# CODA.

**C**ross**O**vers **D**istribution **A**nalyzer v.1.1

## Quantitative characterization of crossover position patterns along chromosomes.

Authors:
Franck Gauthier
Olivier Martin
Matthieu Falque

CODA web page

# Introduction

The development of CODA is motivated by the study of meiotic recombination mechanisms and notably the regulation of crossover (CO) distribution.

Recent evidence from several distantly related organisms shows that crossovers form in at least two pathways. One (the interfering pathway) lowers the probability that two COs occur close to each other in the same meiosis. The other (non-interfering pathway) leads to independently distributed COs.

The main purpose of CODA is to provide an intuitive interface to powerful and appropriate statistical and computational tools to characterize CO pathways, namely by measuring interference strength in pathway 1 and the proportion of COs formed through pathway 2. The experimental data may come from genetic linkage maps or cytological CO maps on synaptonemal complexes. The detailed algorithm is given in Falque *et al.*(2009).

---

# Operating principle

CODA estimates the proportion of non-interfering COs (*p*) and the interference level by fitting mathematical models to experimental data. For the *gamma* model, we compute the likelihood in order to quantify the goodness of fit between the model's predictions and experimental data. But for complex models such as the *beam-film* model, it is not possible to compute the likelihood, we have thus developed a "projected score" computing method which measures the goodness of fit. Fitting the model then consists in finding the parameters P and I that maximize the projected score/likelihood.

However, computing a complete two dimensional scan on *p* and *I* could easily become time consuming. We have thus designed an efficient "hill-climbing" algorithm that renders the two-dimensional search computationally feasible.

Confidence intervals of optimal parameters

An accurate way to compute the CI is to use the re-simulation method, where the model's predictions are confronted to "pseudo-experimental" data generated using predicted parameter values. However, this method requires a large number of estimations and may be very tedious and "time consuming" on common workstations. Therefore, this option is not provided via the GUI (graphical user interface), but only via the command line as it can easily be executed (even embedded within scripts) on remote server or computation clusters.

Alternatively, when using the *gamma* model, it is possible to compute approximations of confidence intervals using the Fisher's information matrix, thereby avoiding the re-simulation method.

---

# User interface

## **Input data** (“settings” tab)

### Data on gametes or bivalents

CODA can work with CO positions observed at the level of bivalents or at the level of gametes. Bivalent data may be cytological immunolocalization or EM observation on SCs, or tetrad-derived genetic data. Gamete data may come from linkage mapping experiments (segregation data), or sperm typing.

### Using cytological observations of CO positions

CO physical relative positions have to be collected in a text file observing the following nomenclature: the data name ( i.e. the chromosome name) suffixed by “.copos”.
Example: *maize_1.copos*

Data must be stored in tab-delimited format without any header. Each line is a list of COs observed on one gamete or bivalent (SC).

The first column specifies the name or identifier of each gamete or SC.
The second column specifies the number of COs for each gamete or SC.
The next columns indicate CO physical relative positions.

Example:

```
Zm0037re    2     0.563836    0.972040
Zm9889RNre  2     0.932897    0.972040
Zm01073     2     0.608571    0.988816
Zm9979re    4     0.030857    0.061714    0.348244    0.972040
Zm01154     4     0.035265    0.070530    0.202775    0.972040
Zm0005re    4     0.096979    0.123428    0.145469    0.776326
Zm0083      4     0.189551    0.198367    0.361469    0.871387
Zm01135re   3     0.013224    0.216542    0.955265
Zm01108re   3     0.013224    0.308571    0.893755
```

### Using genetic linkage map segregation data

Segregation data (genotype at marker loci) have to be collected in a text file observing the following nomenclature: the data name suffixed by “.raw”

Raw file format, following the conventions used in the MapMaker genetic mapping software (Lander *et al*, 1987):

Header (first two lines):
The first line specifies the type of population. It must be “data type F2 backcross”.
The second line indicates the number of individuals, then the number of loci, then “0” (number of traits for MapMaker/QTL), then “symbols” followed by the desired symbols for homozygotes (“A”) , heterozygotes (“H”, replaced by “B” in the example below), and missing data ("-").
The third line must be empty.
Each of the next lines contains the locus name (max 8 characters, prefixed by an "*" character) and the genotypes at this marker locus for all individuals (a continuous string).
The locus name and the string of genotypes are separated by one or more spaces.

Example:

```
data type F2 backcross
1419 44 0 symbols A=A X=B B=H X=C X=D -=-

*M1 BABABAAAAAAABABAAAAABAAAABBABBBBABAABBAAABBABBAABAABBBBABABBAABAABABA...
*M2 BABABAAAAAAABABAAAAABBAAABBABBBBABAABBAAABBABBAABAABBBBABABBAABABBABA...
*M3 BABABAAAABAABABAAAAABBAAAABABBBBABAABBAAABBABBAABAABBBBABABBAABABBABA...
*M4 BBBABAAAABAABABBAAAABBAAAABABBBBABAABBAAABBABBAABAABBBBABABBAABABBABA...
*M5 BBBABAAAABAABBBBAAAABBAAAABABBBBABAABBAAABBABBAABAAABBBABABBAABABAABA...
*M6 BBBABAAAABAABBBBAAAABBAAAAAABBBBABAAABAAABBABBAABAAABBBABABBAABABAABA...
*M7 BBBABAAAABAABBBBAAAABBAAAAAABBBBABAAABAAABBABBAABAAABBBABABBAABABAABA...
*M8 BBBABAAAABAABBBBAAAABBAAAAAABBBBABAAABAAABAABBAABAAABBBABABBAABBBAABA...
*M9 BBBABAAAABAABBBBAAAABBAAAAAABBBAABAAABAAABBABBAABAAAABBABABBAABBBAABA...
...
*M44 AAABBAABABABAAAABBBBAAAABABBABBBAAAABABABBBBBAAAABABBBBBBBBABBBBBAABB...
```

The genetic map file obeys the following nomenclature: the data name suffixed by ".map".

Data must be stored in tab-delimited format without any header.

The first column specifies the number of the locus.
The second one specifies the locus name.
Last column indicates the locus genetic map position in centiMorgan. The position of the first locus must be zero.

Example:

```
1     M1    0
2     M2    4
3     M3    7
4     M4    11.6
5     M5    17
6     M6    18.8
7     M7    19.9
8     M8    21.4
9     M9    23.3
10    M10   24.8
...
44    M44   91
```

Access to the experimental data is done via the *working directory* browser.

---

## Parameter fitting ("settings" tab)

### Type of model

For modeling interference, we use the *gamma* model (McPeek and Speed, 1995) which is the most frequently used SRP-based statistical model, and also the *beam-film* (BF) model (Kleckner *et al*., 2004) which is a mechanistically motivated physical model.

To include a second pathway (non-interfering pathway), we use the sprinkling procedure (Copenhaver *et al*., 2002) whereby non-interfering COs are simply added to those of the interfering pathway.

### Number of simulations:

This is the number of simulated gametes or bivalents generated by the model for each combination of

parameters.
Using likelihood with the *gamma* model, simulations are not necessary to fit parameters, but they are used to generate the theoretical histograms (red curves). If you care about these graphs, use at least 1000, otherwise choose 1 (0 is forbidden).
Using the "projected score" to fit the model, the number of simulations should be large enough to ensure a smooth score surface ($10^6$ is usually sufficient).

**Search algorithm for determining the optimum parameters**

Two algorithms are available to explore the parameter space.

- "*2D scan*": Scans all possible combinations of the values of the two parameters (may be extremely slow). 2D scan is required for 3D likelihood (or score) landscape visualization (see Graphs tab paragraph).
- "*Hill-climbing*": Direct search for optimal parameters values (quicker but no likelihood landscape graph is available).

**Score type:**

With the *beam-film* model, no likelihood is available, so the projected score must be used.
With the *gamma* model, using likelihood is more powerful and much quicker.

**Interference model parameters:**

- *nu*: Interference intensity of pathway one (interfering) COs, in the *gamma* model. The "no-interference" situation corresponds to *nu*=1.
- *lambda*: Interference intensity of pathway one (interfering) COs, in the *beam-film* model. The "no-interference" situation corresponds to *lambda*=0.
- *p*: Proportion of COs from pathway 2 (non interfering).
- *Min*, *max*: Range within which the algorithm will search the optimal value of each parameter.
- Precision/step:
  "*Precision*": Using hill-climbing algorithm, absolute precision required to meet the convergence criterion.
  "*Step*": Using 2D scan algorithm, increment of parameter value at each step of the scan.

**Advanced:**

- *Randseed*: Seed value of random number generator. May be useful to perform replicates of parameter estimations.
- *Bin nb of histograms*: For CO-CO interval length distributions.
- *Starting point for hill-climbing*: Initial conditions (value of nu (or lamda) and p) for the hill climbing algorithm, specified via the relative position in the *Min – Max* search interval.

---

## Output data

### *Graphs* tab

The graphs tab is made up of two parts:

The lower part that provides a panel displaying the current and best parameters values, with their

confidence intervals (*gamma* model only), as well as the average inter-crossover distance and variance in all experimental gametes (right side of the panel).
The upper part that provides three types of graphs.
Both are real-time updated throughout the duration of the fit.
At any time, one can switch from a graph to an other, keeping the parameters panel visible.

Hill-climbing trajectory: Using the hill-climbing algorithm, displays the trajectory (in the parameters space) from the starting point (initial parameter values) to the peak of the likelihood (or score) hill (best parameter values).

3D surface plot: Using 2D scan algorithm, displays a 3D visualization of score/likelihood.

The user can easily manipulate the 3D view using mouse and/or keyboard:

- To rotate the view: move mouse cursor while keeping left mouse button pressed, or use up/down/left/right keyboard arrows.
- To translate the view vertically or horizontally: move mouse cursor while pressing left mouse button + Ctrl key together, or use keyboard arrows while pressing left mouse button + Ctrl key together.
- To zoom in/out the view: move the mouse wheel over the plot, or use Ctrl+Alt+left button then move mouse vertically.
- To stretch or compress Z axis (likelihood/score): move the mouse wheel while pressing Ctrl+Maj (=shift) keys together.
- Cut-off sliders on x y and z axes allow the user to display a subset (sub-matrix) of the 3D surface. NB: Only min cut-off is available on Z axis.

WARNING The 2D scan will take a long time if the value of *Step* is too small. In practice, start with a coarse mesh (10 points in each direction). The graph will begin to be displayed after the first three rows have been computed.

Histograms: Give the possibility to graphically visualize agreement between simulated data generated by the model (red curves) and experimental data (green histogram bars) on the same plot.

Three categories of histograms are available:

- Distribution of the number of COs per chromosome.
- Distribution of the positions of COs along the chromosome. You can choose to look at all chromosomes or to select only chromosomes with 1, 2, 3 etc COs.
- Distribution of distances between successive COs. You can choose to look at all chromosomes or to select only chromosomes with 1, 2, 3 etc COs.

In addition, one can press the “export graph” button to save a bitmap (png, jpeg, bmp) or vector (pdf, eps, ps, svg) image of the currently displayed graph.
NOTE: pdf and svg exports are not available for 3D scan graphs.

### *Output* tab

Displays the standard text output while the fit is running.

#### Output text files

The fitting program generates several output files in the working directory.

- *route.txt*: numerical values corresponding to the hill-climbing trajectory graph.

- *inter_histograms.txt*: numerical values corresponding to all experimental and simulated histograms.
- *scan_dataname.txt*: lists all the parameter values tested with the corresponding likelihood or score values all along the fitting procedure (either for hill climbing or 2D scan).

Values displayed in the parameters panel are taken from this file and are updated whenever a new line is added to the file.

*dataname.gen2phy* and *dataname.phy2gen*: gives the correspondence between genetic (cM) and relative physical positions along the chromosome (available only if datatype is set to *CO physical relative positions*).
*dataname.comap*: genetic map created from experimental physical relative CO positions (*dataname.copos* file), if datatype is set to *CO physical relative positions*.

---

# Command line interface

Analyses can also be launched from a command-line terminal by calling the *interfit.x* program with a quite long list of ordered arguments:

*14* : The first argument has to be 14 !
*model_type*: 2 for gamma or 6 for beam-film
*ipar_min*: the lower bound for the first parameter (interference level), a floating point value (≥0)
*ipar_max*: its upper bound ...
*ipar_step* (or precision): a floating point positive or negative value (see the NOTE below)
*pnip_min*: the lower bound for the fraction of non-interfering COs, a floating point value (≥0)
*pnip_max*: its upper bound (≤1)
*pnip_step* (or precision): a floating point positive or negative value (see the NOTE below)
*ipar_prec*: has to be 0
*centro_size*: has to be 0
*centro_begin*: has to be 0
*thinning*: "th0" or "th1", indicate that the data come from bivalent or gamete observations respectively.
*win_beg*: has to be 0
*win_end*: has to be 50000
*initx_distrib*: has to be "uniform"
*numsim*: the number of simulations (integer)
*randseed* seed value for random number generator
*dataname*: "maize_1" for example
*datatype*: cyto (for CO physical relative positions) or genetic (for segregation data)
*score_type*: gamma (likelihood) or ccdall (projected score)

Optional arguments can be specified at the end of this list to simulate and analyze "pseudo-experimental" data (see **Operating principle** section above). Optional arguments are the following:

"*simdata"* : the first optional argument has to be "simdata" !
*ipar_simdata*, *pnip_simdata*: parameters values used to generate "pseudo-experimental" data (i.e. values obtained via a previous analysis of true experimental data).
*nb_chrom_simdata* : number of "pseudo-experimental" chromosomes to generate.

NOTE: the algorithm to search for the optimum is set through the signs of *ipar_step* and *pnip_step*. Two positive signs leads to complete 2D scan, whereas two negative signs lead to the "hill-climbing" algorithm.

---

# License

# References


Falque, M. *et al*. (2009) Two Types of Meiotic Crossovers Coexist in Maize. *Plant Cell*, **1**, 3915-3925.

Lander, E.S. *et al*. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **2**, 174181.

McPeek, M.S. and Speed, T.P. (1995) Modeling Interference in Genetic Recombination. *Genetics*, **139**, 1031-1044.

Kleckner, N. *et al*. (2004) A mechanical basis for chromosome function. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12592–12597.

Copenhaver, G.P., Housworth, E.A. and Stahl, F.W. (2002) Crossover Interference in Arabidopsis. *Genetics*, **160**, 1631-1639.