Multi-objective evolutionary algorithms for the identification of master regulators in pancreatic cancer

Claude Pasquier[1]

May 31, 2017

[1]Team SPARKS - I3S - Sophia Antipolis

claude.pasquier@unice.fr

Objective

Highlighting main regulators of disturbed signaling pathways involved in cancer cell phenotype.

Several criteria:

- altered expression between control PDAC[1] and CAF[2]-simulated cells
- acting in a coherent way to common biological processes
- involvment in the process under study

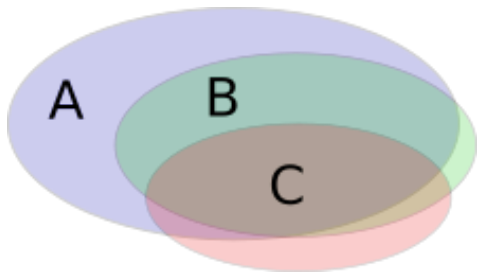
The analysis has to be done at both transcriptomic and proteomic level

[1] Pancreatic Ductal Adenocarcinoma
[2] Cancer-Associated Fibroblasts

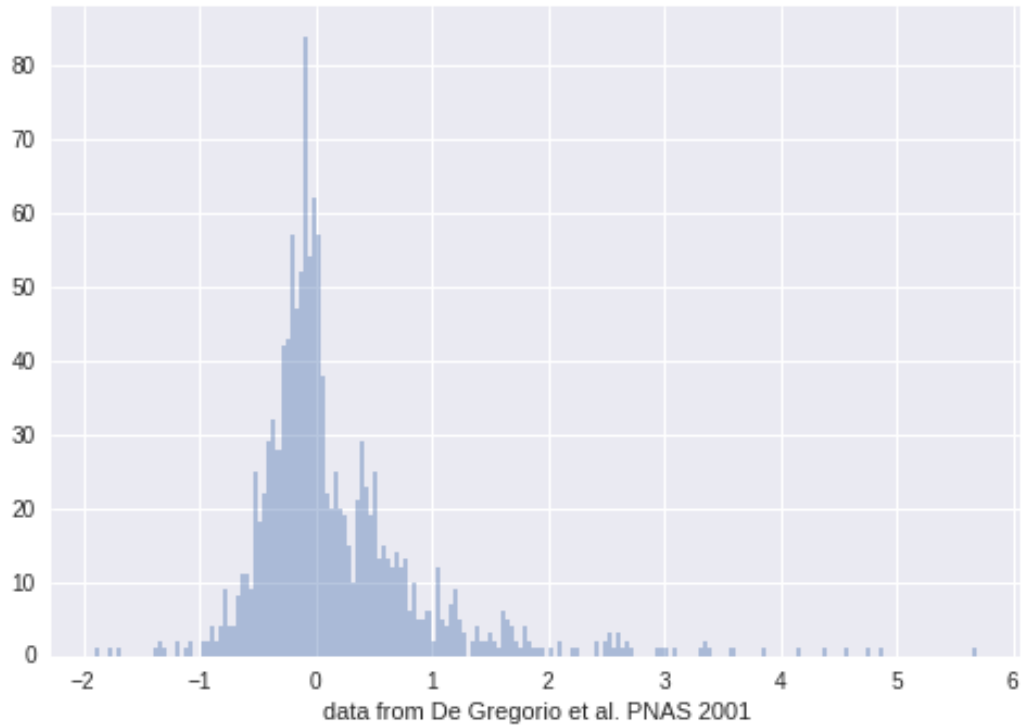
Standard Procedure

The processing is done sequentially by successive refinements

1. Select a set of differentially expressed molecules
   → candidate set A
2. Select members of A acting in a coordinated way
   → candidate set B
3. Select members of B involved in biological processes of interest
   → candidate set C

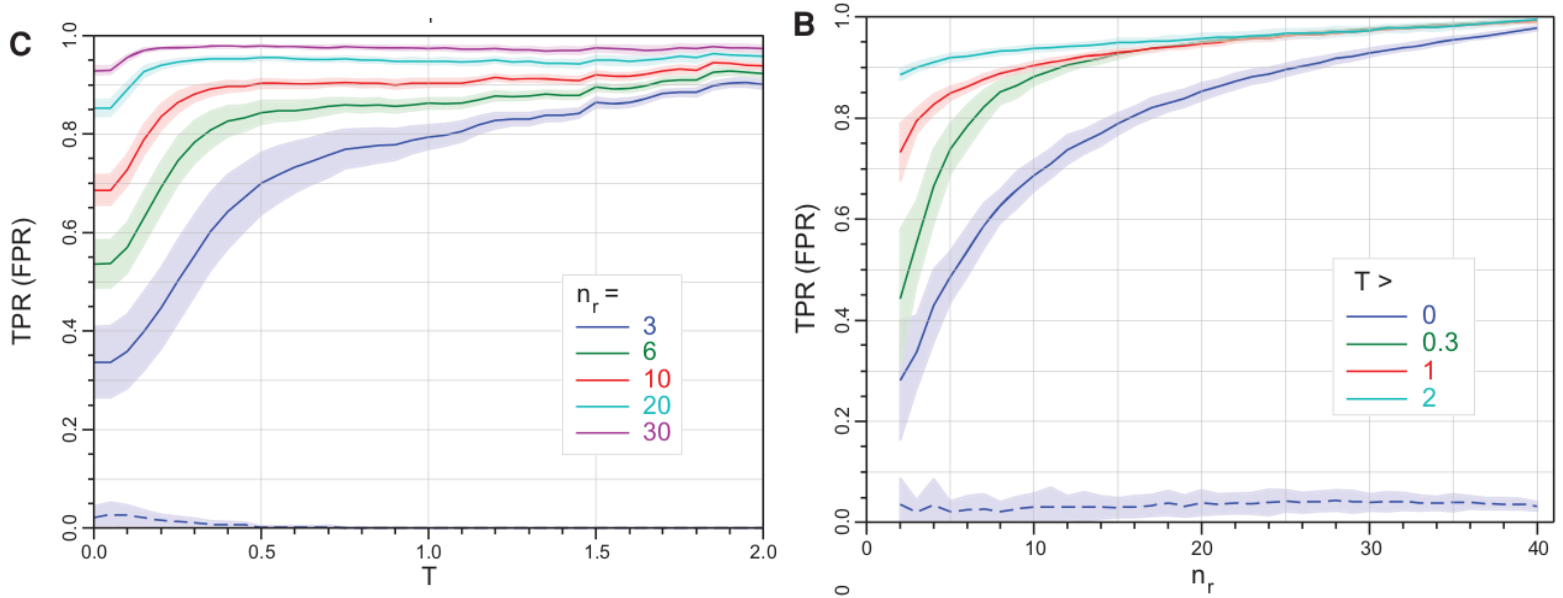


Ident. of differentially expressed molecules

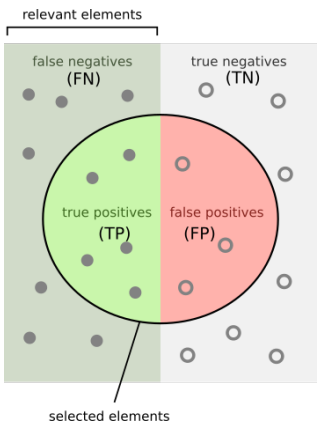


- maximization of true positives
- minimization of false positives
- how many replicates?

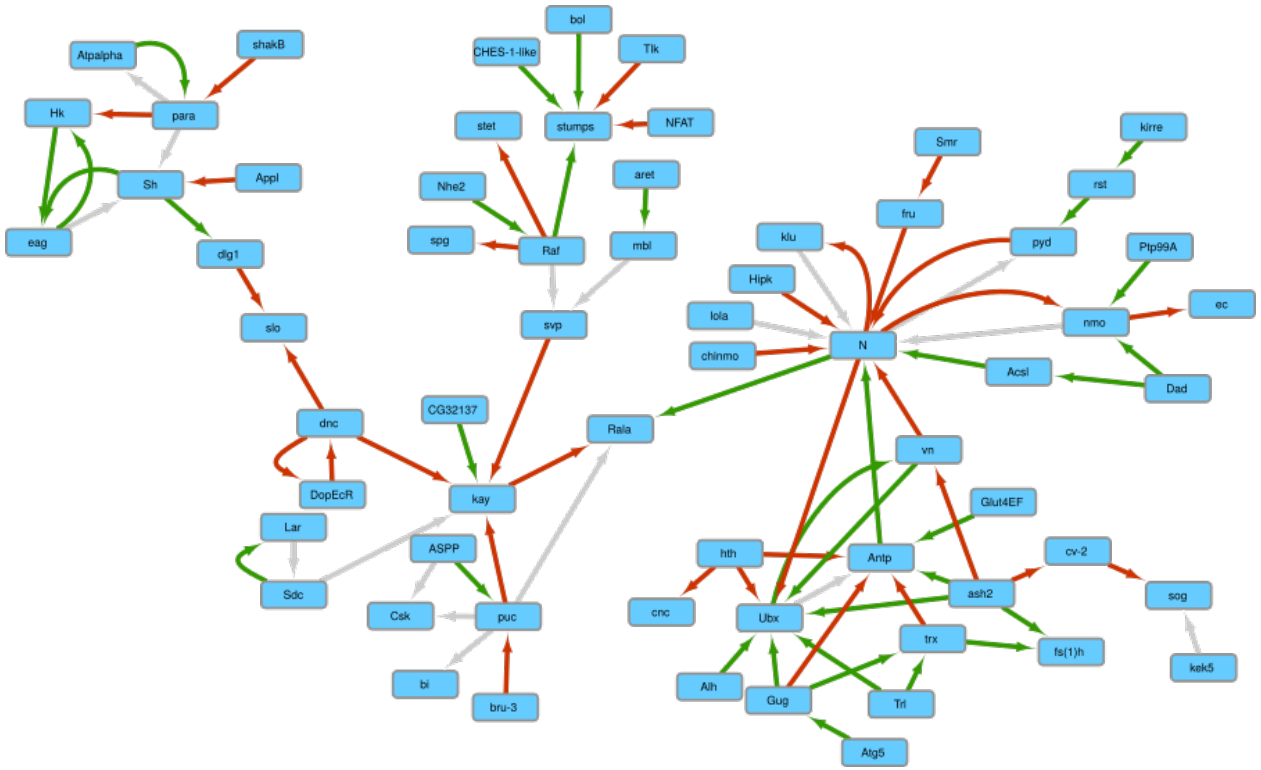
True positives vs. cutoff vs. replicates



study based on 48 biological replicates and 11 tools
Schurch et al. RNA 2016

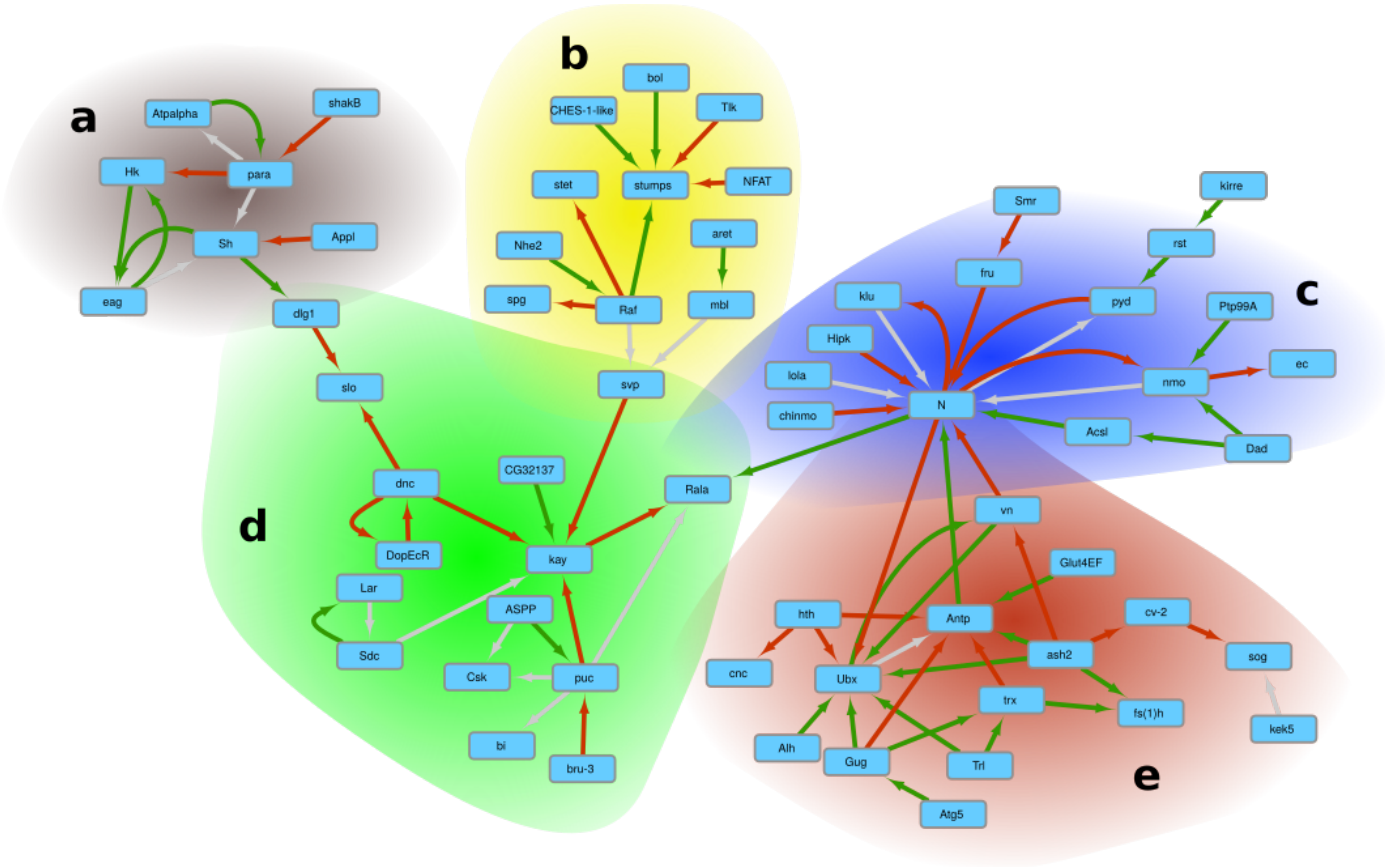


Identification of cluster of interacting genes



Pasquier et al. G3: Genes, Genomes, Genetics, 2017

Cluster annotation (with GO for example)

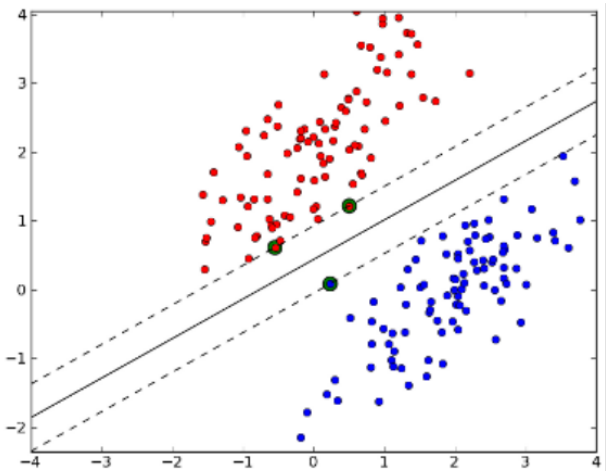


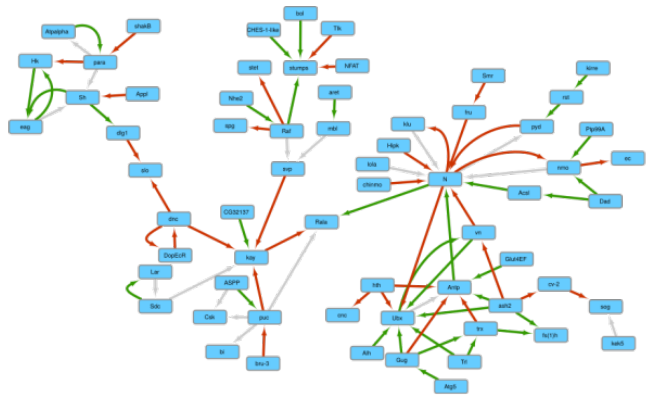
Pasquier et al. G3: Genes, Genomes, Genetics, 2017

Problem with the sequential approach

the reliability of the first step
shapes the accuracy of the following steps

Alternative: multi objective optimization





Annotations

expression based metrics

how appropriate are the selected genes to discriminate between control PDAC and CAF-simulated cells?

Metrics based on the accuracy of supervised classification methods (SVM)

network based metrics

how connected are the selected genes?

Metrics based on graph measures (diameter, density, etc)

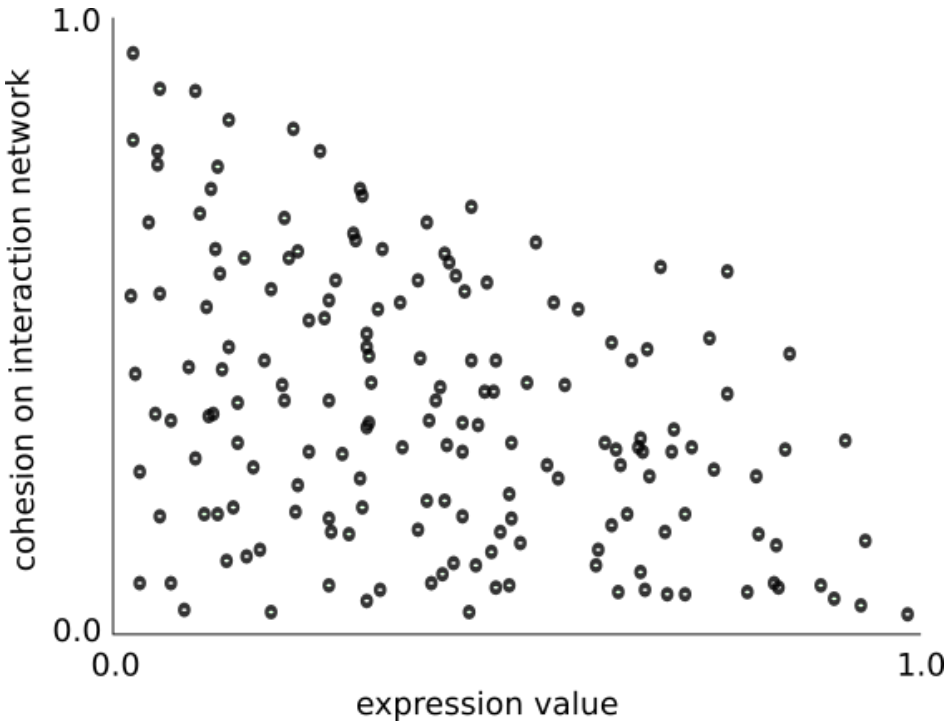
annotation based metrics

how relevant are the selected genes?

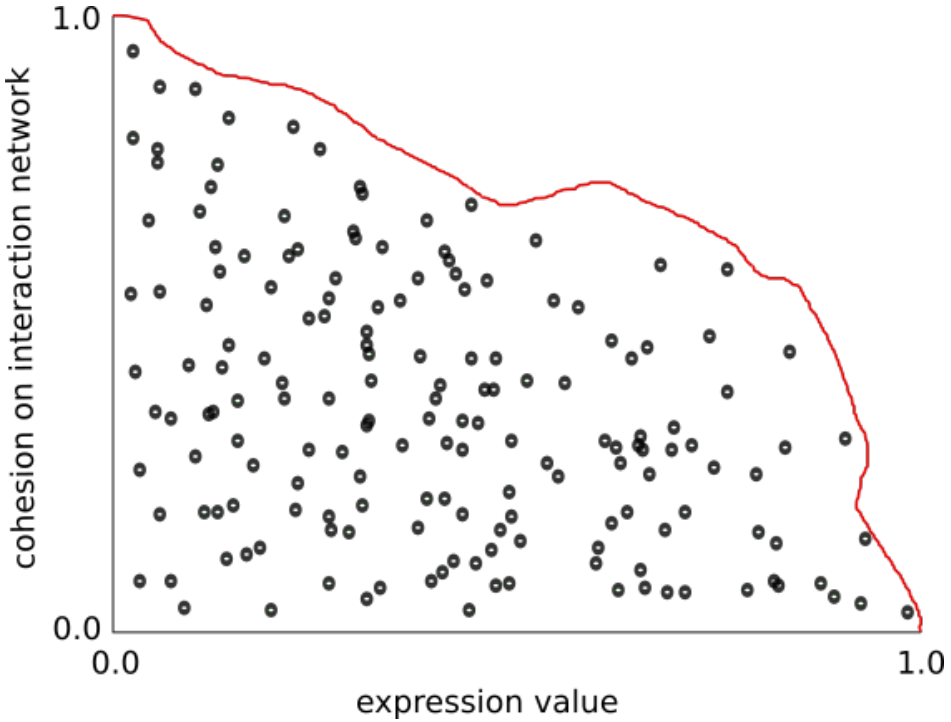
to be defined

performed at proteomic and transcriptomic level

Example with two objectives

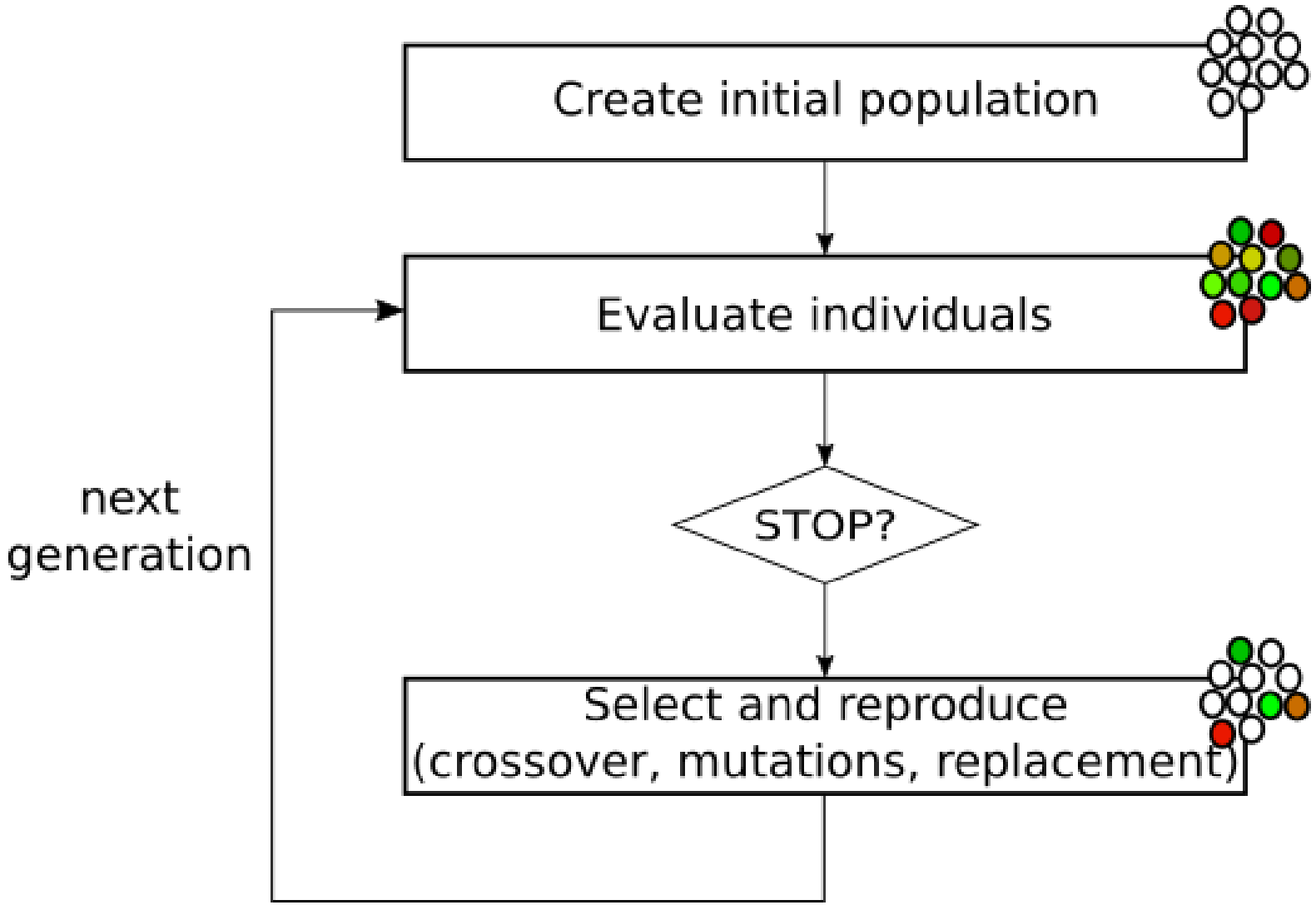


Example with two objectives



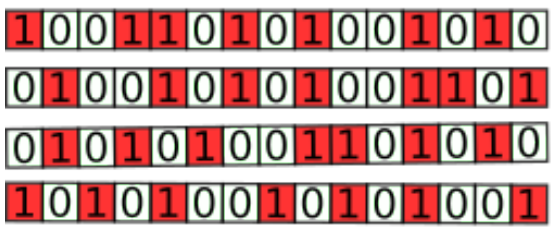
- potientially several best solutions
- solutions are probably not independent
- finding the best solutions involves evaluating 2^{nb_genes} possibilites
- our proposition: the use of evolutionary algorithms

Evolutionary Algorithms (EA)

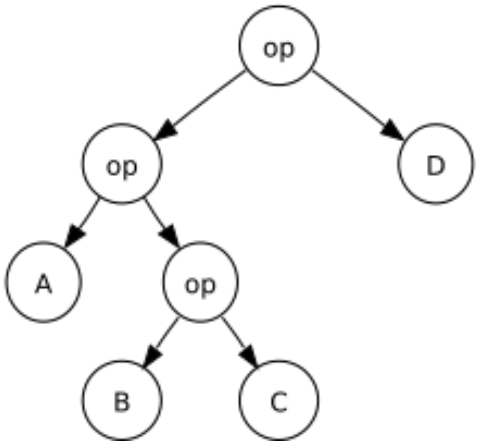


representation of individuals

Genetic Algorithms (GA)



Genetic Programming (GP)



Challenges and opportunities

Difficult points:

- number of features (tens of thousand vs. several hundred)
- number of objectives ($\sim$6 vs. $\sim$3)
- using GP for multi objective optimization is an open research area
- the design of good metrics

However, the specificities of the problem gives us opportunities:

- to tune initialization step and operators (crossover, mutation) using background knowledge
- to combine stochastic global search (EA) with deterministic local search (gradient descent)
- to use hashing-based approximation for neighbor search (e.g. LSH - Locality-Sensitive Hashing)