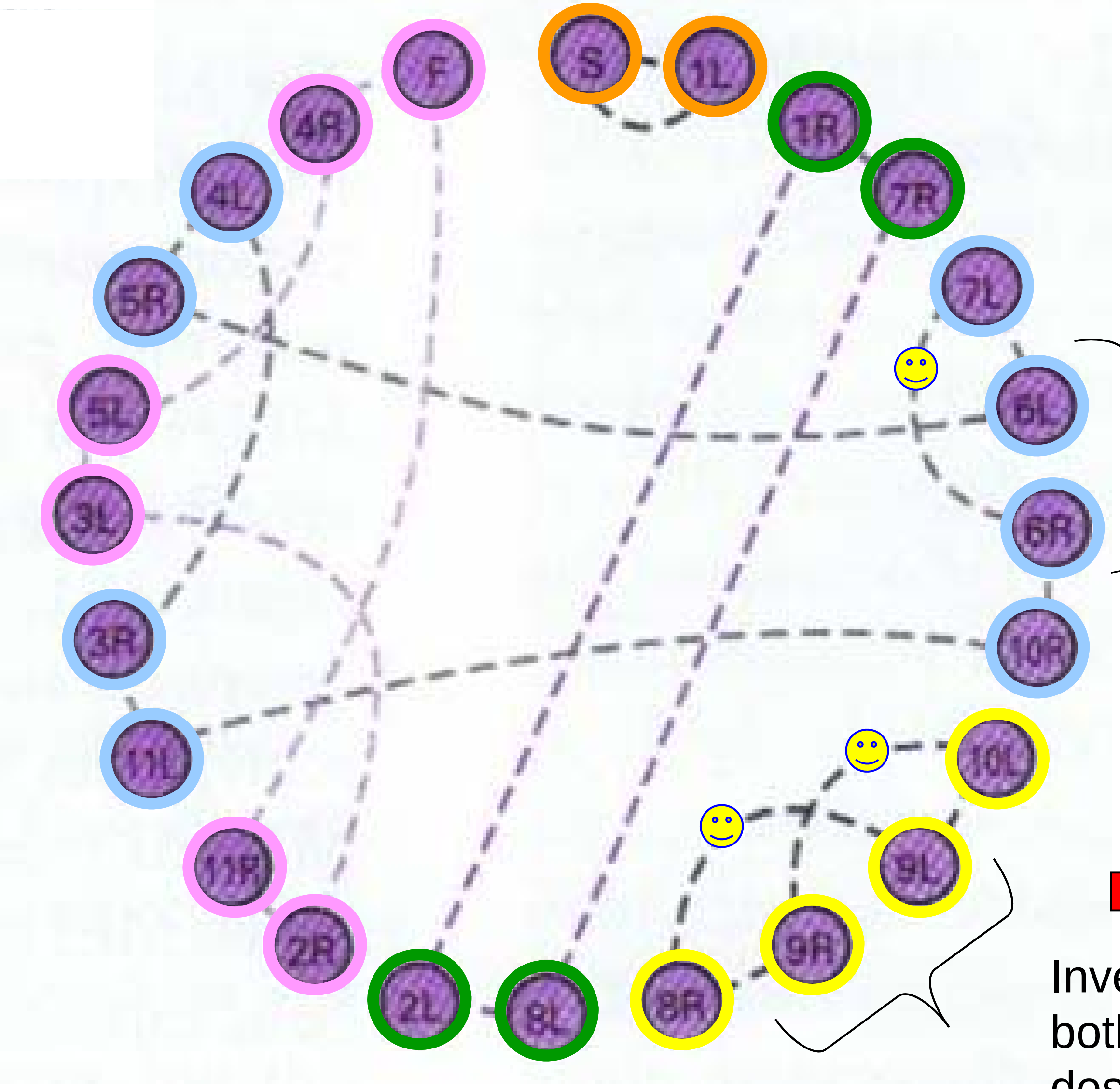
Inversion #6 ☺☺

1 -7 -6 -5 ☺ - 4 -3 -2 8 9 10 11 ☺

Inversion #7 ☺☺

Human 1 ☺ 2 3 4 5 6 7 ☺ 8 9 10 11

n+1=12 intervals. *12* satisfaction marks ☺ (correct gene orders) required in total.
7 inversion operations. 4 double satisfactions. 1 satisfaction from its beginning.

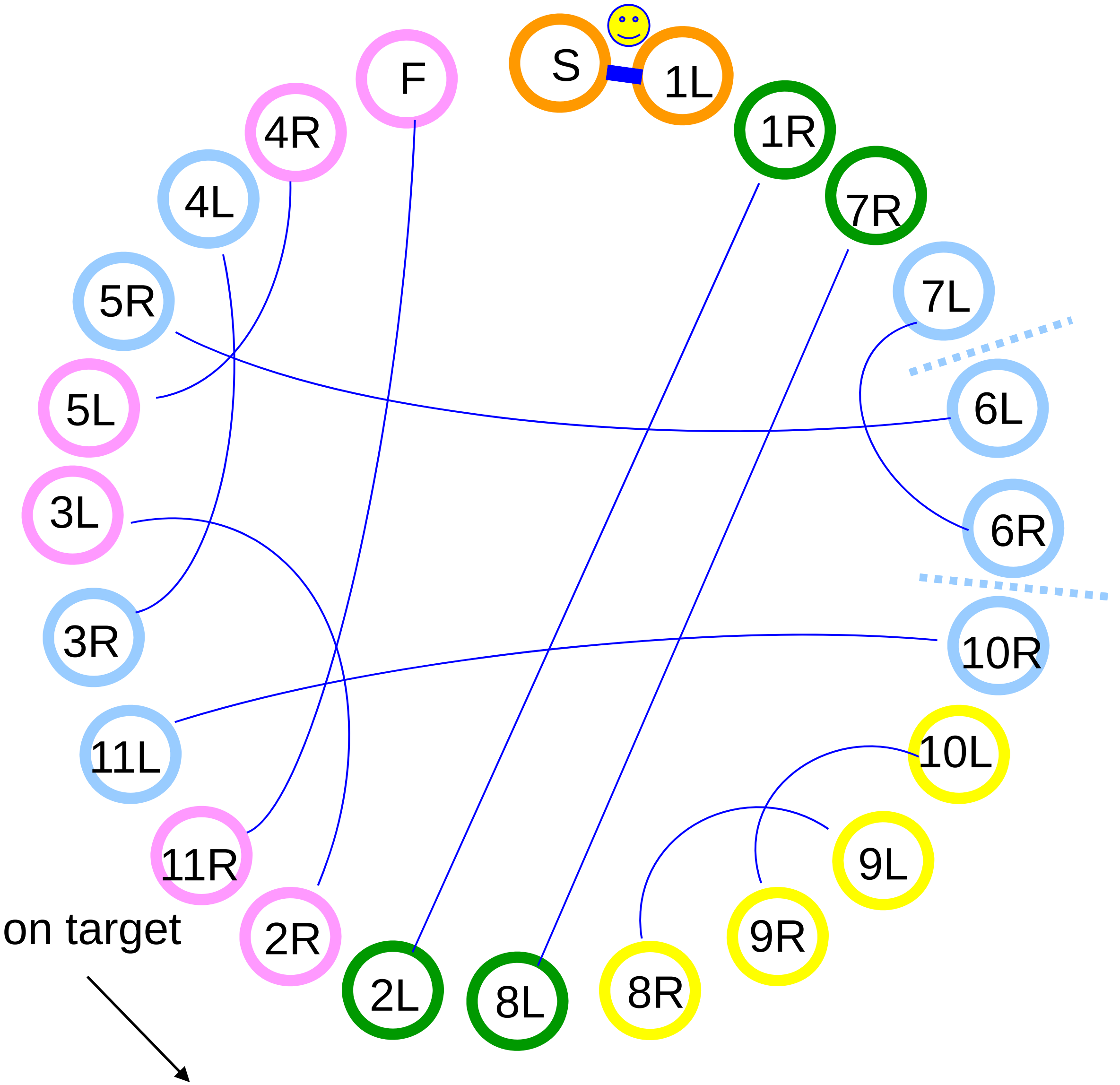


6

Invert gene 6, and then 7L - 6R desired edge will come to outer circumference.

9

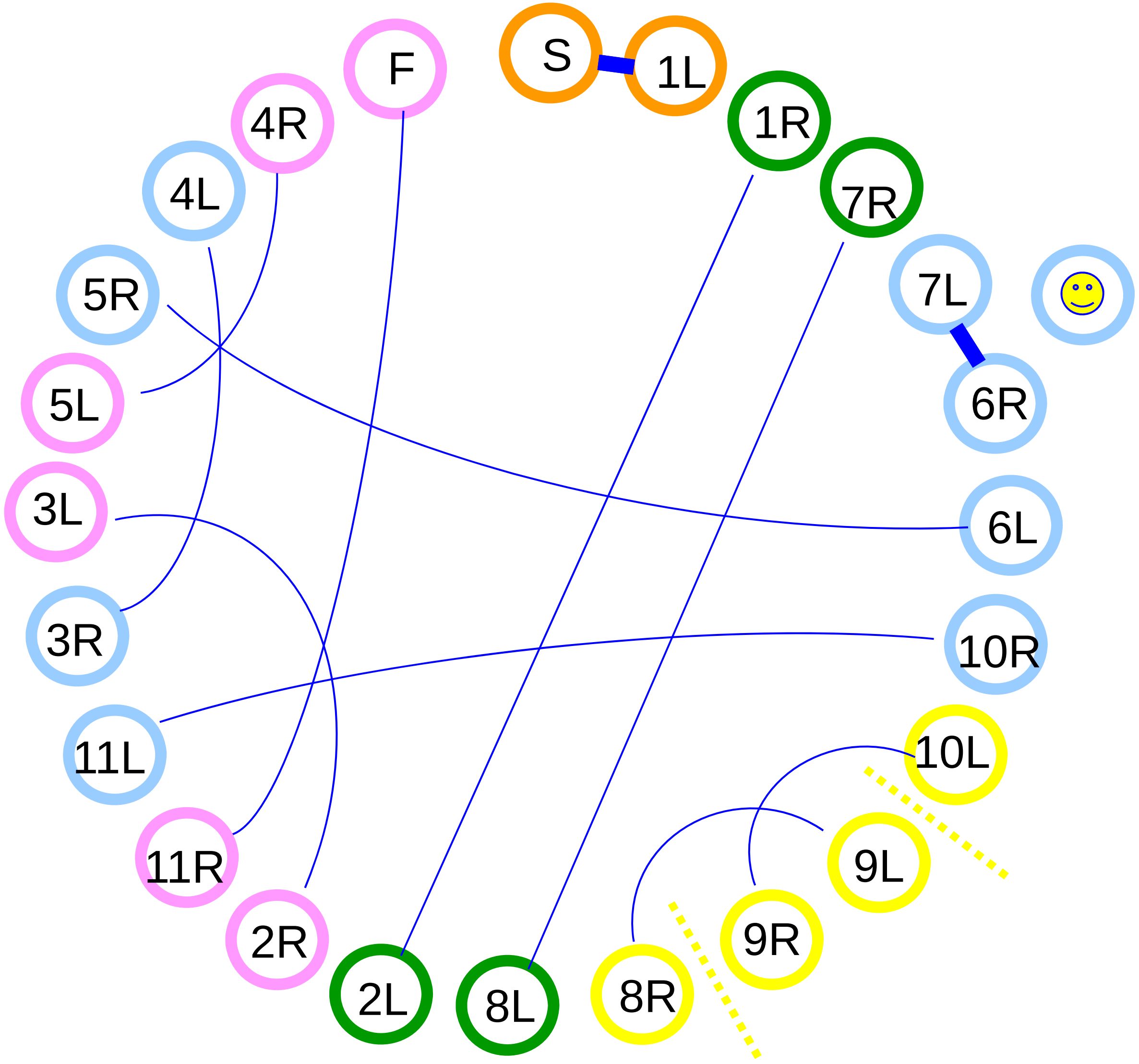
Invert gene 9, and then both 10L - 9R & 9L - 8R desired edges will come to outer place.



First inversion target

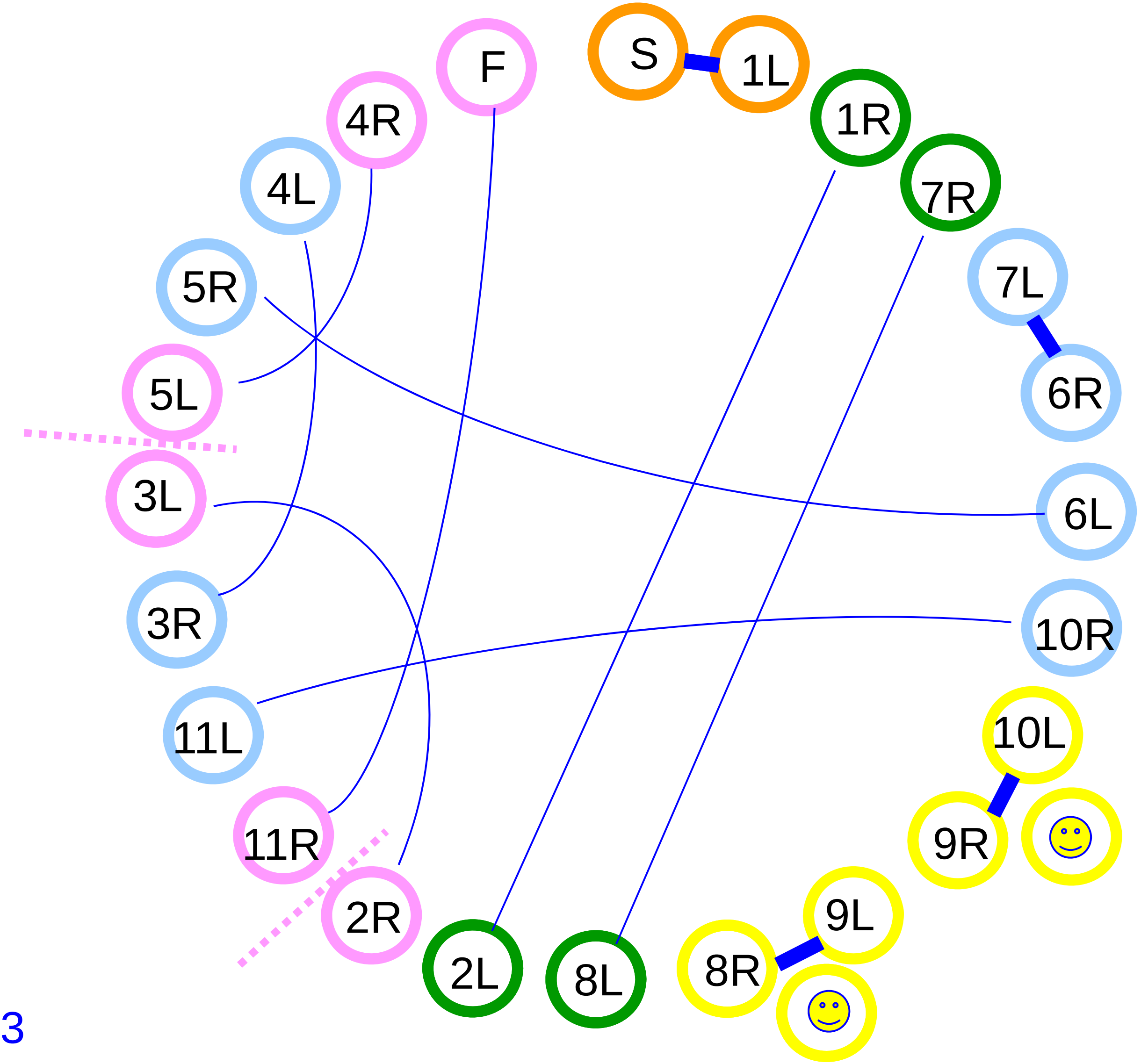
Inversion #1

S 1 -7 6 -10 9 -8 2 -11 -3 5 4 F



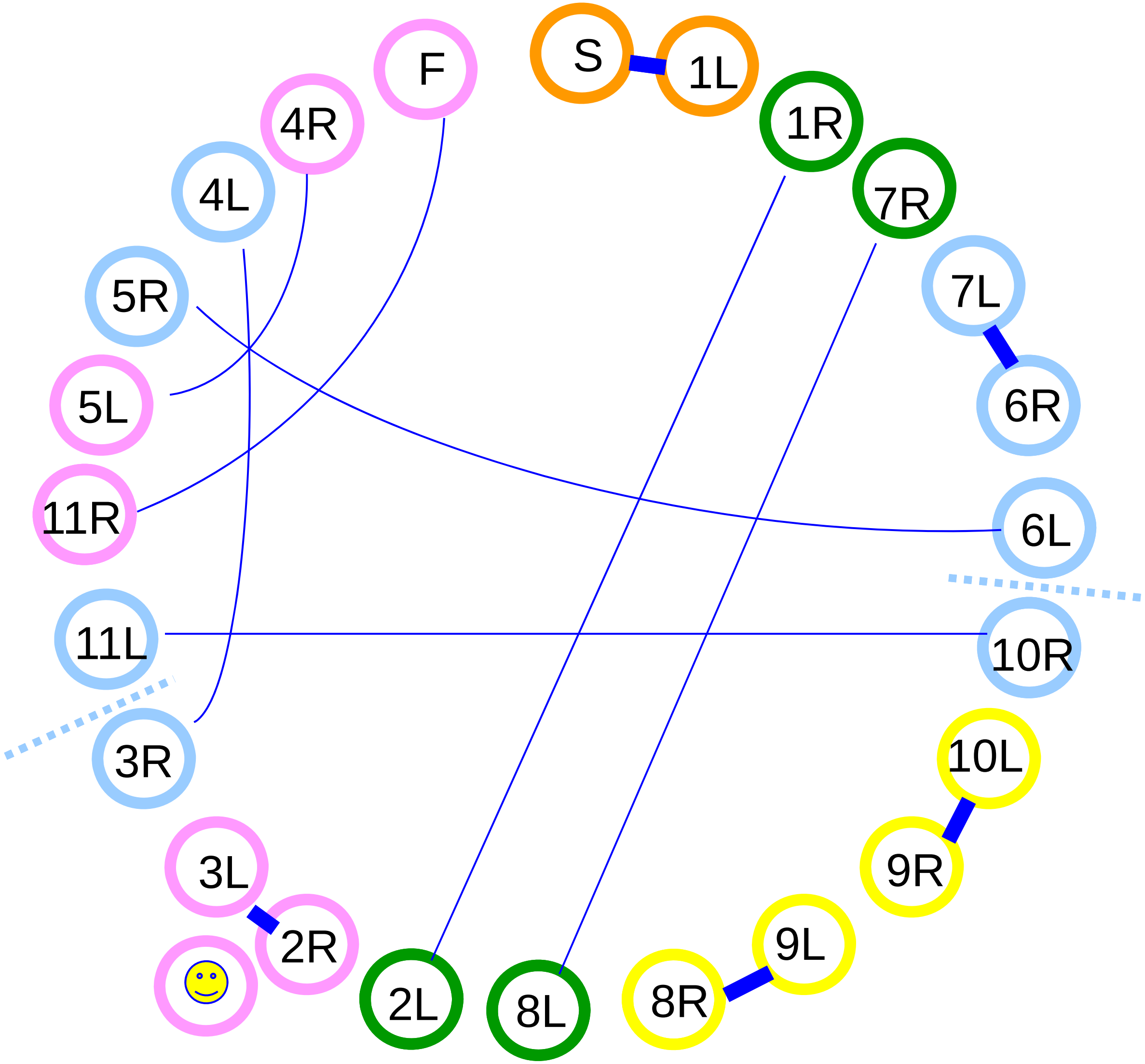
Inversion #2

1 -7 -6 -10 **9** **-8 2 -11 -3 5 4**



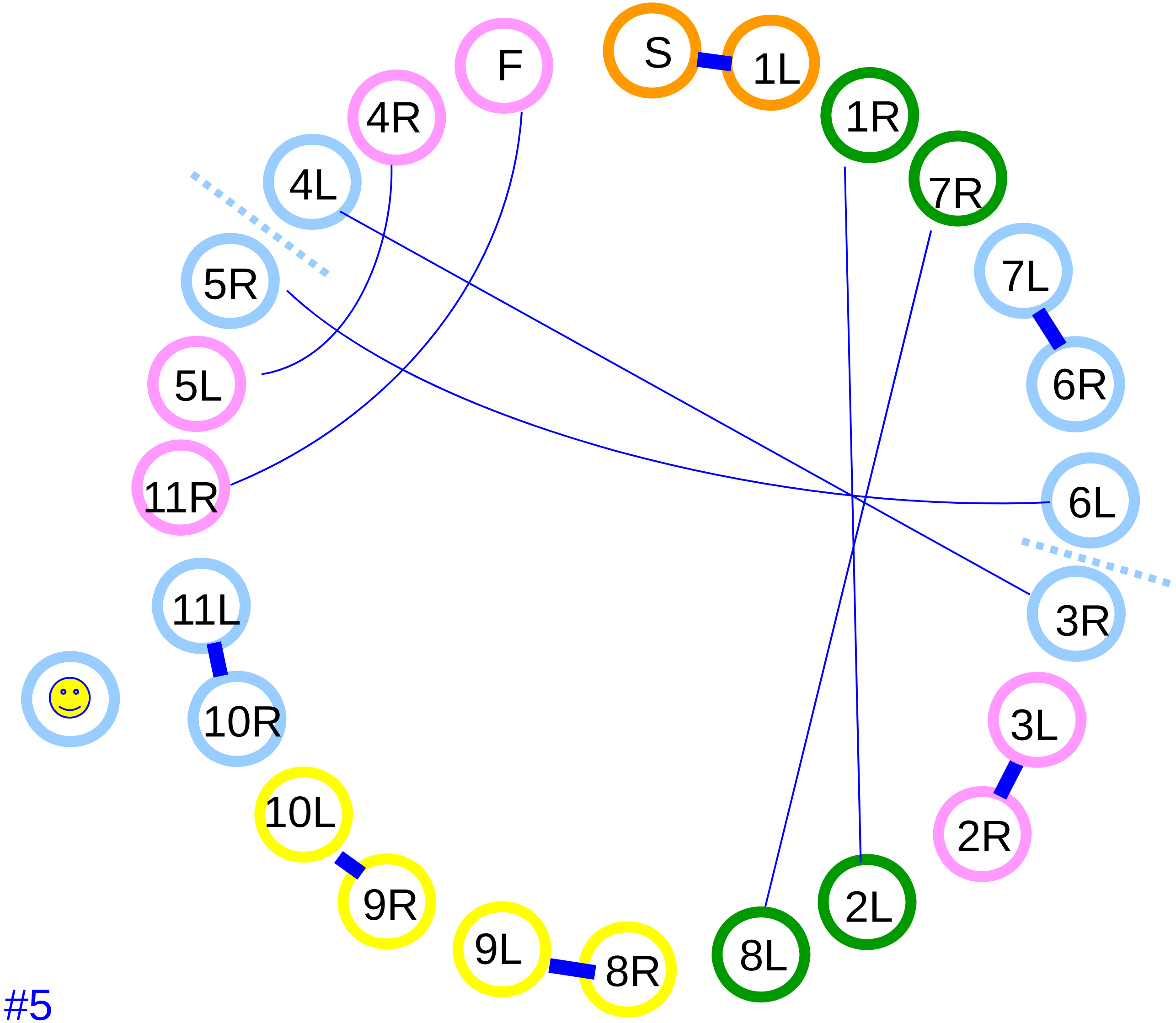
Inversion #3

1 -7 -6 -10 -9 -8 2 -11 -3 5 4



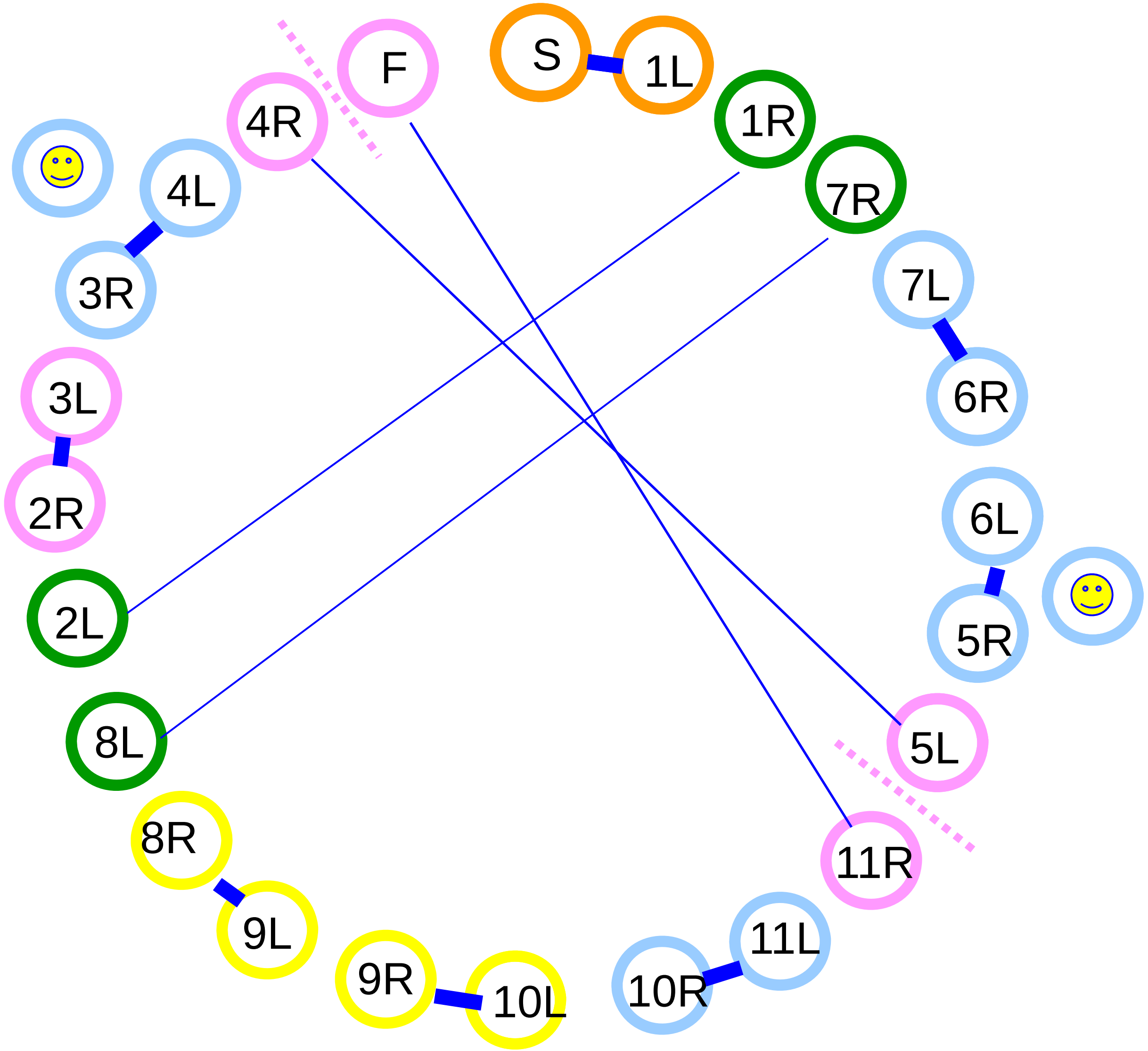
Inversion #4

1 -7 -6 -10 -9 -8 2 3 11 5 4



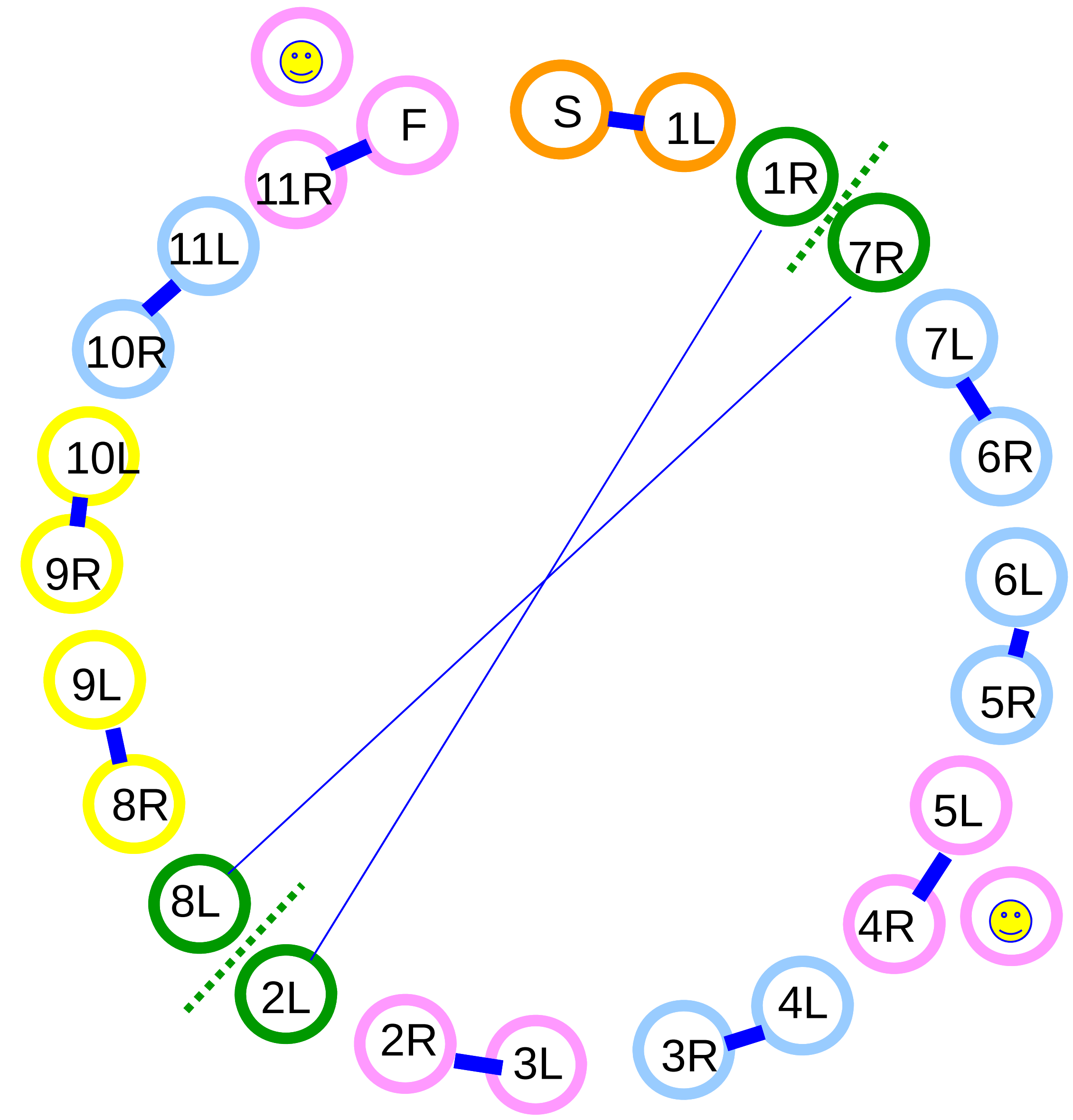
Inversion #5

1 -7 -6 -3 -2 8 9 10 11 5 4



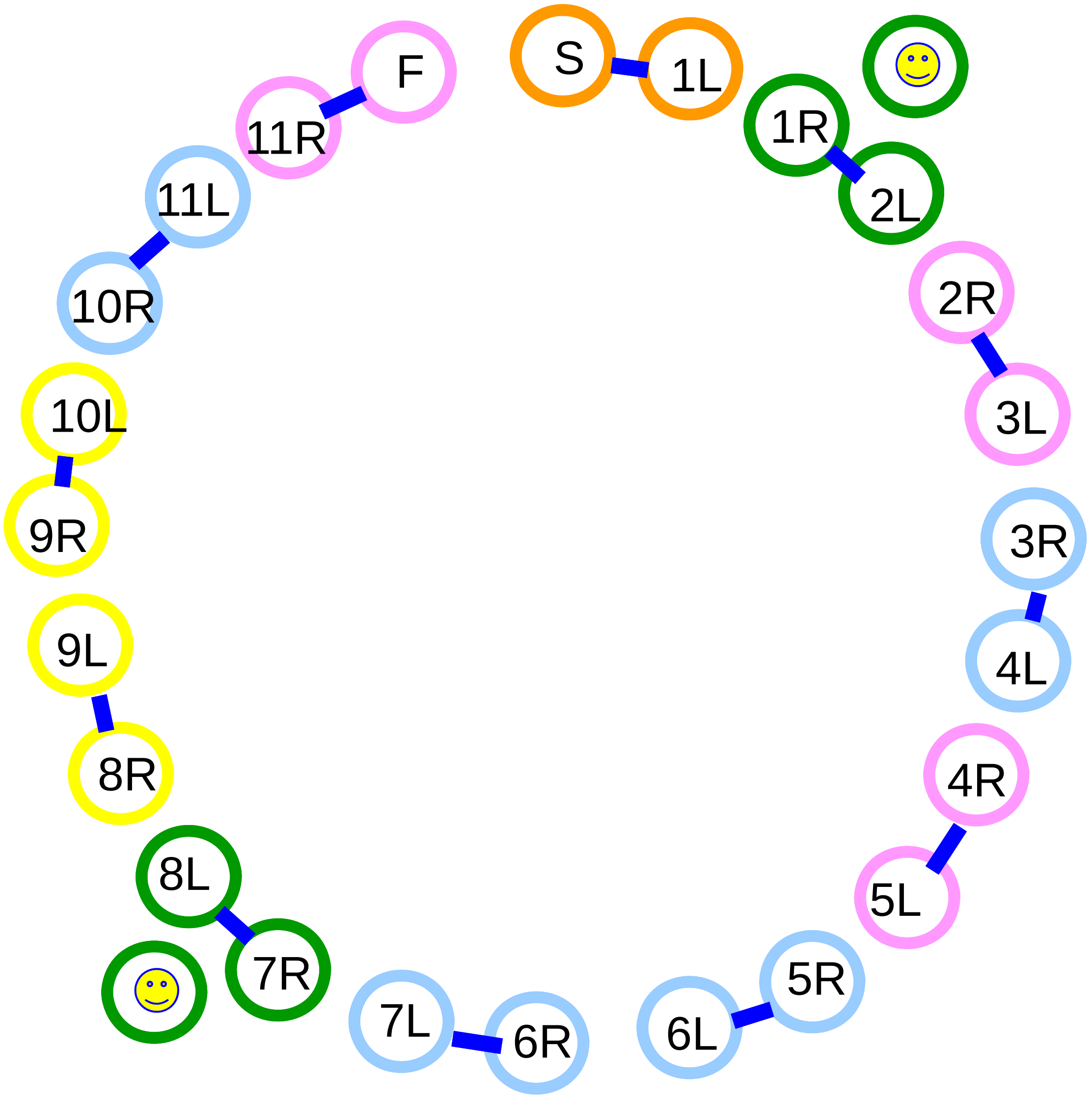
Inversion #6

1 -7 -6 -5 -11 -10 -9 -8 2 3 4



Inversion #7

1 **-7** **-6** **-5** **- 4** **-3** **-2** **8** **9** **10** **11**



Finish

1 2 3 4 5 6 7 8 9 10 11