Validation of a morphogenesis model of *Drosophila* early development by a multi-objective evolutionary optimization algorithm

Rui Dilão[1] Daniele Muraro[1]
Miguel Nicolau[2] Marc Schoenauer[2]

[1]Nonlinear Dynamics Group, IST
Department of Physics, Av. Rovisco Pais, Lisbon, Portugal

[2]INRIA Saclay - Île-de-France
LRI - Université Paris-Sud, Paris, France

EvoBIO 2009

GENNETEC

Outline

Validation of mathematical models of real Complex Systems

- Search for the set of parameters that best approaches model output with available real data;
- Usually a hard, multi-modal problem:
  - Potential experimental errors on available data;
  - Data may originate from several experiments with different setups;
  - Gradient-based techniques fail to give reliable solutions.
- Evolutionary Algorithms are a better choice.

Calibration of a Morphogenesis Model of *Drosophila*

- Distribution of *Bicoid* and *Caudal* proteins along the antero-posterior axis of the embryo of *Drosophila*.
- Ideal optimisation will find parameters fitting the distribution of both proteins through minimisation of sum of MSE;
  - Infeasible given noise and different experimental setups.
- Multi-objective algorithms a better approach for model calibration and validation.

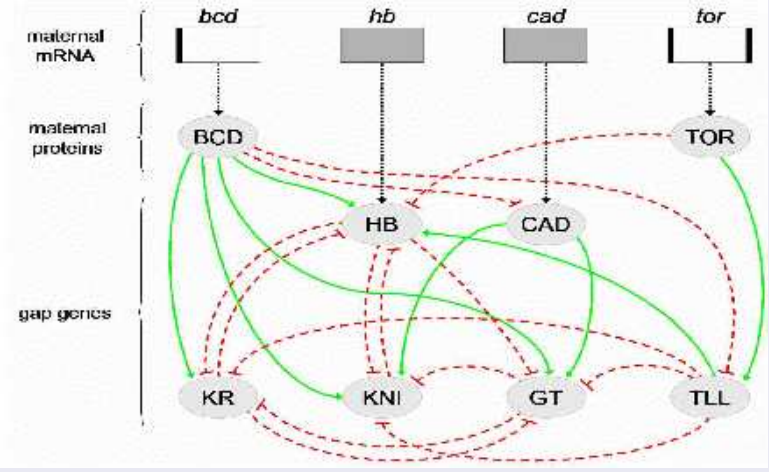
Morphogenesis in *Drosophila* early development

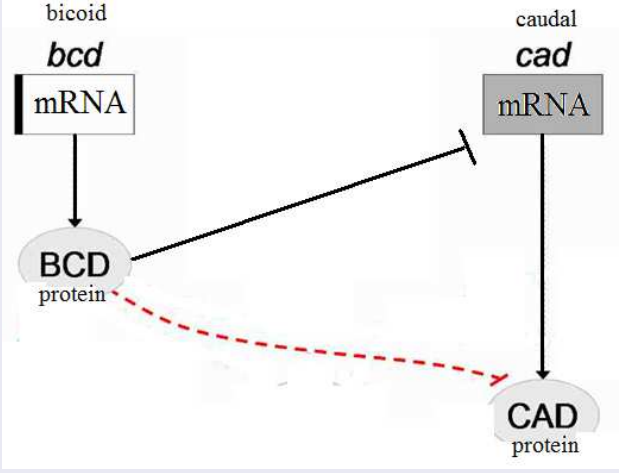
First 2h of development

- Begins with deposition of *bicoid* mRNA of maternal origin near pole of embryo:
- 14 mitotic nuclear replication cycles (first 2h);
- Nuclear membranes appear at end of 14^{th} mitotic cycle;
- Absence of membranes facilitates diffusion of proteins:
  - stable gradients are established.

Morphogenesis in *Drosophila* early development

Regulation Network responsible for first 95 minutes



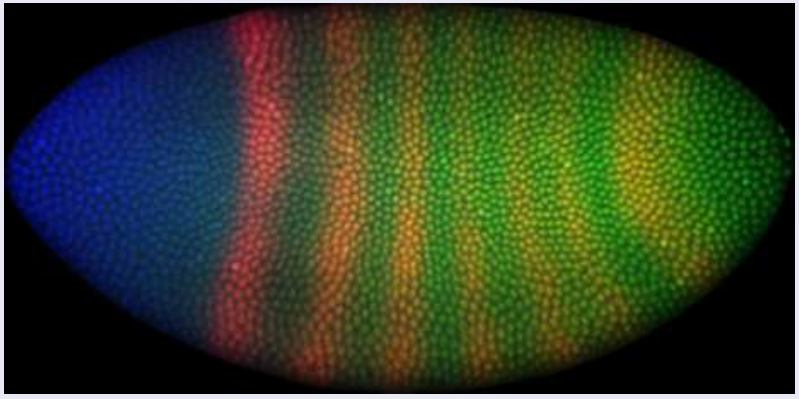


- Model repression mechanism between *Bicoid* and *Caudal*;
- Interested in spatial gradients of both proteins.

(From: F. Alves and R. Dilão, J. Theoretical Biology, 241 (2006) 342-359.)

Morphogenesis in *Drosophila* early development

After 14^{th} replication cycle

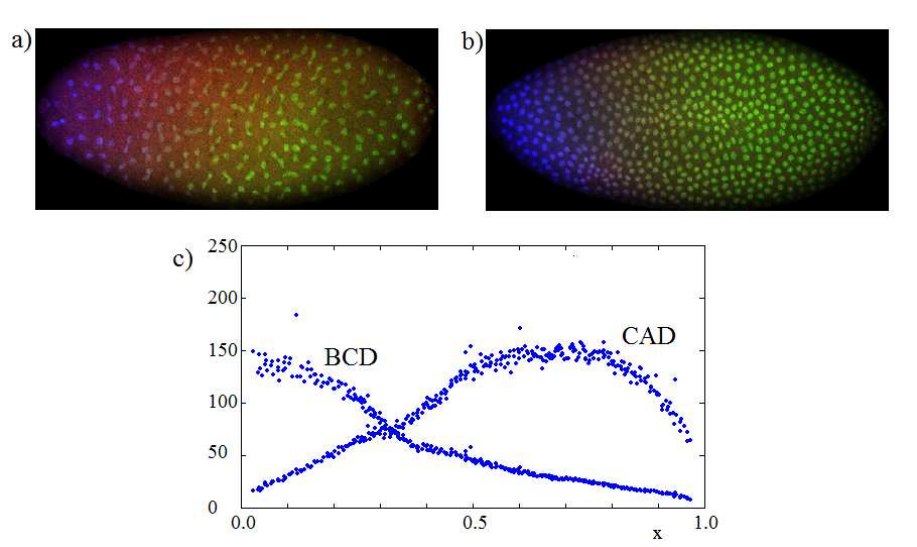


- Fluorochrome measurement marking protein concentrations proportional to intensity;
- Blue: Bicoid; Green: Caudal.

(experimental data, FlyEx database)

Morphogenesis in *Drosophila* early development

11^{th} (a) and 12^{th} (b) replication cycles



- From 1a to 1b the nuclei have divided by mitosis, but proteins keep apparent gradient;
- 1c shows concentrations of BCD and CAD along the antero-posterior axis (x) of embryo.

(experimental data, FlyEx database, datasets ab18 (a) and ab17 (b))

Mathematical Model of Protein Diffusion

- The *bicoid* and *caudal* mRNA of maternal origin have initial distributions given by:

$$bcd(x, t = 0) = \begin{cases} B > 0, & \text{if } 0 < L_1 < x < L_2 < L \\ 0, & \text{otherwise} \end{cases}$$

$$cad(x, t = 0) = \begin{cases} C > 0, & \text{if } 0 < L_3 < x < L_4 < L \\ 0, & \text{otherwise} \end{cases}$$

L_1, L_2, L_3 and L_4 are constants representing intervals of localisation of the corresponding mRNA; B and C are concentration constants.

Mathematical Model of Protein Diffusion

- During first stage of development, *bicoid* and *caudal* are transformed into proteins with rate constants a_{bcd} and a_{cad}:

$$bcd \xrightarrow{a_{bcd}} BCD \qquad cad \xrightarrow{a_{cad}} CAD$$

- *Bicoid* prevents expression of *Caudal* through repression mechanism described by the mass action type transformation:

$$BCD + cad \xrightarrow{r} BCD$$

r is rate of degradation

Mathematical Model of Protein Diffusion

- From mass action law, model equations are deduced:

$$\begin{cases} \frac{\partial bcd}{\partial t} &= -a_{bcd} bcd(x) + D_{bcd} \frac{\partial^2 bcd}{\partial x^2} \\ \frac{\partial BCD}{\partial t} &= a_{bcd} bcd(x) \\ \frac{\partial cad}{\partial t} &= -a_{cad} cad(x) - rBCD.cad + D_{cad} \frac{\partial^2 cad}{\partial x^2} \\ \frac{\partial CAD}{\partial t} &= a_{cad} cad(x) \end{cases}$$

- System of non-linear parabolic partial differential equations;
- Diffusion of *bicoid* and *caudal* mRNA is added.

Mathematical Model of Protein Diffusion

Calibrate model just derived with experimental data

- Parameters to calibrate:
  - L_1, L_2, L_3 and L_4;
  - B and C;
  - a_{bcd} and a_{cad};
  - D_{bcd} and D_{cad};
  - r and t (time).

Hard optimisation problem

- Model is an approximation;
- Biological data is noisy;
- Optimise with single- or multi-objective algorithms.

Single-Objective Approach

CMA-ES: state of the art in evolutionary computation

- $(\mu, \lambda)-$Evolutionary Strategy:
  - Population of μ parents to generate λ offspring;
  - Deterministically choose the best μ offspring to become parents for the next generation;
  - Offspring generated by sampling Gaussian distribution centered on weighted recombination of parents;
  - Multi-dimensional Gaussian distributions determined by their covariance matrix;
  - Notion of cumulated path to separately update stepsize and covariance matrix.

Multi-Objective Approach

MO-CMA-ES

- Multi-objective version of CMA-ES:
  - Based on a specific (1+1)-CMA-ES algorithm;
  - λ_{MO}(1+1)-CMA-ES are run in parallel, each with its own stepsize and covariance matrix;
  - At each step, set of λ_{MO} parents and their λ_{MO} offspring are ranked, according to selection criterion;
  - Fleisher algorithm used for selection - based on hyper-volume measure.

Fitness Functions

Optimise MSEs of model with experimental data of distribution of *BCD* and *CAD*

- Optimise two fitness functions:

$$FitBCD(\vec{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}(BCD(x_i,\vec{\alpha}) - BCD_{exp}(x_i))^2$$

$$FitCAD(\vec{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}(CAD(x_i,\vec{\alpha}) - CAD_{exp}(x_i))^2$$

(α = set of parameters to be optimised)

  - CMA-ES optimises function:

$$Fit(\vec{\alpha}, c_i) = FitCAD(\vec{\alpha}) + c_i \cdot FitBCD(\vec{\alpha})$$

  - 12 different c_i slopes sample Pareto front.

Parameters

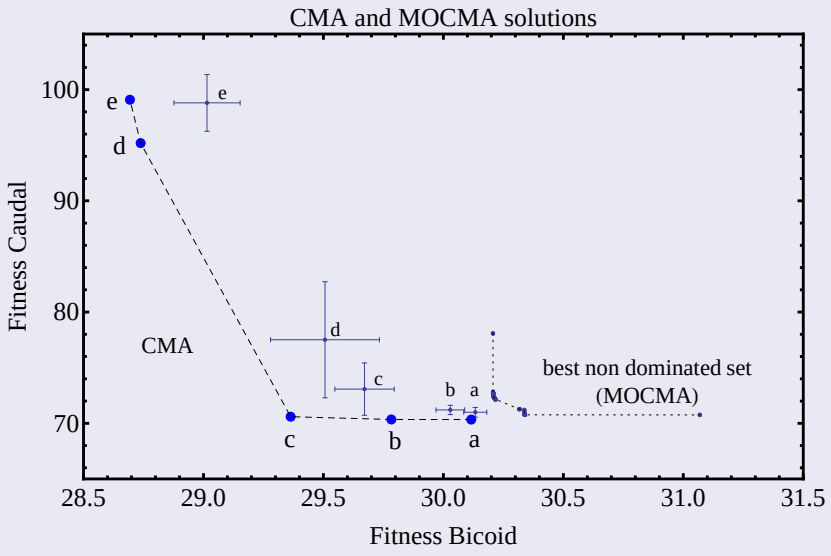
MO-CMA-ES

- Population size $\lambda_{MO} = 100$;
- Gradual penalisation to reduce spread of Pareto front;
  - Sample Pareto front in range $[0, 40] \times [0, 80]$;
  - Penalise *FitBCD* by amount which *FitCAD* overpassed upper bound.
- 100 runs: best non dominated points extracted;

CMA-ES

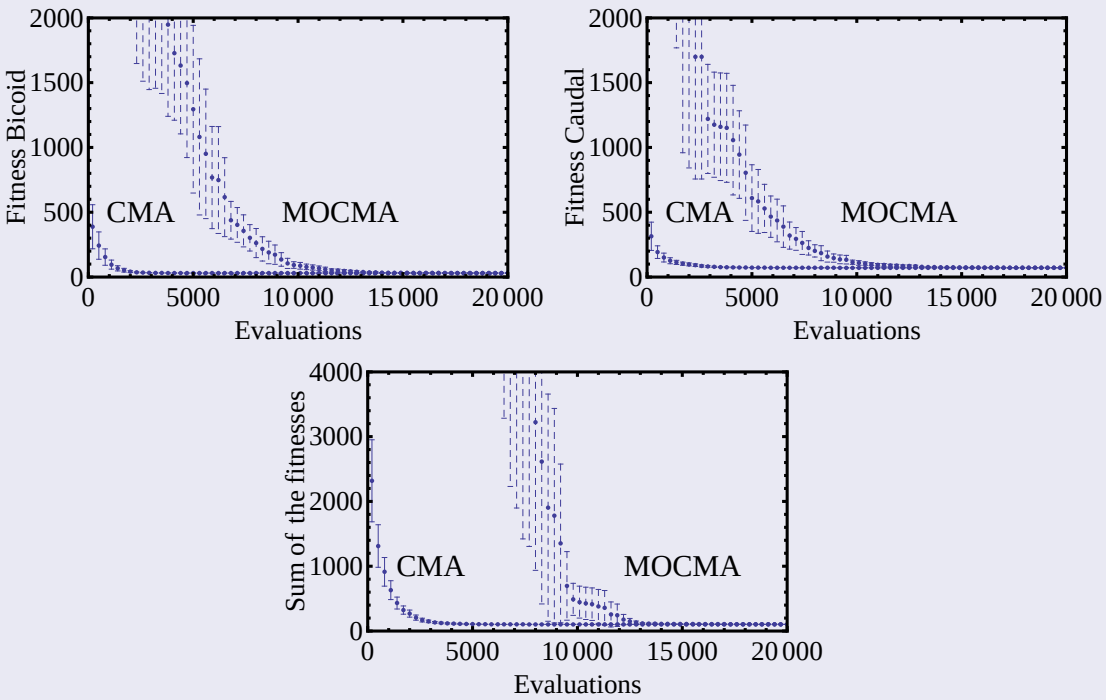
- Population size $\lambda_{CMA} = 4 + \lceil 3 \times \log n \rceil$;
- Fitness function: $Fit(\vec{\alpha}, c_i) = FitCAD(\vec{\alpha}) + c_i \cdot FitBCD(\vec{\alpha})$
  - 12 slopes used $(0.01, 1, 5, 10, \ldots, 100)$, 10 runs per slope;
- Best non-dominated results from each slope gathered to form Pareto front.

Pareto Front Approximation



- Best non-dominated sets found by both algorithms;
- CMA-ES results for slopes $(1, 5, 25, 50, 100)$;
- Asymmetrical relationship between *FitCAD* and *FitBCD*:
  - In accordance with biology.

Fitness Evolution over time



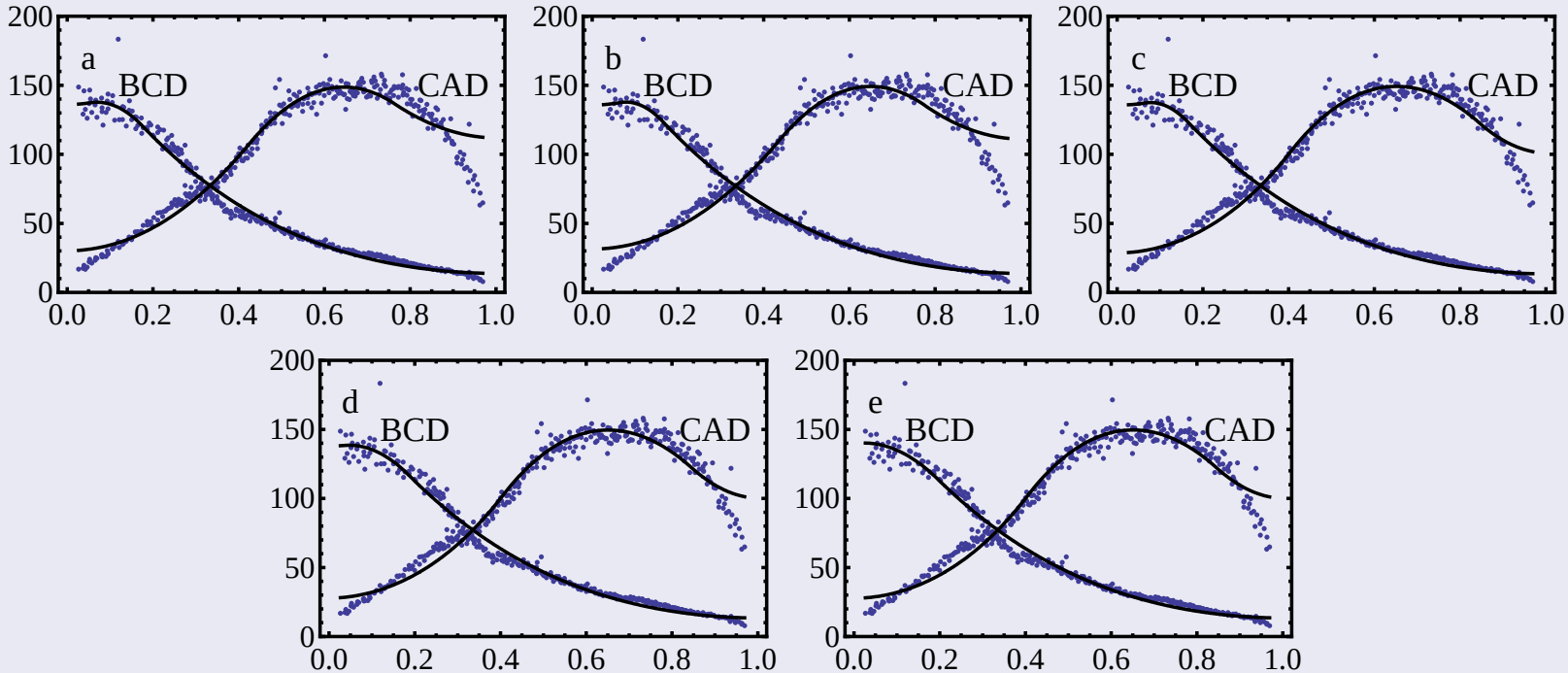
- Evolution of MSEs on *BCD* and *CAD*;
- Similarity between runs on CMA-ES, but not on MO-CMA-ES.

Best Sets of Parameters Found

	a	b	c	d	e	mean
L_1	$5.68 \cdot 10^{-2}$	$6.72 \cdot 10^{-2}$	$6.25 \cdot 10^{-2}$	$3.29 \cdot 10^{-2}$	$1.43 \cdot 10^{-2}$	$4.67 \cdot 10^{-2}$
L_2	$1.73 \cdot 10^{-1}$	$1.68 \cdot 10^{-1}$	$1.62 \cdot 10^{-1}$	$1.84 \cdot 10^{-1}$	$1.94 \cdot 10^{-1}$	$1.76 \cdot 10^{-1}$
L_3	$4.28 \cdot 10^{-1}$	$4.35 \cdot 10^{-1}$	$4.04 \cdot 10^{-1}$	$4.07 \cdot 10^{-1}$	$4.04 \cdot 10^{-1}$	$4.16 \cdot 10^{-1}$
L_4	$7.63 \cdot 10^{-1}$	$7.74 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$	$8.48 \cdot 10^{-1}$	$8.15 \cdot 10^{-1}$
B	$1.53 \cdot 10^{+3}$	$1.98 \cdot 10^{+3}$	$3.47 \cdot 10^{+3}$	$2.36 \cdot 10^{+3}$	$1.98 \cdot 10^{+3}$	$2.26 \cdot 10^{+3}$
C	$1.06 \cdot 10^{+3}$	$1.08 \cdot 10^{+3}$	$1.26 \cdot 10^{+3}$	$1.28 \cdot 10^{+3}$	$1.28 \cdot 10^{+3}$	$1.19 \cdot 10^{+3}$
D_{bcd}	$1.00 \cdot 10^{-2}$	$1.09 \cdot 10^{-2}$	$1.99 \cdot 10^{-2}$	$2.03 \cdot 10^{-2}$	$2.04 \cdot 10^{-2}$	$1.63 \cdot 10^{-2}$
D_{cad}	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$
a_{bcd}	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$
a_{cad}	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$	$9.99 \cdot 10^{+4}$
r	$8.64 \cdot 10^{+3}$	$6.74 \cdot 10^{+3}$	$3.34 \cdot 10^{-2}$	$5.74 \cdot 10^{-2}$	$6.71 \cdot 10^{-4}$	$3.07 \cdot 10^{+3}$
Iterations	$9.84 \cdot 10^{+3}$	$9.79 \cdot 10^{+3}$	$9.37 \cdot 10^{+3}$	$9.35 \cdot 10^{+3}$	$9.36 \cdot 10^{+3}$	$9.54 \cdot 10^{+3}$

- Parameters from 5 best non-dominated solutions of CMA-ES;
- All valid solutions from a Biological point of view.

Fitting the Experimental Data



- Best 5 non-dominated solutions of CMA-ES;
- (a): best *FitBCD*, (e): best *FitCAD*.

Error of Calibrated Model

Accuracy of Model

- Can be measured by fitness function;
- For BCD, if BCD_{max} is maximum value of experimental values:

$$\sqrt{\frac{FitBCD}{(BCD_{max})^2}}$$

- For current experimental data, error in range $3\% - 6\%$.

Computer Science point of view

- Striking difference in performance between both algorithms:
  - Multi-objective lacks pressure toward Pareto front;
  - Concentrates on spreading, even after penalisation.
- Future work: test how well identified parameters generalise over other datasets.

Biological point of view

- Applicability of an mRNA diffusion model to describe protein gradients in early *Drosophila* development;
- **Non-dominated variability provided by multi-objective approaches intrinsic to biological systems**:
  - Helps explain phenotypic plasticity of living systems.

Acknowledgements

- Work funded by the European project GENNETEC (FP6 STREP IST 034952).
- Project partners:

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
INRIA
ICTP
INSTITUTE FOR SCIENTIFIC INTERCHANGE FOUNDATION
INSTITUTO SUPERIOR TÉCNICO
noraybio
GENNETEC

Bibliography
Relevant publications

- F. Alves and R. Dilão: Modelling segmental patterning in Drosophila: Maternal and gap genes. *Journal of Theoretical Biology*, 241 (2006) pp. 342–359
- N. Hansen, and A. Ostermeier: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: *ICEC96*. IEEE Press. (1996) pp. 312–317
- C. Igel, N. Hansen, and S. Roth: Covariance Matrix Adaptation for Multi-objective Optimization. Evolutionary Computation, Vol. **15**, No. 1. (2007) pp. 1–28
- R. Dilão and D. Muraro and M. Nicolau and M. Schoenauer: Validation of a morphogenesis model of Drosophila early development by a multi-objective evolutionary optimization algorithm. In EvoBio'09, Tübingen, 2009.