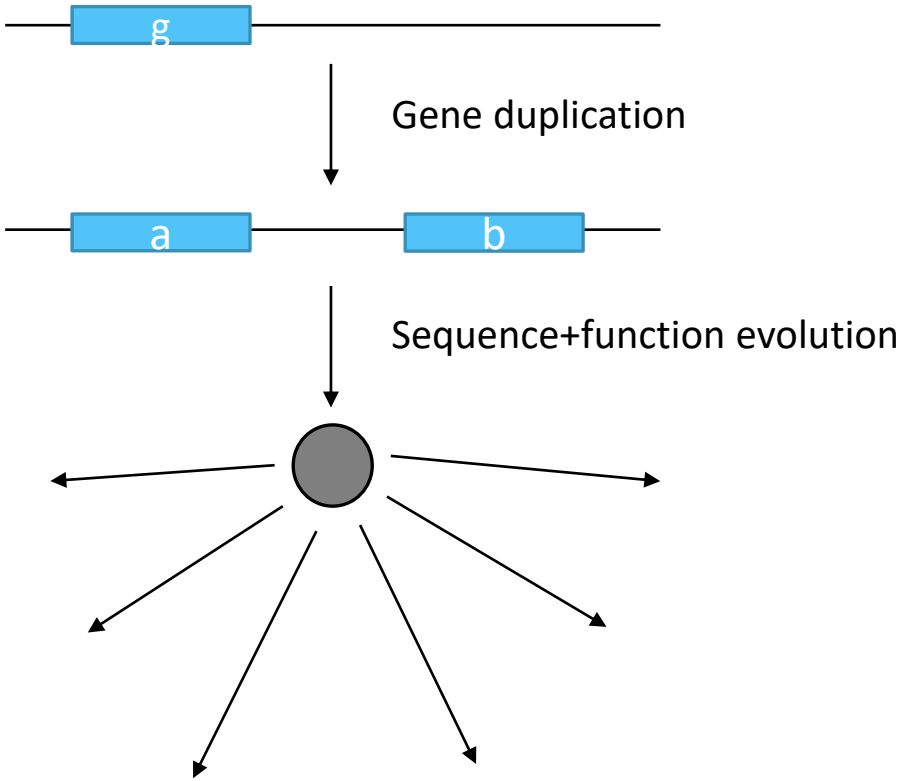
CLASSIFYING THE POST-DUPLICATION FATE OF PARALOGOUS GENES

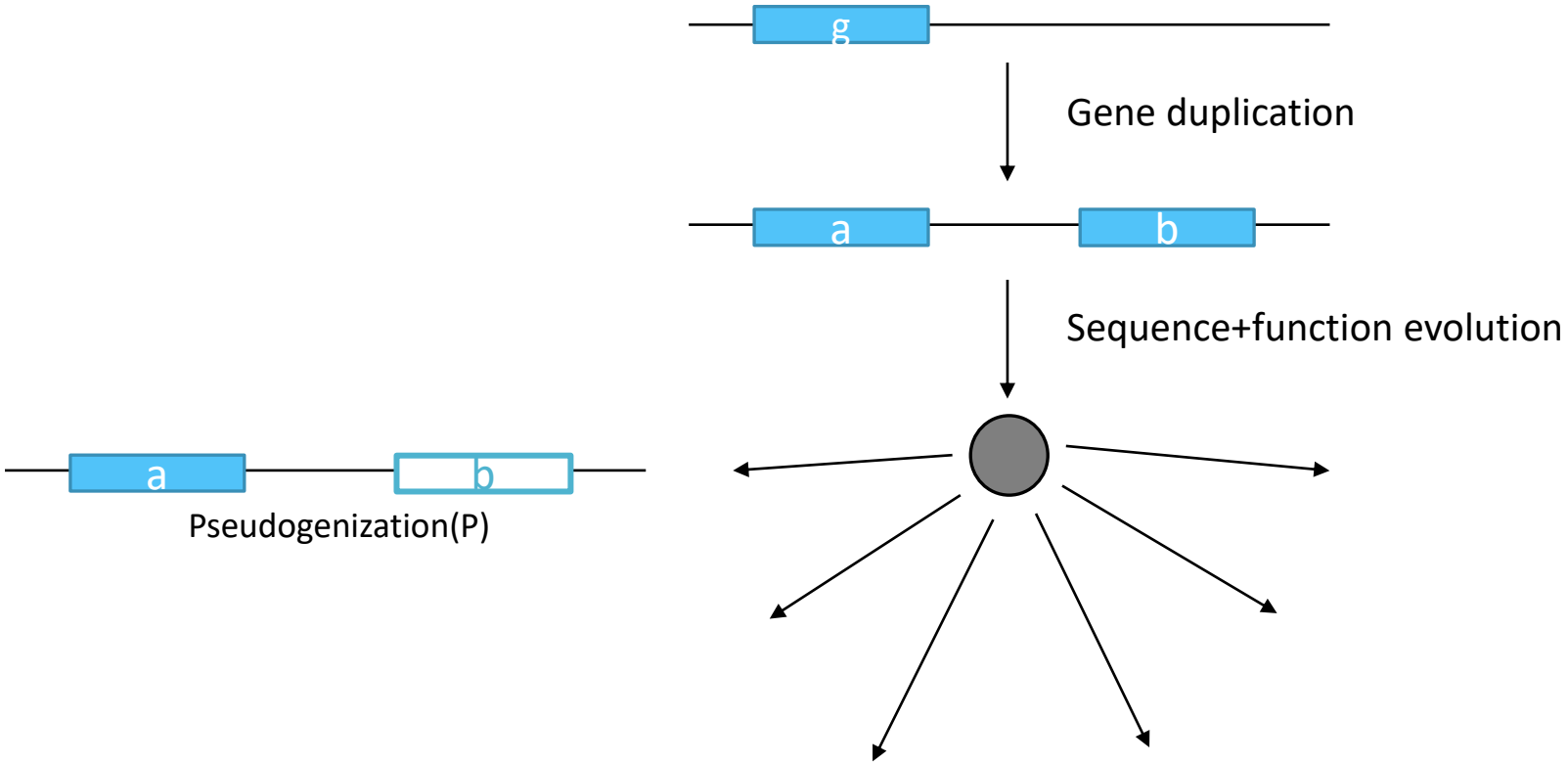
Reza Kalhor[1], Guillaume Beslon[2], Manuel Lafond[1], and Celine Scornavacca[3]

[1] Université de Sherbrooke, Canada
[2] Université de Lyon, France
[3] Université de Montpellier, France

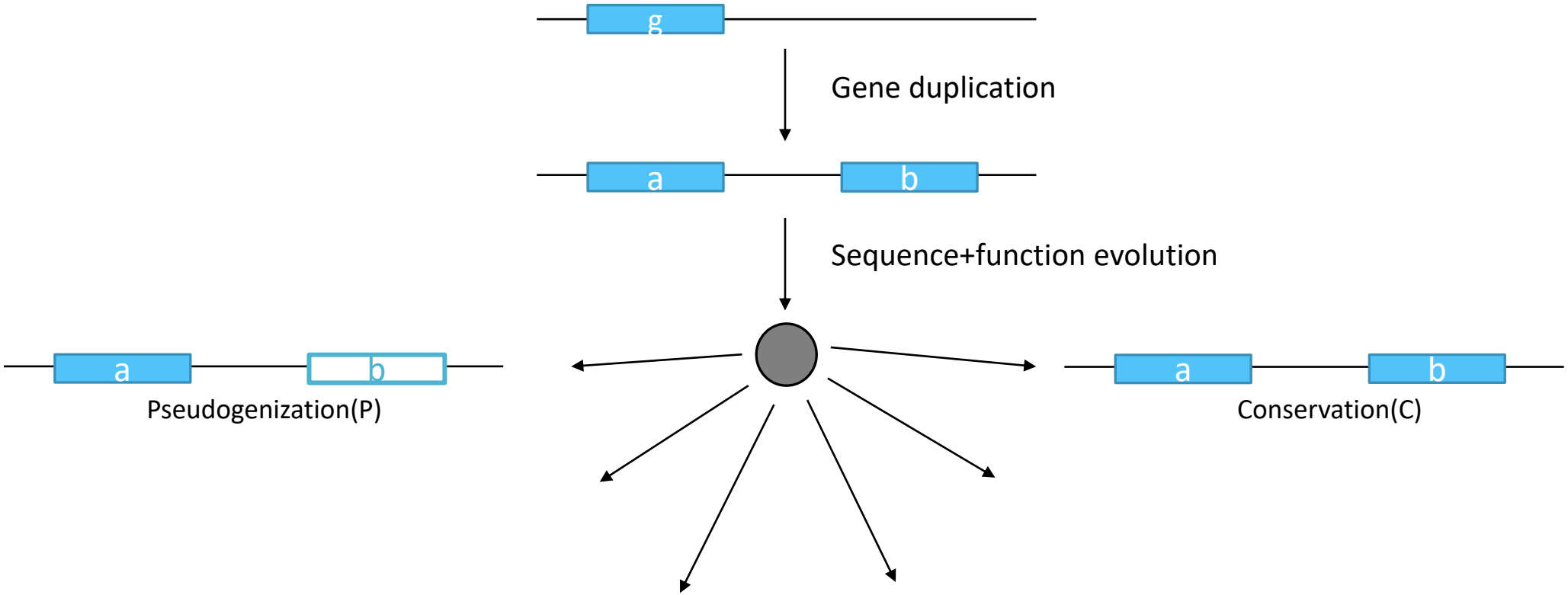
Post-duplication fates



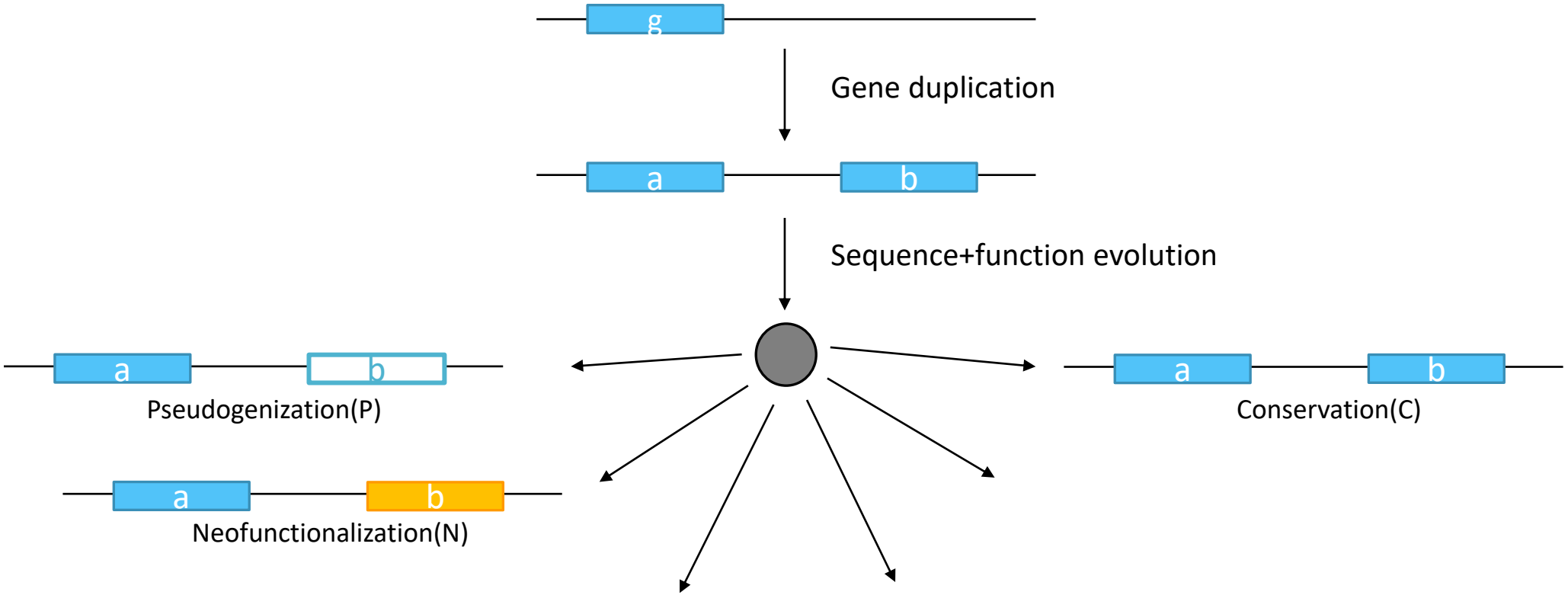
Post-duplication fates



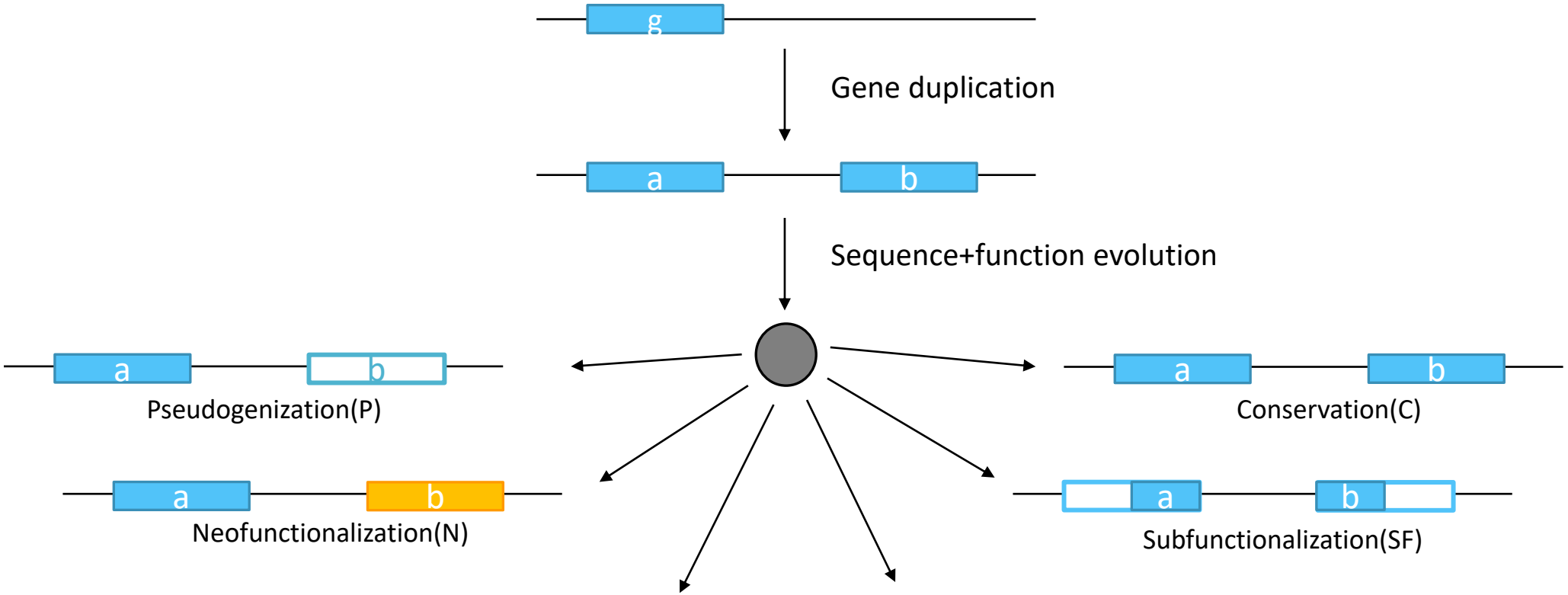
Post-duplication fates



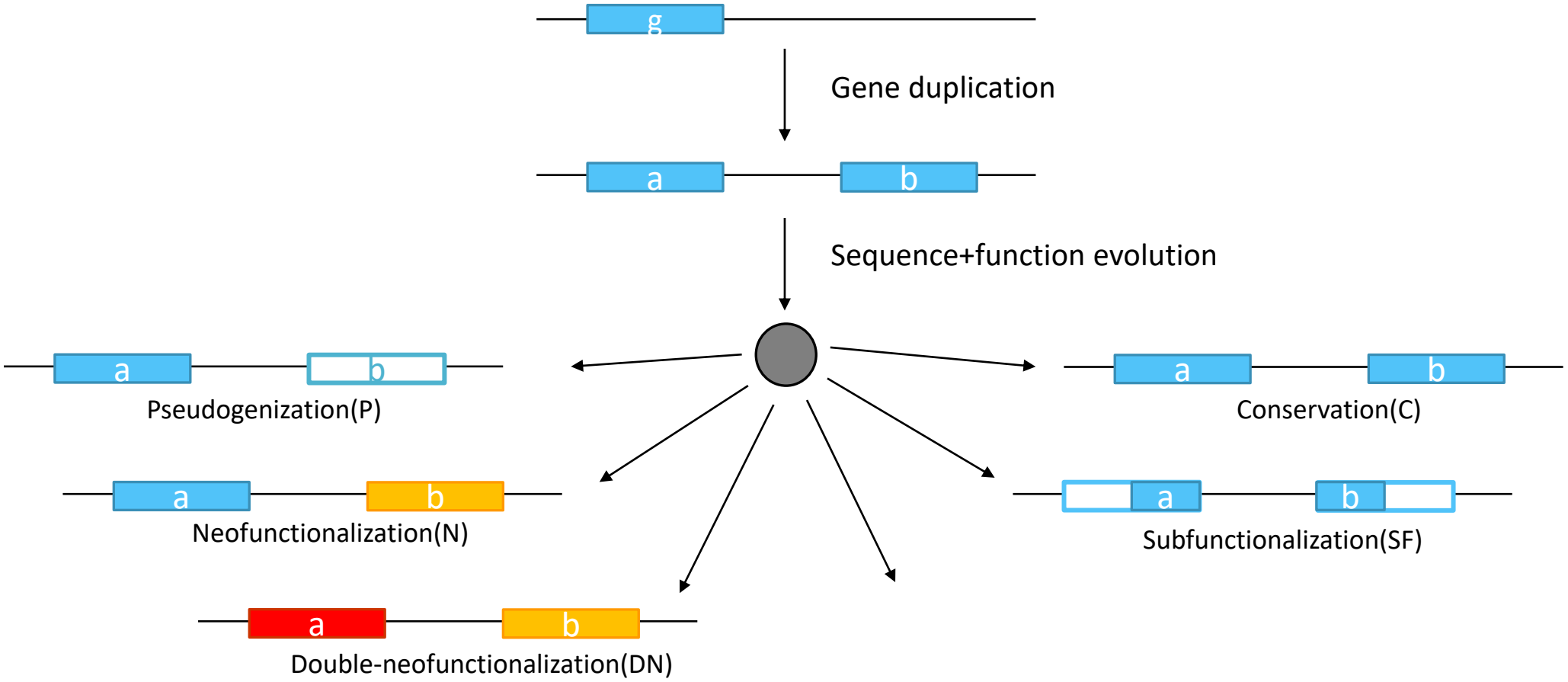
Post-duplication fates



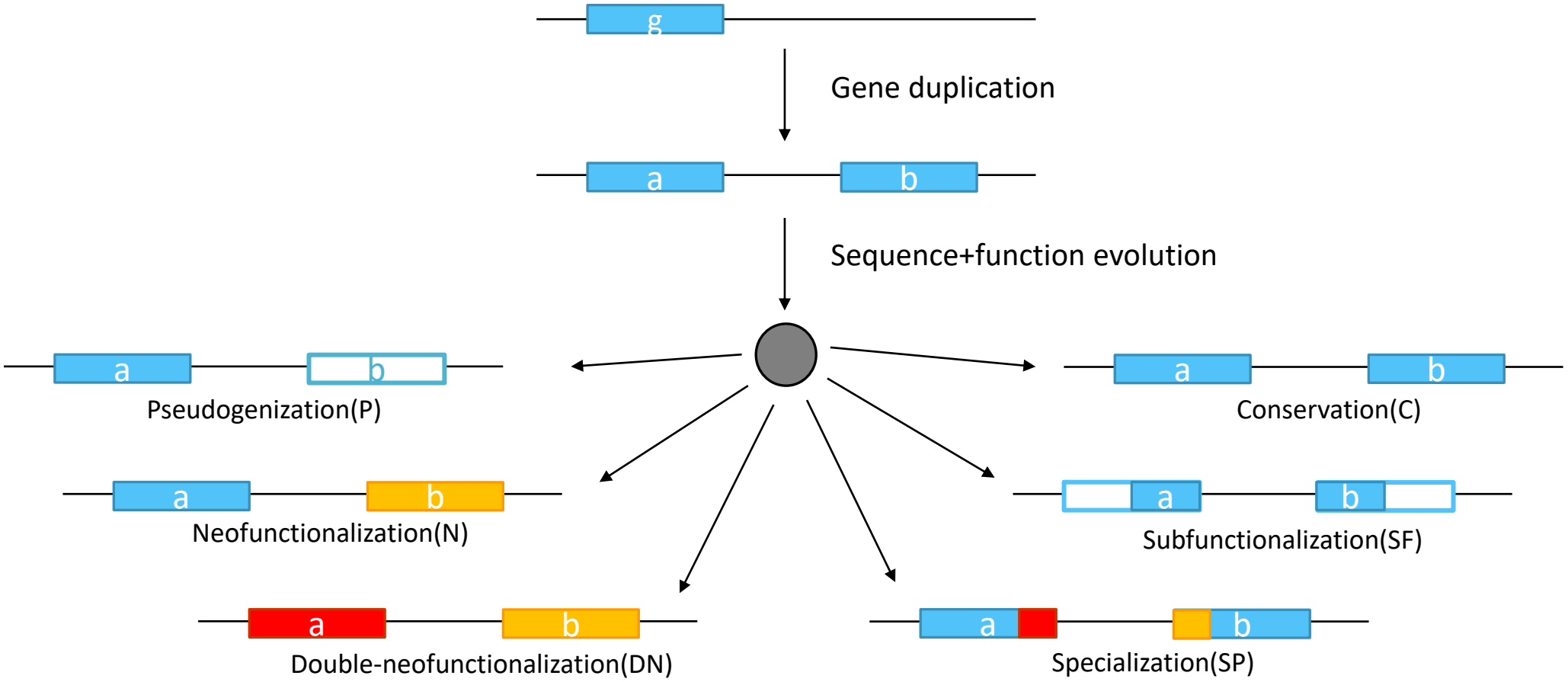
Post-duplication fates



Post-duplication fates



Post-duplication fates



Challenges and questions

- Are some fates **more frequent** than others? What do they depend on?
- Predicting fates is difficult, **even if we assume complete knowledge of gene functions**.
  - hybrid/unknown fates may occur

Challenges and questions

- Are some fates **more frequent** than others? What do they depend on?
- Predicting fates is difficult, **even if we assume complete knowledge of gene functions**.
  - hybrid/unknown fates may occur

- In this work:
- A mathematical framework to **quantify** post-duplication fates
- Validation on extensive population genetics simulations, using aevol (https://fr.wikipedia.org/wiki/Aevol).

Some related work

- Theoretical :
  - Lynch et al. (2000) : modeled genes as discrete sets of functions and proposed a population-based model of **subfunctionalization**.
  - Walsh (2003) : compared **pseudogenization** against other fates.
  - Stark et al. (2017) : compared **subfunctionalization** and **pseudogenization** using a mechanistic model based on Markov chain.
  - Lafond et al. (2018) : discussed theoretical impacts of **neofunctionalization** on orthology prediction.
  - Diao et al. (2020) : used Markov chains to predict the evolution of gene families undergoing duplications, loss, and partial gain/loss of function.

Some related work

- Practical :
  - Jaillon et al. (2004), and Brunet et al. (2006) : focused on **pseudogenization**, based on sequence comparisons and homology detection, showing that this fate is very likely in certain species.
  - Gu et al. (2004), Conant et al. (2008), and Nguyen et al. (2014) : studied the role of **neofunctionalization** and functional changes that can result in neofunctionalization.
  - Assis et al. (2013) : on drosophilia, distinguish between the fates (**neofunctionalization, subfunctionalization, conservation and specialization**) using **Euclidean distances between gene expression profiles**.

Some related work

- Practical :
  - Jaillon et al. (2004), and Brunet et al. (2006) : focused on **pseudogenization**, based on sequence comparisons and homology detection, showing that this fate is very likely in certain species.
  - Gu et al. (2004), Conant et al. (2008), and Nguyen et al. (2014) : studied the role of **neofunctionalization** and functional changes that can result in neofunctionalization.
  - Assis et al. (2013) : on drosophilia, distinguish between the fates (**neofunctionalization, subfunctionalization, conservation and specialization**) using **Euclidean distances between gene expression profiles**.
  - At ISMB: *Unveiling the Functional Fate of Duplicated Genes Through* ***Expression Profiling***
    - ***Least Divergent Orthologs conjecture.***
    - **** also just after me***

Our mathematical framework

How we model a gene (ideally)

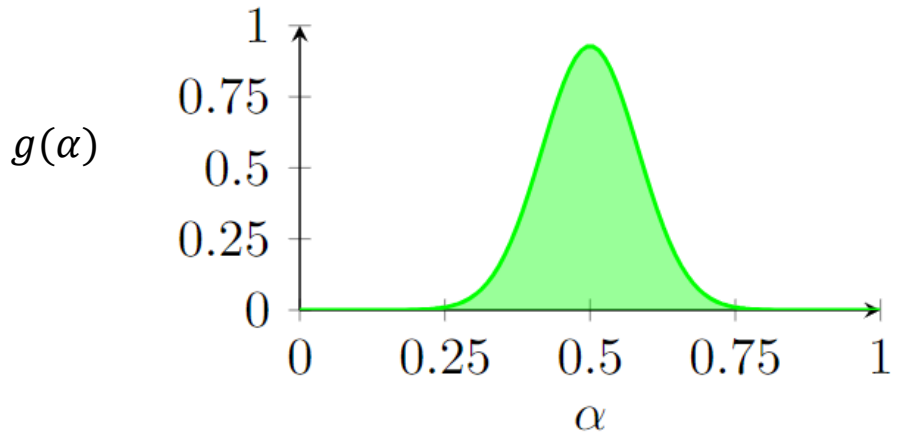
- Assume a set of functions $\mathcal{F}$ which can be discrete or continuous.
- $\mathcal{F}$ could be discrete, e.g. $\mathcal{F} = \{\text{"amino acid binding"}, \text{"transporter protein"}, \text{"GO:0016597"}, \ldots\}$
- $\mathcal{F}$ could be continuous, e.g., $\mathcal{F} = [0 \,..\, 1]$
- A gene g contributes to, or expresses, some functions of $\mathcal{F}$ to some degree. A gene is modeled as a function:

$$g : \mathcal{F} \rightarrow \mathbb{R}$$

- $g(\alpha)$ represents the contribution level of g to function $\alpha \in \mathcal{F}$
  - for example, expression level
  - can be NEGATIVE → inhibition

Example

- $g(\alpha)$ represents the contribution level of g to function $\alpha \in \mathcal{F}$



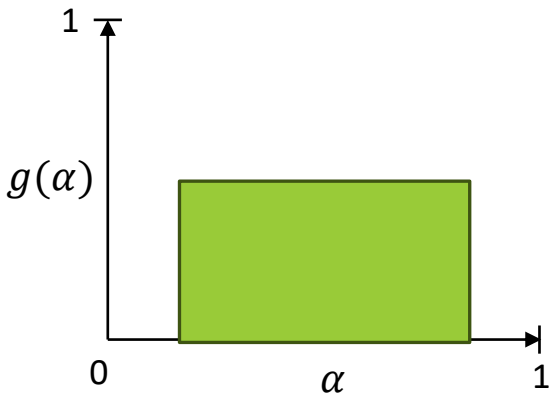
$g(0.5) = 1$ --> g contributes maximally to function "0.5"

$g(0) = 0$ --> g does not contribute to function "0"

$g(0.375) = 0.5$

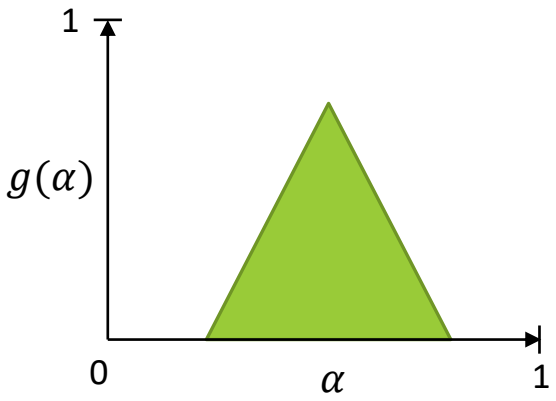
Example

- $g(\alpha)$ represents the contribution level of g to function $\alpha \in \mathcal{F}$



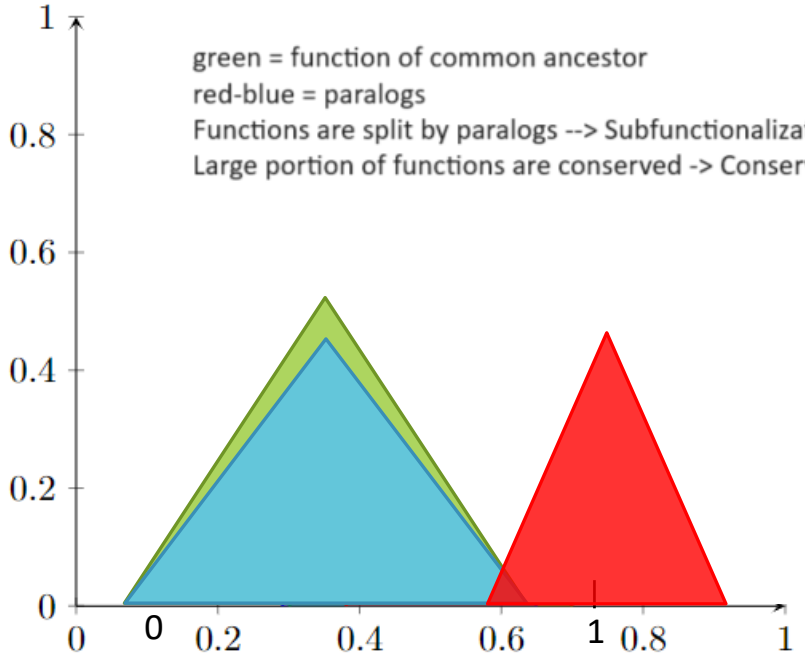
Example

- $g(\alpha)$ represents the contribution level of g to function $\alpha \in \mathcal{F}$



Inferring the fates of paralogs

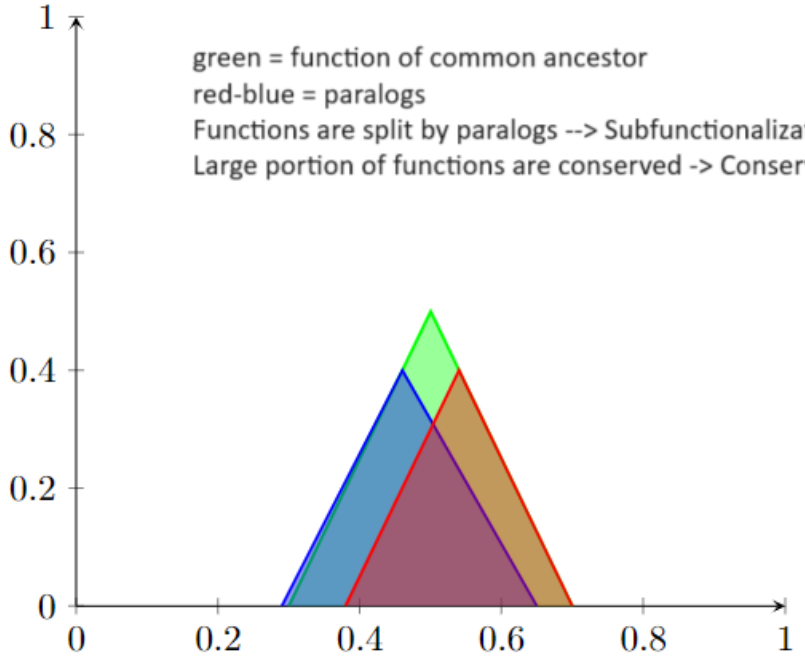
- Given two paralogs g, h, we predict an "amount" of each fate between 0 and 1



Prediction:
0.99 Neofunc
0.01 Conservation

Inferring the fates of paralogs

- Given two paralogs g, h, we predict an "amount" of each fate between 0 and 1



Prediction:
0.5 Subfunc
0.5 Conservation

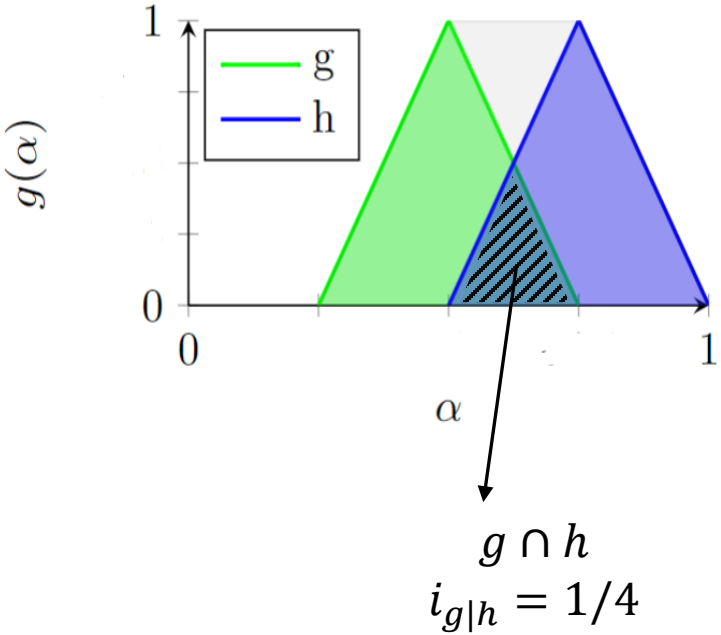
Inferring the fates of paralogs

- Consider paralogous genes g and h. Assume knowledge of all their functional contributions.

- Variables of interest:
  - $g \cap h$: function intersection of g and h. Essentially an area.
  - $i_{g|h}$: proportion of the area of g that is covered by h.

$$i_{g|h} = \frac{area(g \cap h)}{area(g)}$$

  - $g + h$: function addition of g and h.



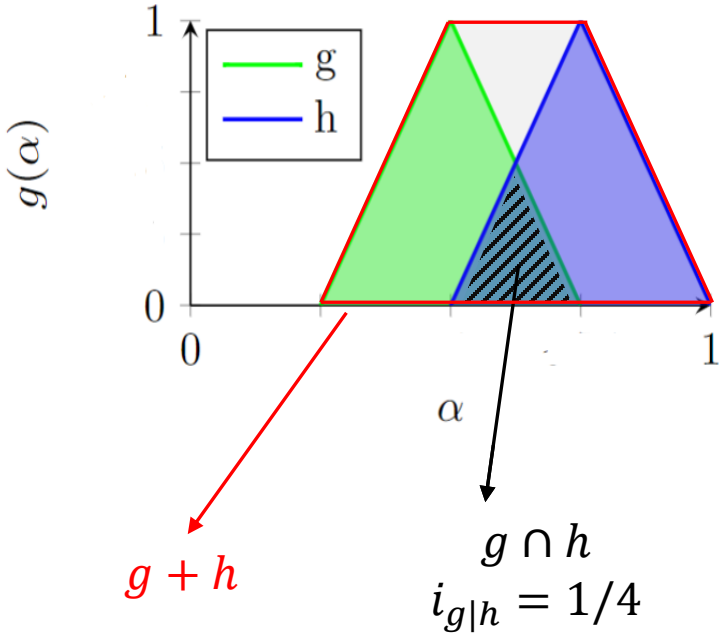
Inferring the fates of paralogs

- Consider paralogous genes g and h. Assume knowledge of all their functional contributions.

- Variables of interest:
  - $g \cap h$: function intersection of g and h. Essentially an area.
  - $i_{g|h}$: proportion of the area of g that is covered by h.

$$i_{g|h} = \frac{area(g \cap h)}{area(g)}$$

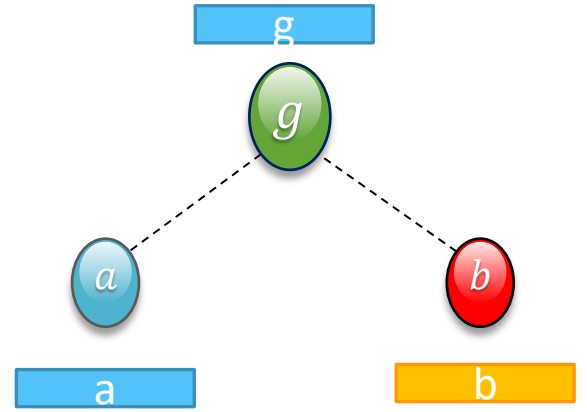
  - $g + h$: function addition of g and h.



Fate	Formula
Pseudogenization (P)	$P_a = i_{a\|g} \cdot (1 - \frac{i_{g\|a}}{\rho})$ $P_b = i_{b\|g} \cdot (1 - \frac{i_{g\|b}}{\rho})$ $P = \max(0, P_a, P_b)$
Neofunc. (N)	$N_a = (1 - i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|b}$ $N_b = (1 - i_{b\|g}) \cdot \delta_\nu(i_{a\|g}) \cdot i_{g\|a}$ $N = \max(N_a, N_b) \cdot (1 - P)$
Double-neo. (DN)	$DN = (1 - i_{a\|g})(1 - i_{b\|g})(1 - i_{g\|a+b})(1 - P)$
Conservation (C)	$C = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot (1 - \delta_{0.5}(i_{a+b\|g})) \cdot (1 - P)$
Subfunc. (SF)	$SF = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot \delta_{0.5}(i_{a+b\|g}) \cdot (1 - P)$
Specialization (SP)	$SP = i_{g\|a+b} \cdot (1 - \delta_\nu(i_{a\|g})) \cdot (1 - \delta_\nu(i_{b\|g})) \cdot (1 - P)$

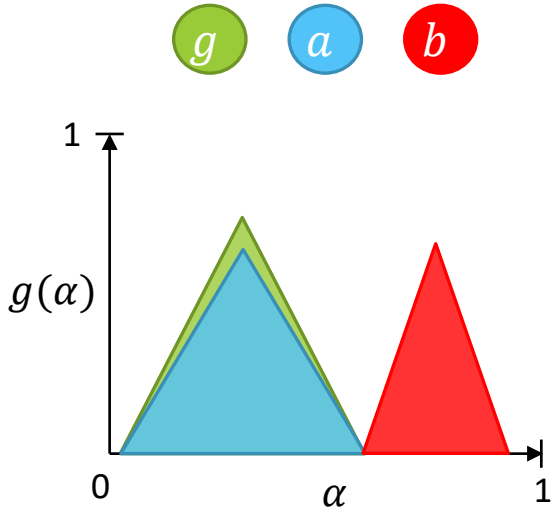
Calculating the probability of fates

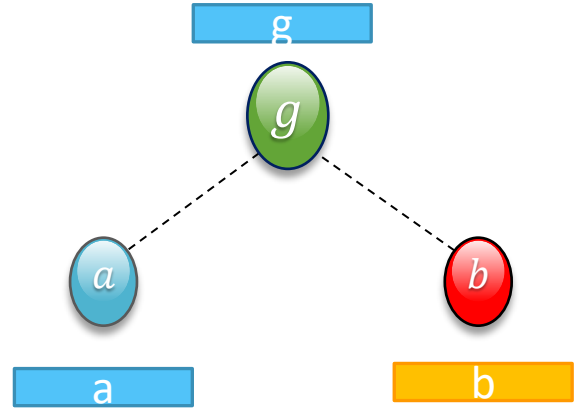
- Neofunctionalization (N) of b happens when:
- paralog a has conserved the function of ancestor g, but
- paralog b has diverged from ancestor g.



Calculating the probability of fates

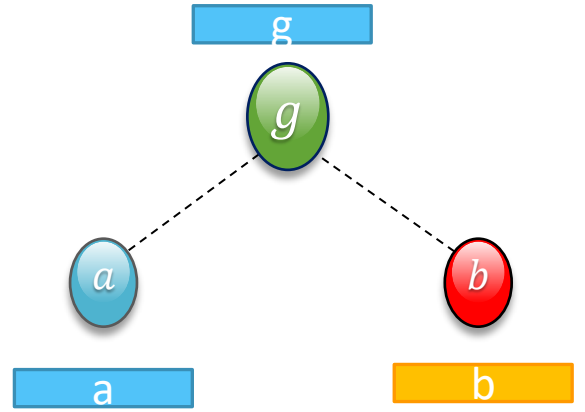
- Neofunctionalization (N) of b happens when:
- paralog a has conserved the function of ancestor g, but
- paralog b has diverged from ancestor g.

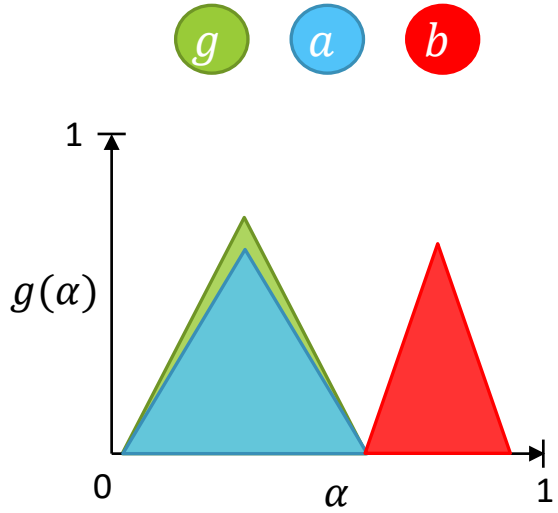




Calculating the probability of fates

- Neofunctionalization (N) of b happens when:
- paralog a has conserved the function of ancestor g, but
- paralog b has diverged from ancestor g.





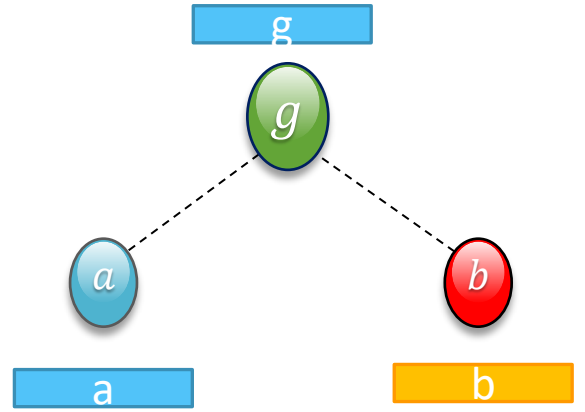
$i_{b|g} = 0$ *(b not covered by g)*

$i_{g|a} = 1$ *(g covered by a)*

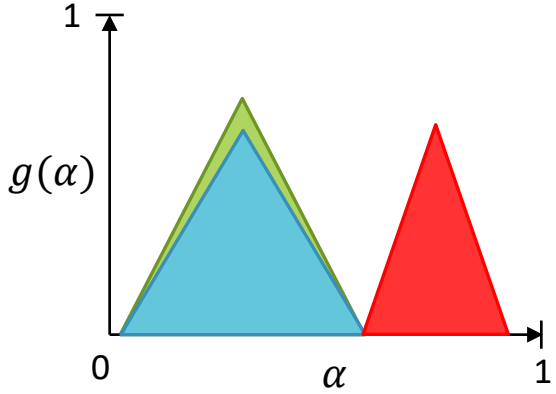
$i_{a|g} = 1$ *(a covered by g)*

Calculating the probability of fates

- Neofunctionalization (N) of b happens when:
- paralog a has conserved the function of ancestor g, but
- paralog b has diverged from ancestor g.



$$N_b = (1 - i_{b|g}).(i_{a|g}).i_{g|a}$$



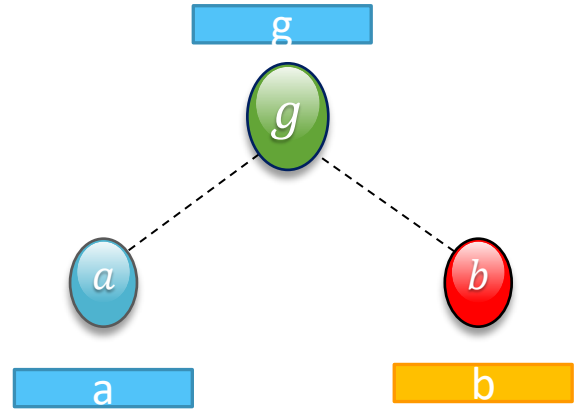
$i_{b|g} = 0$ *(b not covered by g)*

$i_{g|a} = 1$ *(g covered by a)*

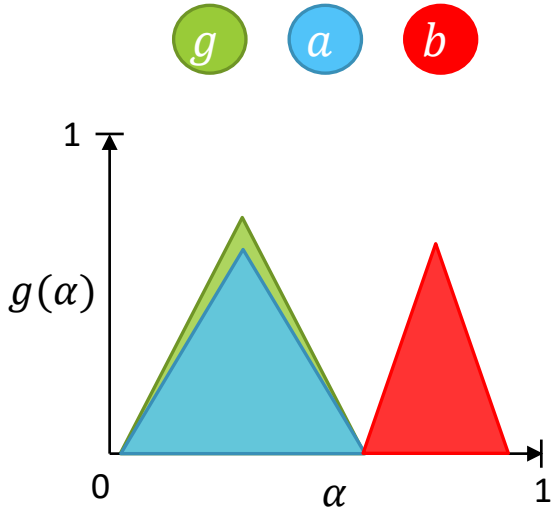
$i_{a|g} = 1$ *(a covered by g)*

Calculating the probability of fates

- Neofunctionalization (N) of b happens when:
- paralog a has conserved the function of ancestor g, but
- paralog b has diverged from ancestor g.



$$N_a = (1 - i_{b|g}).(i_{a|g}).i_{g|a}$$
$$N_b = (1 - i_{b|g}).(i_{a|g}).i_{g|a}$$
$$N = \max(N_a, N_b)$$



$i_{b|g} = 0$ *(b not covered by g)*
$i_{g|a} = 1$ *(g covered by a)*
$i_{a|g} = 1$ *(a covered by g)*

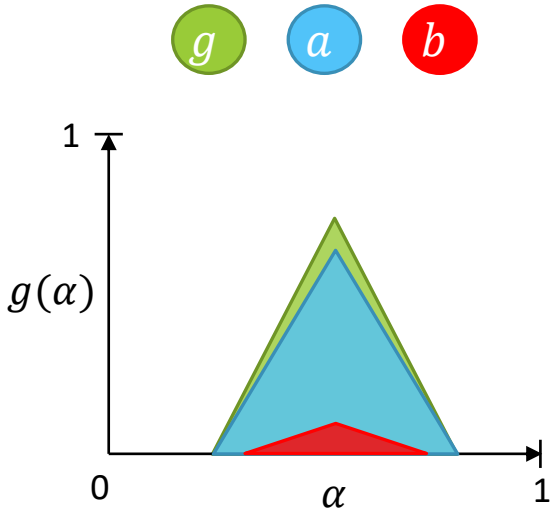
Calculating the probability of fates

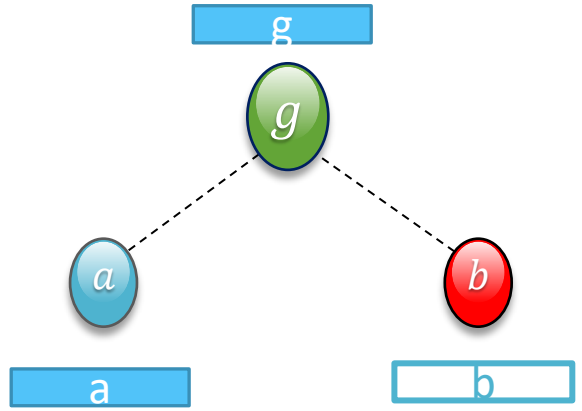
- Pseudogeneization(P)

$$P_a = i_{a|g}.(1 - i_{g|a})$$

$$P_b = i_{b|g}.(1 - i_{g|b})$$

$$P = \max(0, P_a, P_b)$$

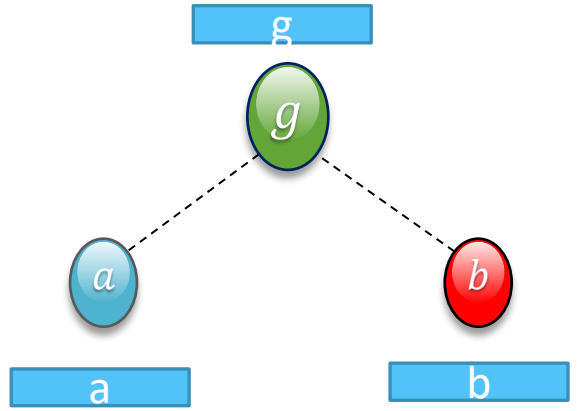


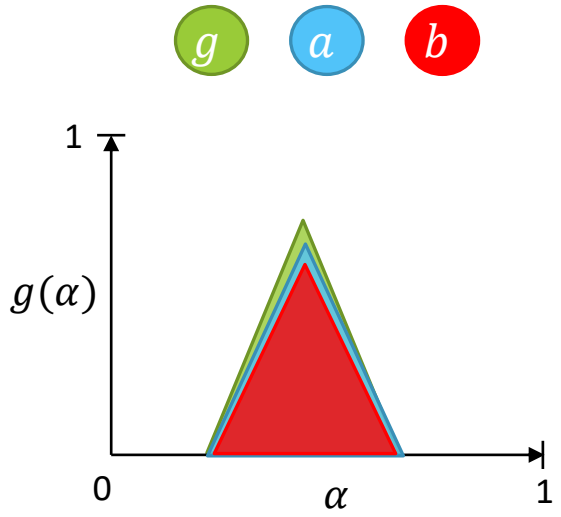


Calculating the probability of fates

- Conservation(C)

$$C = (i_{a|g}).(i_{b|g}).(i_{g|a+b}).\left(1 - \delta_{0.5}(i_{a+b|g})\right).(1 - P)$$

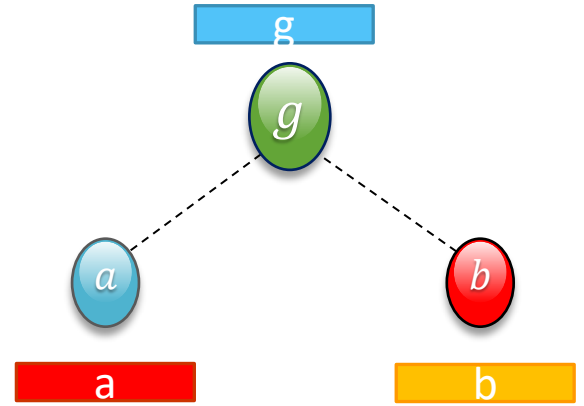


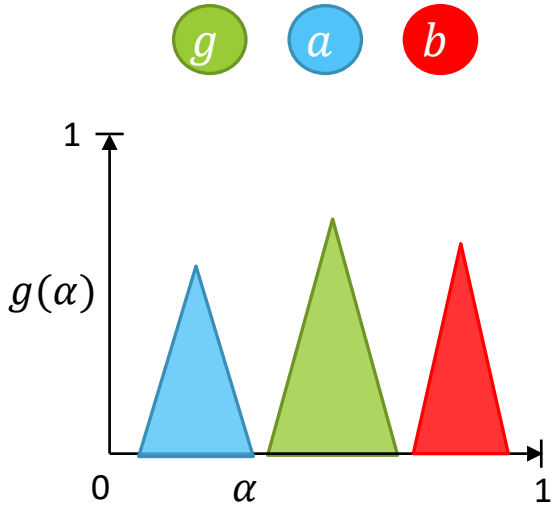


Calculating the probability of fates

- Double-neofunctionalization(DN)

$$DN = (1 - i_{a|g}).(1 - i_{b|g}).(1 - i_{g|a}).(1 - i_{g|b}).(1 - P)$$

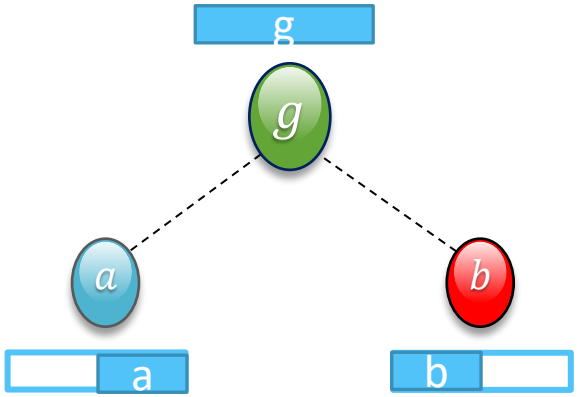


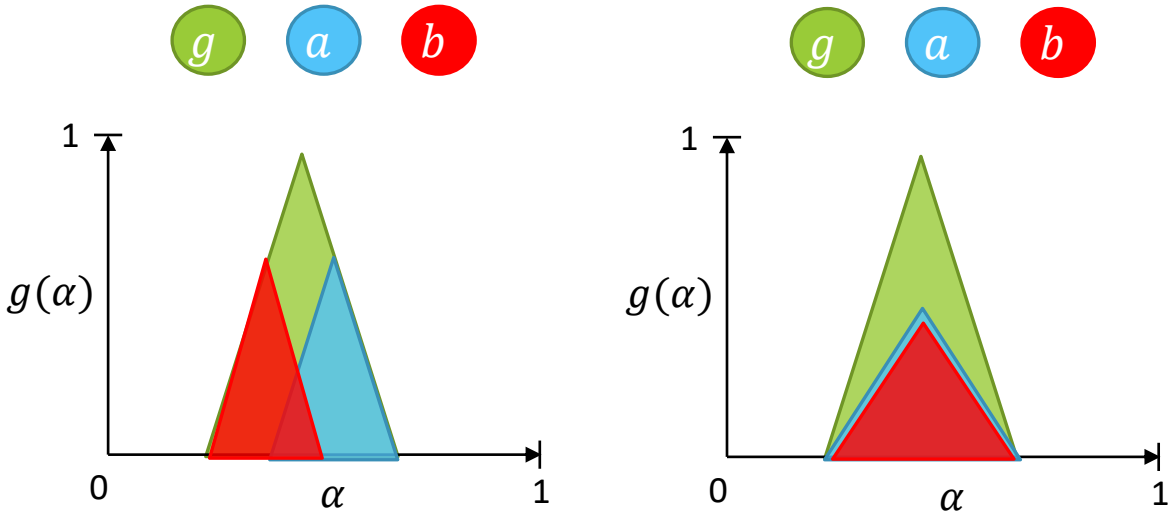


Calculating the probability of fates

- Subfunctionalization(SF)

$$SF = (i_{a|g}).(i_{b|g}).(i_{g|a+b}).\left(\delta_{0.5}(i_{a+b|g})\right).(1-P)$$





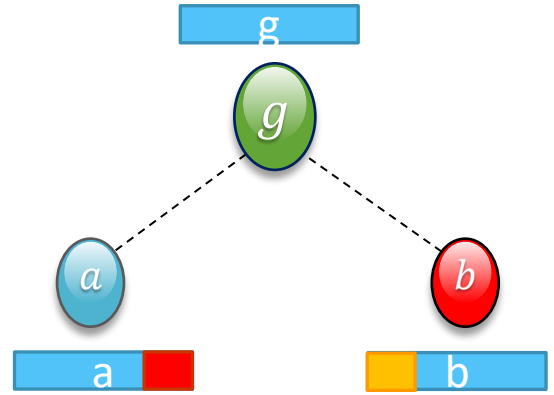
$$i_{a|g} = 1$$
$$i_{b|g} = 1$$
$$i_{g|a+b} = 1$$
$$i_{a+b|g} = 1$$
$$\left(\delta_{0.5}(i_{a+b|g})\right) = 1$$

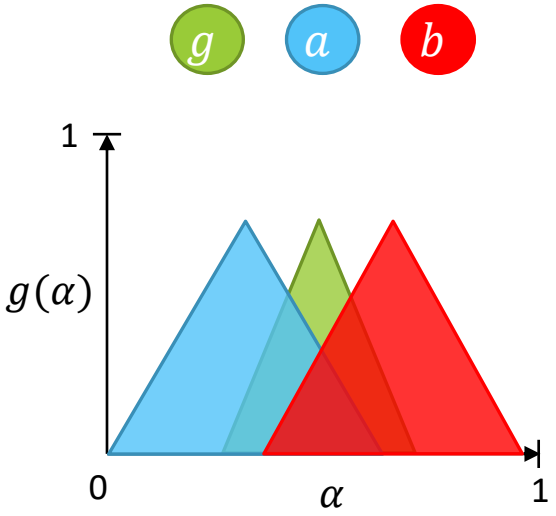
$$SF = 1$$

Calculating the probability of fates

- Specialization(SP)

$$SP = (1 - (i_{a|g})).(1 - (i_{b|g})).(i_{g|a+b}).(1 - P)$$





$$i_{a|g} = 0.25$$
$$i_{b|g} = 0.25$$
$$i_{g|a+b} = 1$$

$$SP = 1$$

Fate	Formula
Pseudogenization (P)	$P_a = i_{a\|g} \cdot (1 - \frac{i_{g\|a}}{\rho})$ $P_b = i_{b\|g} \cdot (1 - \frac{i_{g\|b}}{\rho})$ $P = \max(0, P_a, P_b)$
Neofunc. (N)	$N_a = (1 - i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|b}$ $N_b = (1 - i_{b\|g}) \cdot \delta_\nu(i_{a\|g}) \cdot i_{g\|a}$ $N = \max(N_a, N_b) \cdot (1 - P)$
Double-neo. (DN)	$DN = (1 - i_{a\|g})(1 - i_{b\|g})(1 - i_{g\|a+b})(1 - P)$
Conservation (C)	$C = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot (1 - \delta_{0.5}(i_{a+b\|g})) \cdot (1 - P)$
Subfunc. (SF)	$SF = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot \delta_{0.5}(i_{a+b\|g}) \cdot (1 - P)$
Specialization (SP)	$SP = i_{g\|a+b} \cdot (1 - \delta_\nu(i_{a\|g})) \cdot (1 - \delta_\nu(i_{b\|g})) \cdot (1 - P)$

Fate	Formula
Pseudogenization (P)	$P_a = i_{a\|g} \cdot (1 - \frac{i_{g\|a}}{\rho})$ $P_b = i_{b\|g} \cdot (1 - \frac{i_{g\|b}}{\rho})$ $P = \max(0, P_a, P_b)$
Neofunc. (N)	$N_a = (1 - i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|b}$ $N_b = (1 - i_{b\|g}) \cdot \delta_\nu(i_{a\|g}) \cdot i_{g\|a}$ $N = \max(N_a, N_b) \cdot (1 - P)$
Double-neo. (DN)	$DN = (1 - i_{a\|g})(1 - i_{b\|g})(1 - i_{g\|a+b})(1 - P)$
Conservation (C)	$C = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot (1 - \delta_{0.5}(i_{a+b\|g})) \cdot (1 - P)$
Subfunc. (SF)	$SF = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot \delta_{0.5}(i_{a+b\|g}) \cdot (1 - P)$
Specialization (SP)	$SP = i_{g\|a+b} \cdot (1 - \delta_\nu(i_{a\|g})) \cdot (1 - \delta_\nu(i_{b\|g})) \cdot (1 - P)$

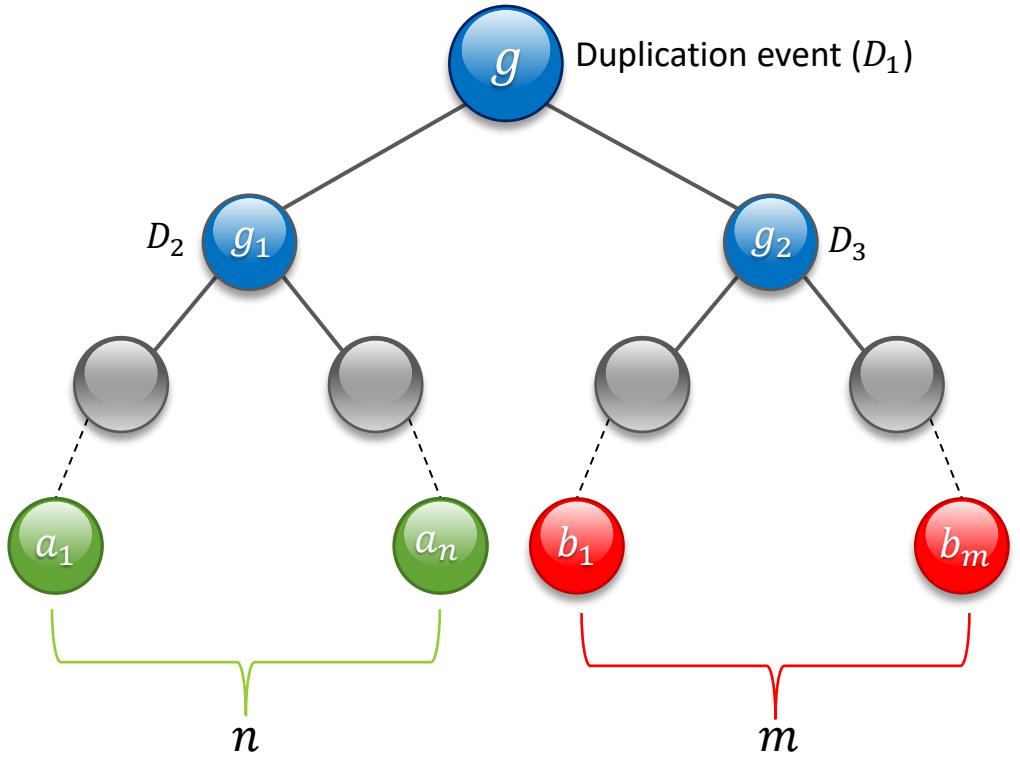
In experiments:
- sum-of-fates ≤ 1
- sum-of-fates not always 1
- one fate = 1 if and only if it "fits our expectation of that fate"

How we compute the fate of paralogs in a gene tree?

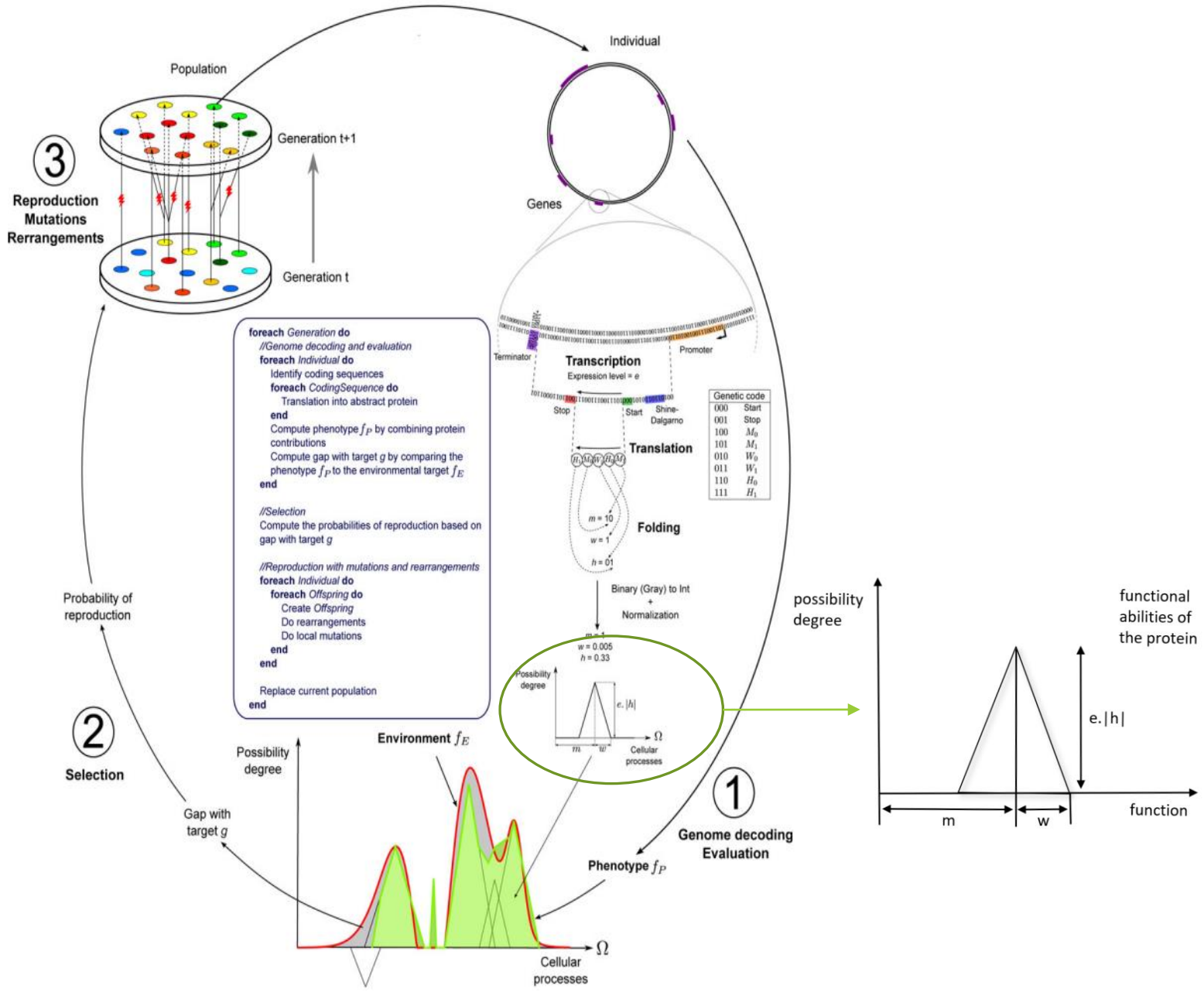
- To compute the probability of each fate, we need a triple combination of genes (g, a, b), g as lowest common ancestor and a, b as paralogous genes.
- To calculate the fate probabilities For duplication event D_1 :

$$ComputeFate\ (g, a_i, b_j),$$
$$where\ 1 \leq i \leq n, 1 \leq j \leq m.$$

$$P_x = \frac{\sum_{k=1}^{n \times m} P_{x_k}}{n \times m},$$
$$where\ x \in \{P, N, DN, C, SF, SP\}$$

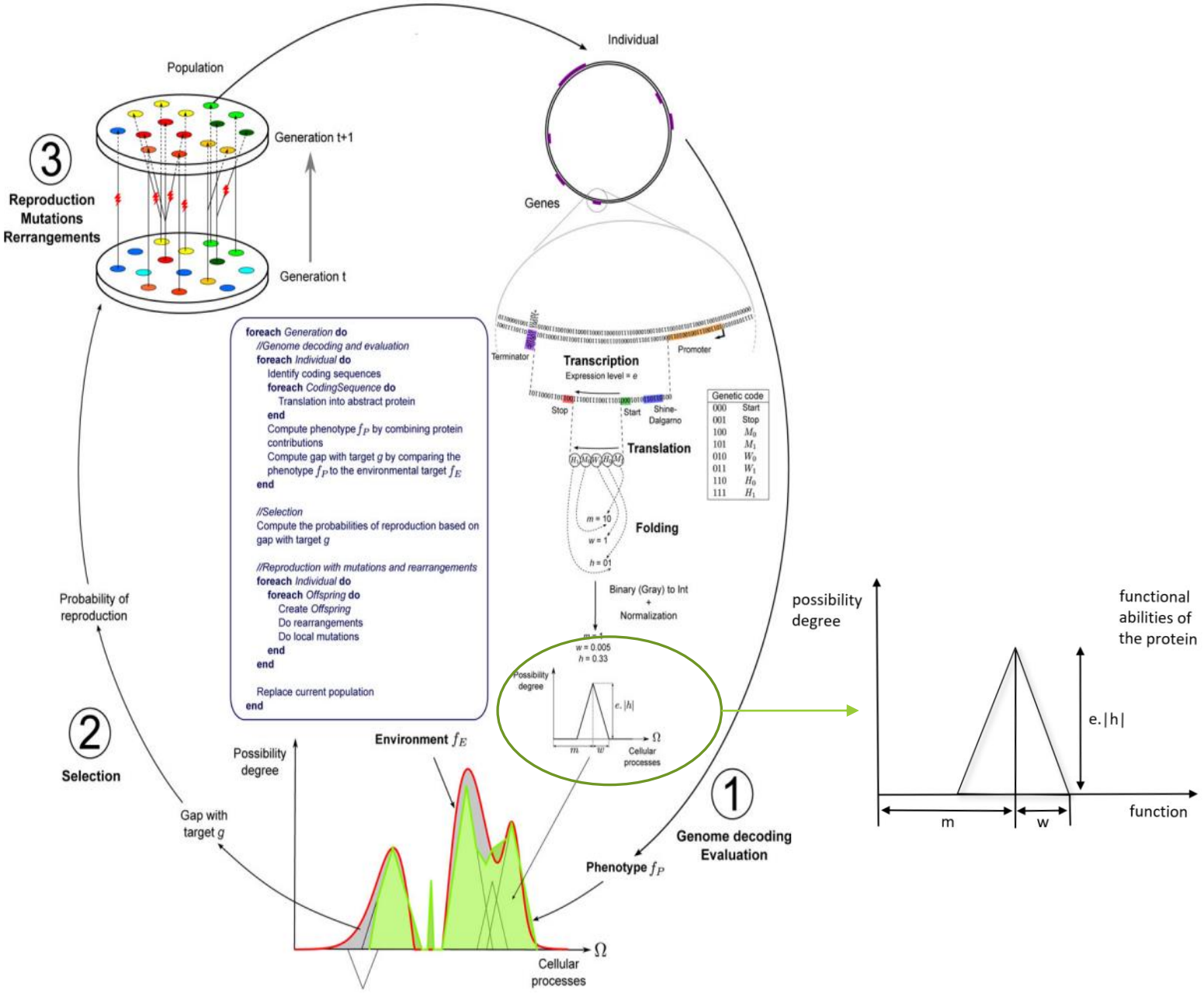


Simulation study using aevol



Aevol (a case study on artificial life)

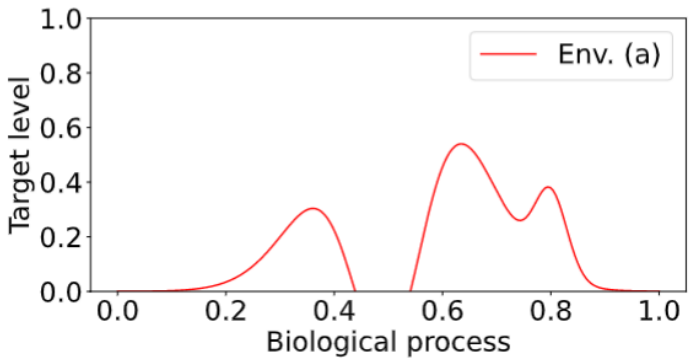
- Aevol : forward-in-time population genetics simulator.
- Simulates whole chromosomes. Has mutations, indels, inversions, transpositions, **duplications**.
- Genes appear if mutations happen to create **promoter + RNA + protein**.
- Protein **sequences encode functions** using gray codes.
  - Function = triangle

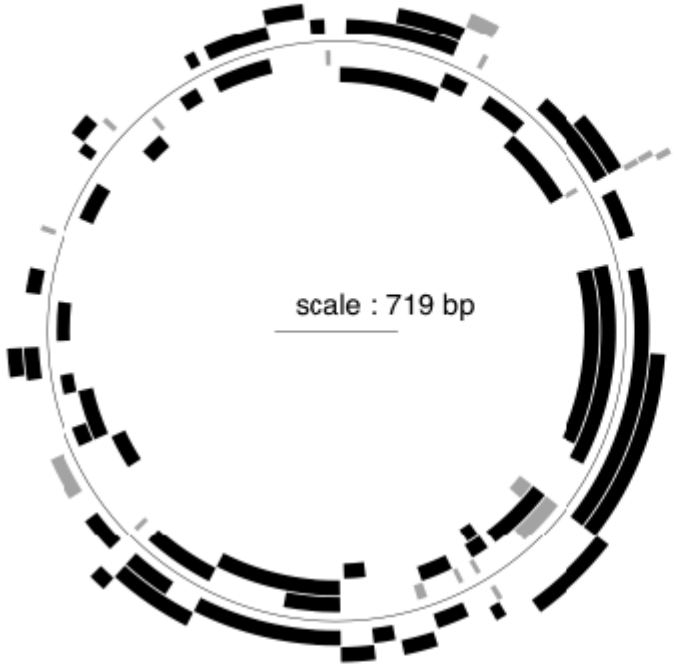


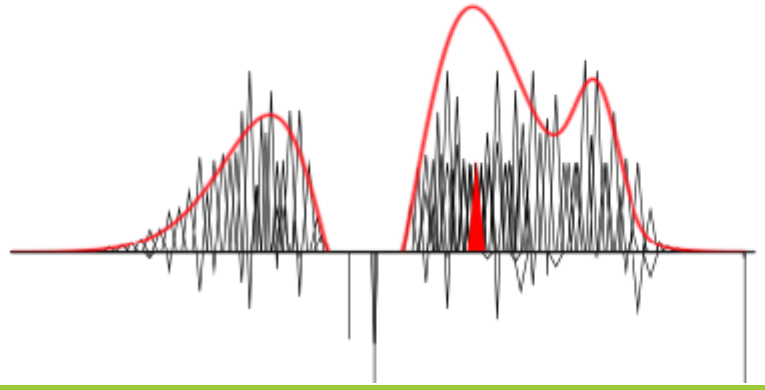

```
foreach Generation do
  //Genome decoding and evaluation
  foreach Individual do
    Identify coding sequences
    foreach CodingSequence do
      Translation into abstract protein
    end
    Compute phenotype f_P by combining protein contributions
    Compute gap with target g by comparing the phenotype f_P to the environmental target f_E
  end

  //Selection
  Compute the probabilities of reproduction based on gap with target g

  //Reproduction with mutations and rearrangements
  foreach Individual do
    foreach Offspring do
      Create Offspring
      Do rearrangements
      Do local mutations
    end
  end

  Replace current population
end
```

Aevol (a case study on artificial life)

- Fitness to a target curve determines probability of creating offsprings.
- Fitness/selection do not determine mutations! Mutations happen randomly, and then fitness is computed.

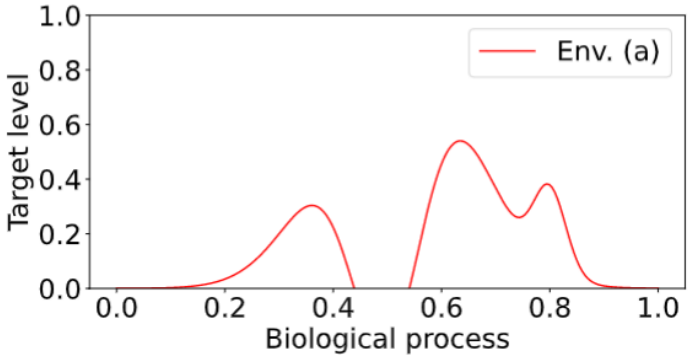



scale : 719 bp



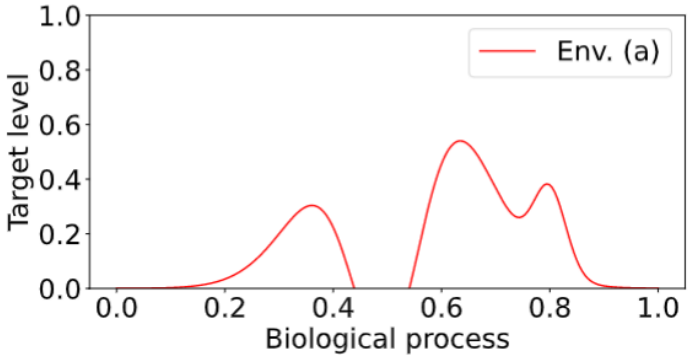
Simulation process

- Initial genomes, called wild-types, are pre-evolved for 1M generations to adapt to a given fitness curve.
- The sum of gene triangles closely matches the curve.
- We then change the environment and track duplication events.

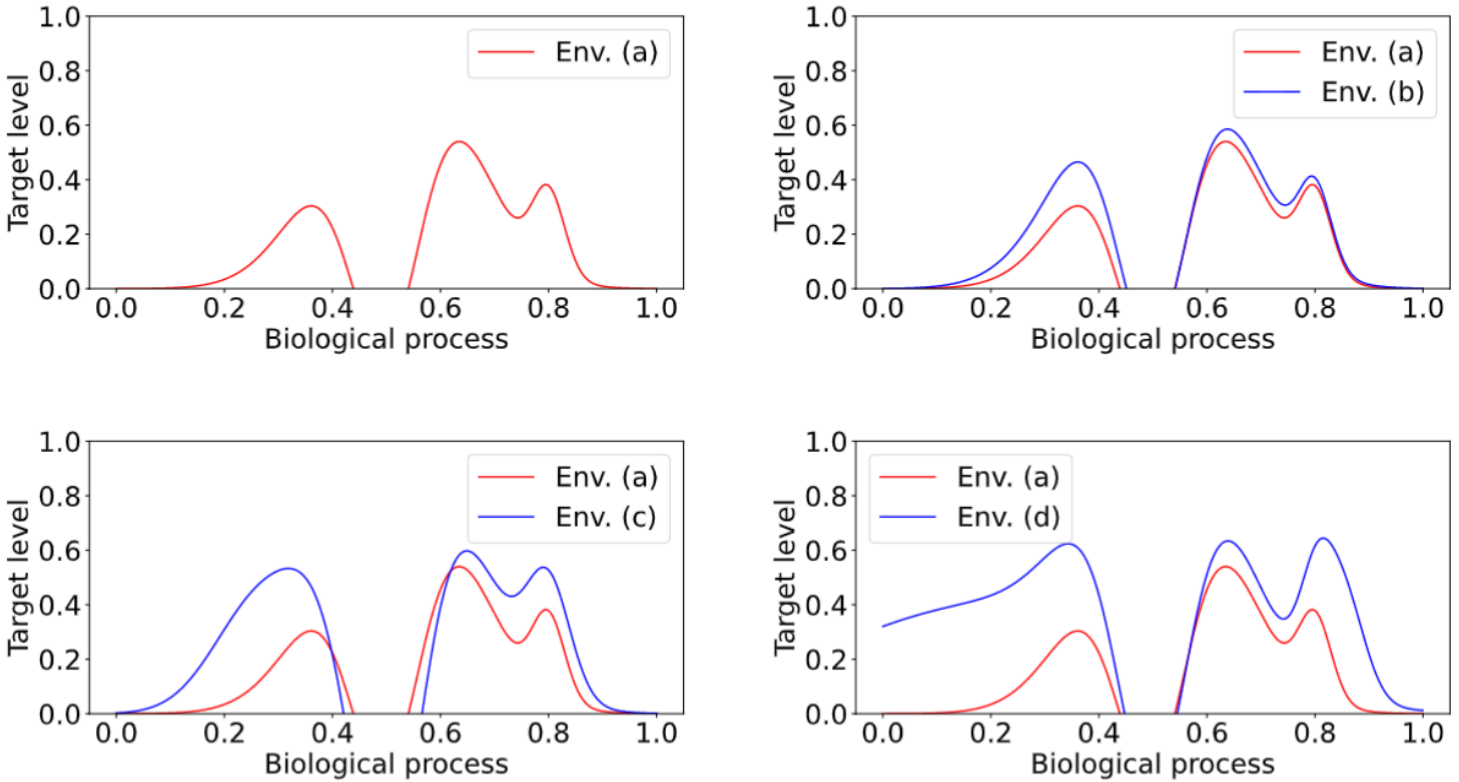


Simulation process

- Initial genomes, called wild-types, are pre-evolved for 1M generations to adapt to a given fitness curve.
- The sum of gene triangles closely matches the curve.
- We then change the environment and track duplication events.
- Parameter w_{max} = max **pleiotropy** of genes.
- Large w_{max} = genes have wider triangles => more functions

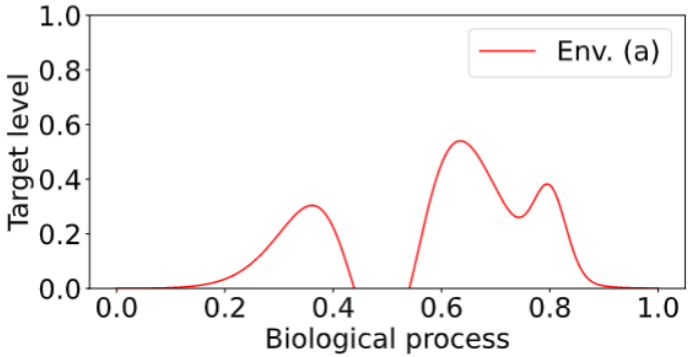


Four environment used in simulations



Changes : $Env.(d) > Env.(c) > Env.(b) > Env.(a)$

- NOTE: we only detect OBSERVABLE paralogs. We do NOT detect entirely lost genes.
  - Pseudogeneization = one of the genes is translated, but performs 0 functions.



Results

- Rate of gene duplications for each environment (dups per 1M generations)

	Env. (a)	Env. (b)	Env. (c)	Env. (d)
Gene dup. rate	3.559	15.825	40.712	55.326

- Rate of gene duplications for each pleiotropy level (dups per 1M generations)

	$w_{max} = 0.01$	$w_{max} = 0.1$	$w_{max} = 0.5$	$w_{max} = 1$
Gene dup. rate	53.735	28.220	17.641	15.825

Results

1) Less adapted = more dups
2) Higher pleiotropy = each gene does more = less dups

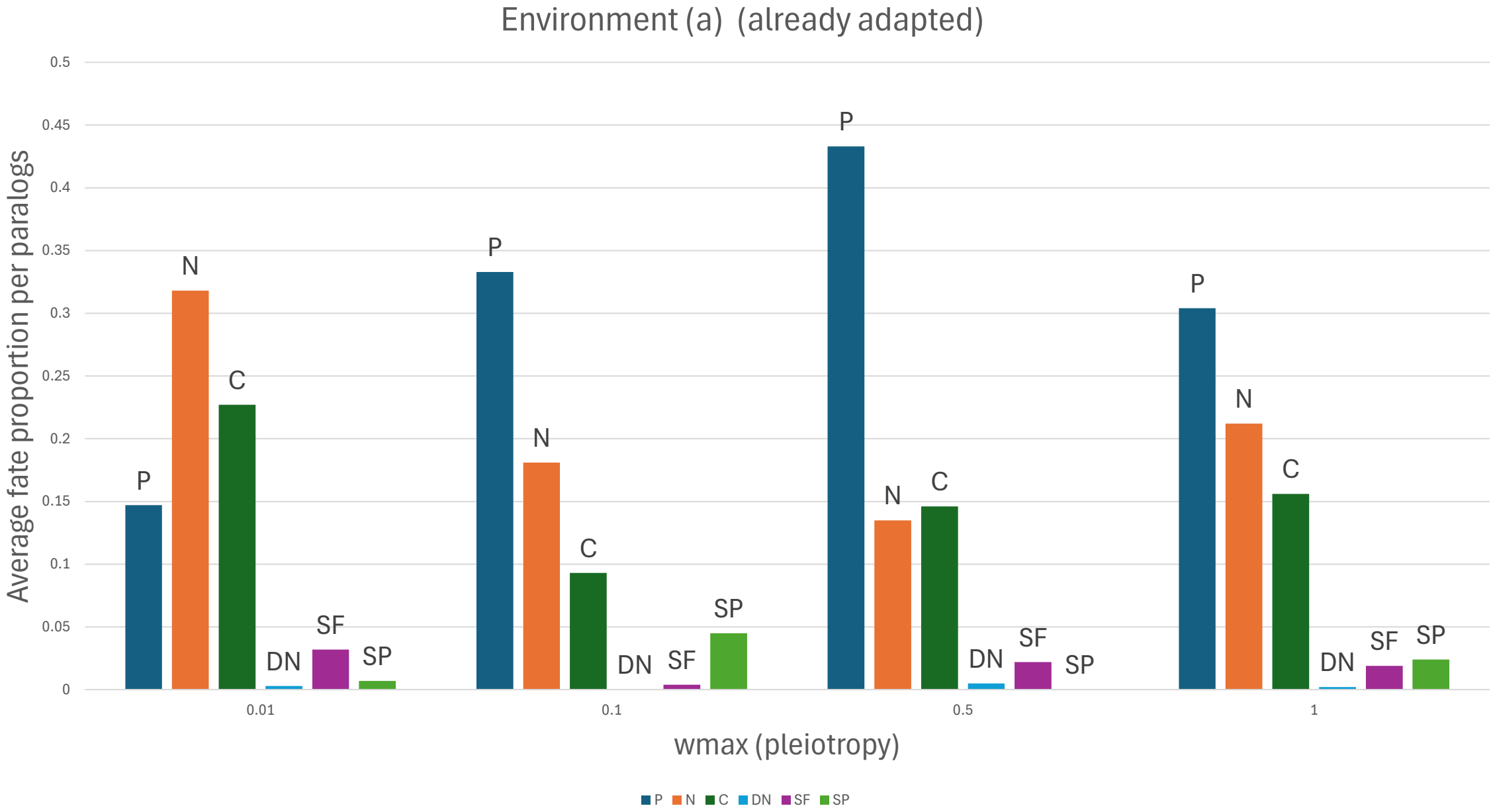
- Rate of gene duplications for each environment (dups per 1M generations)

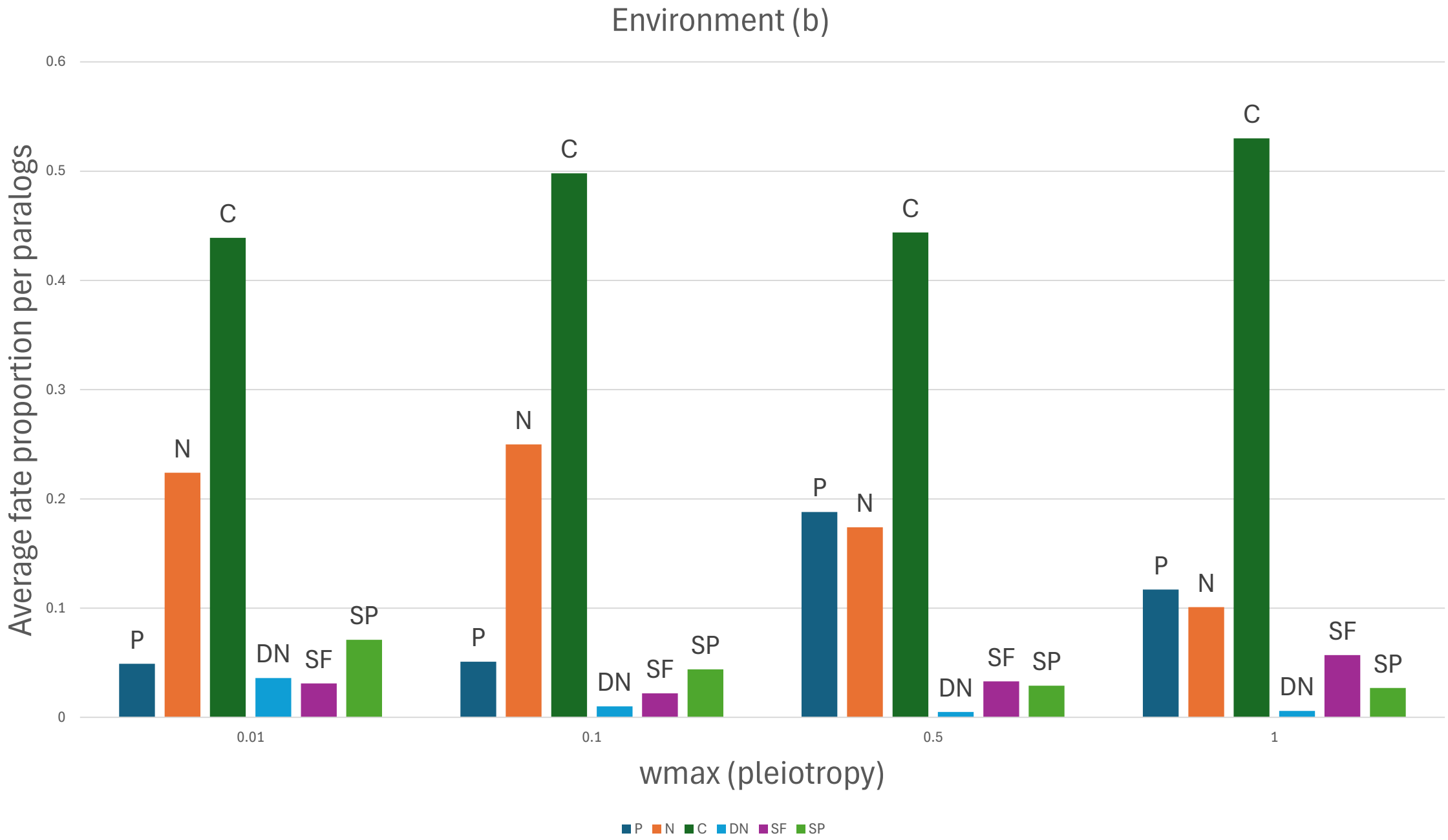
	Env. (a)	Env. (b)	Env. (c)	Env. (d)
Gene dup. rate	3.559	15.825	40.712	55.326

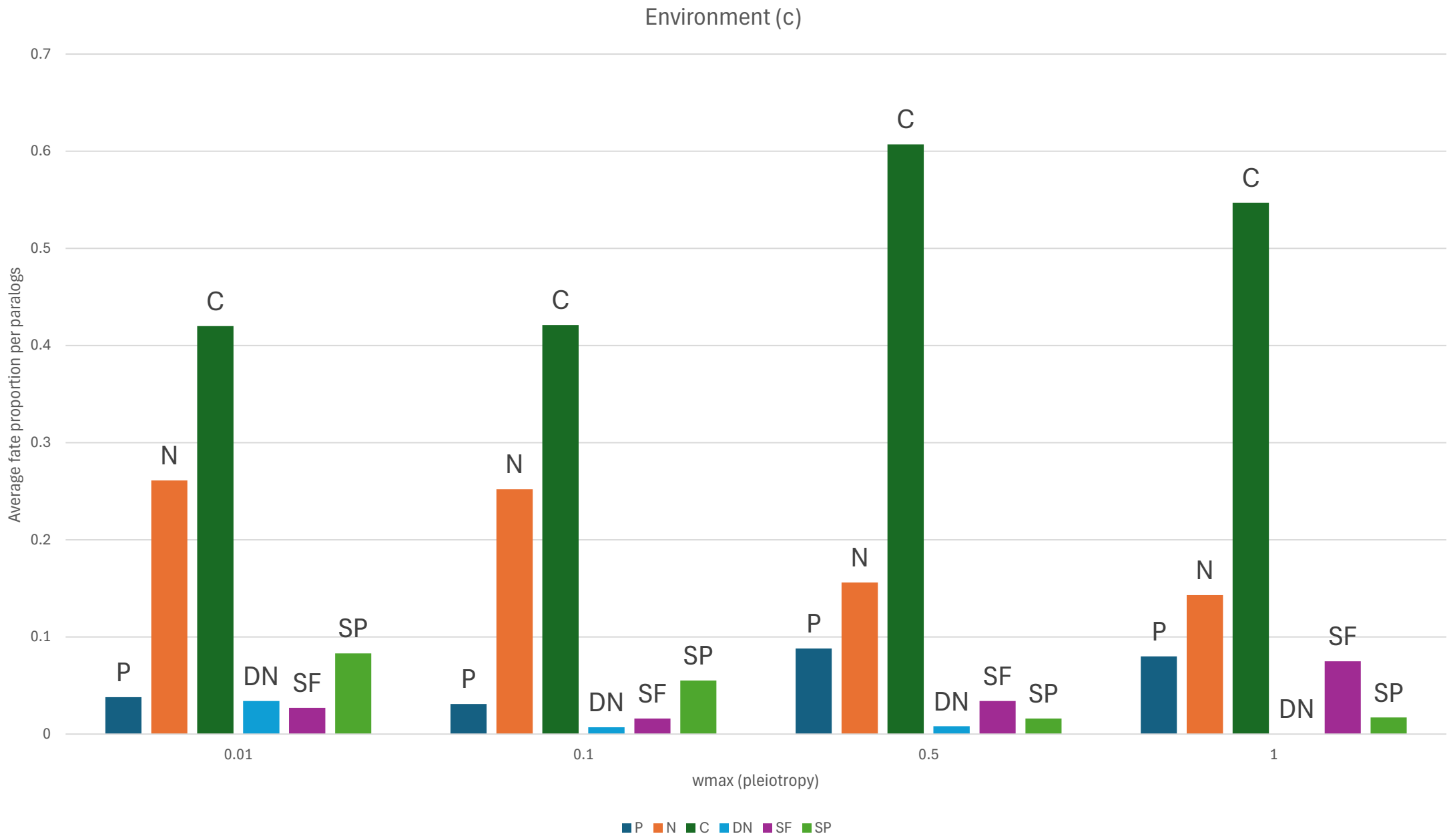
- Rate of gene duplications for each pleiotropy level (dups per 1M generations)

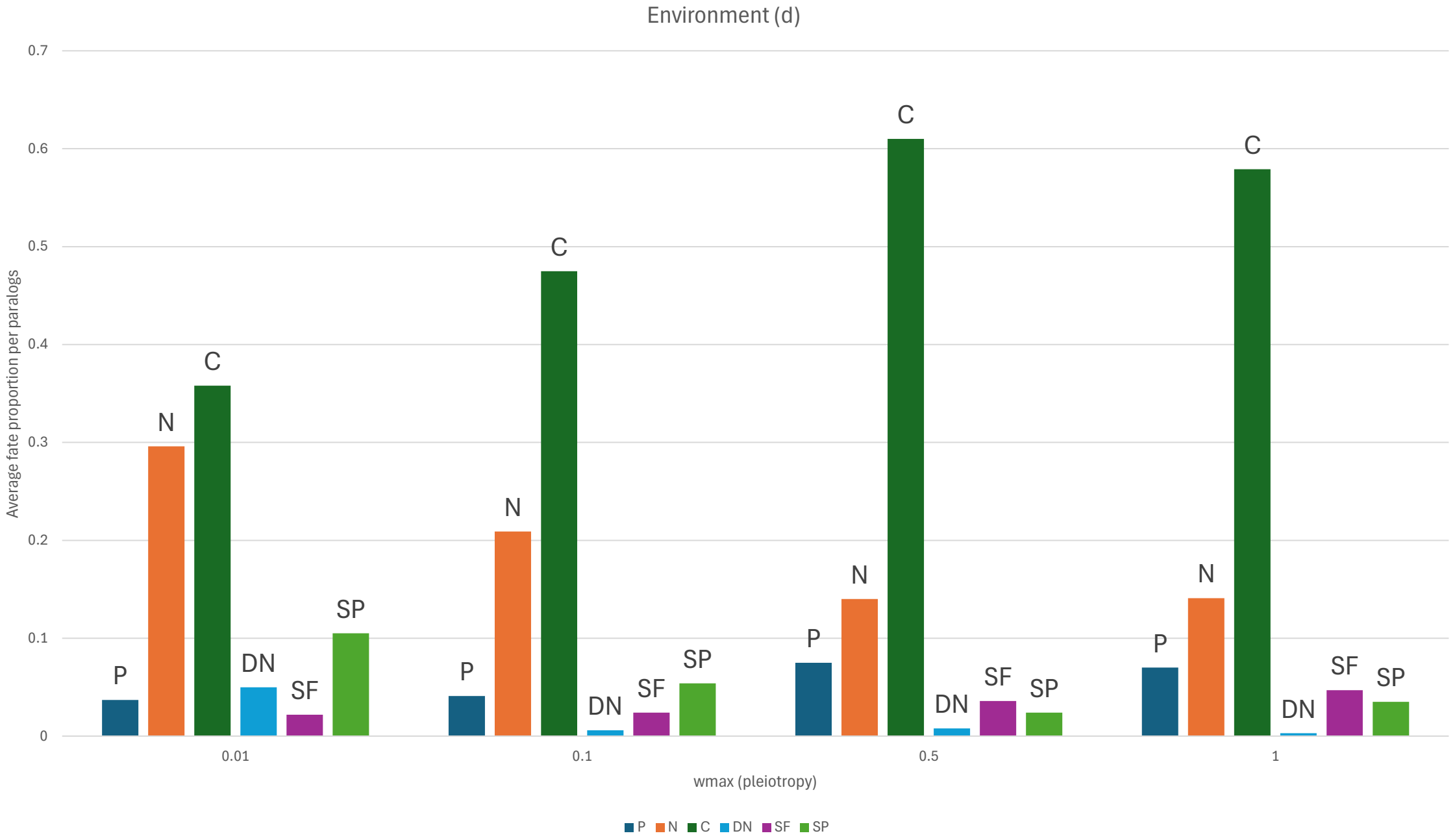
	$w_{max} = 0.01$	$w_{max} = 0.1$	$w_{max} = 0.5$	$w_{max} = 1$
Gene dup. rate	53.735	28.220	17.641	15.825

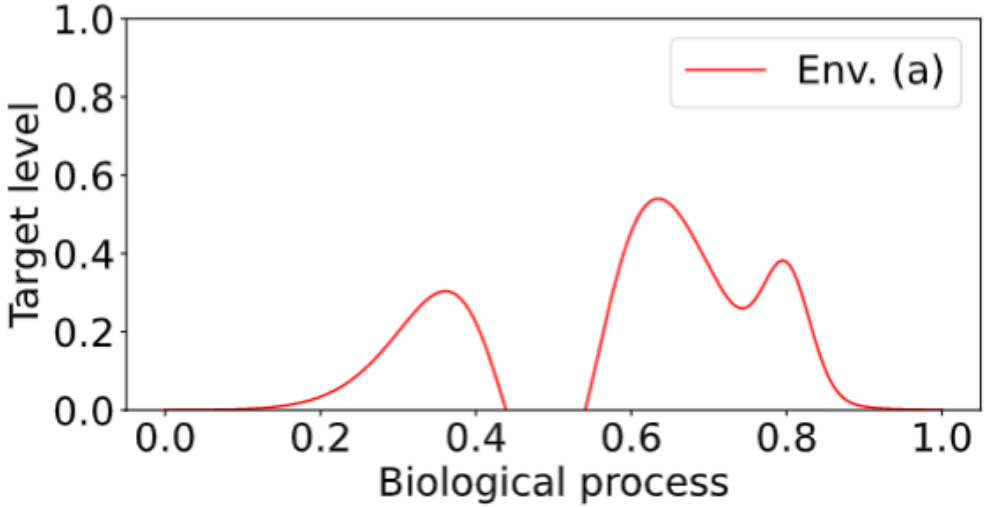
Proportion of fates for each environment

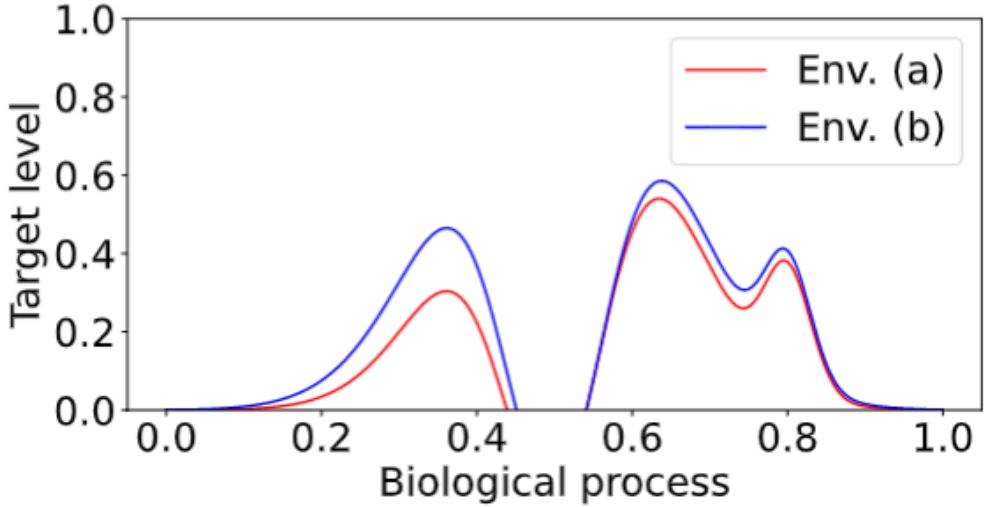

Environment (a) (already adapted)
Average fate proportion per paralogs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
P
N
C
DN
SF
SP
0.01
0.1
0.5
1
wmax (pleiotropy)
P N C DN SF SP

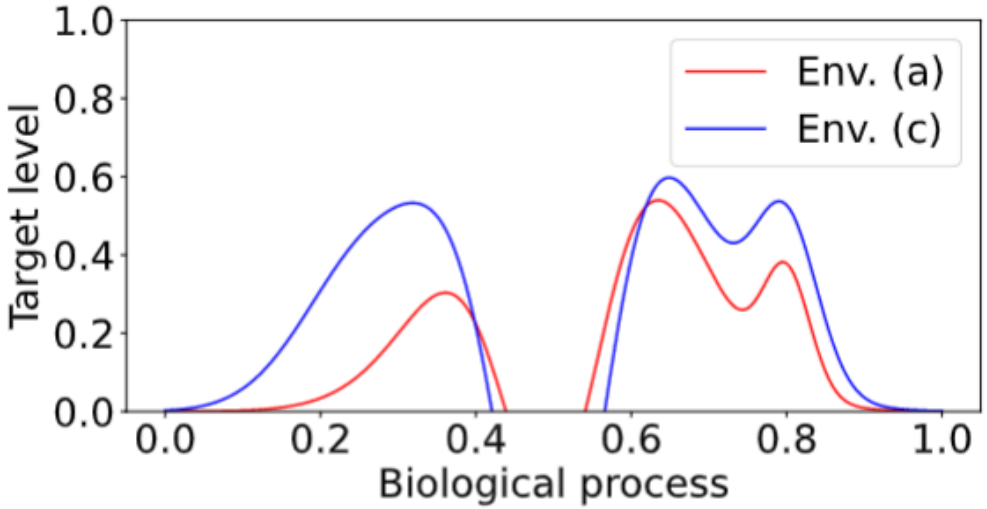

Environment (b)
Average fate proportion per paralogs
0.6
0.5
0.4
0.3
0.2
0.1
0
P
N
C
DN
SF
SP
0.01
0.1
0.5
1
wmax (pleiotropy)
P N C DN SF SP

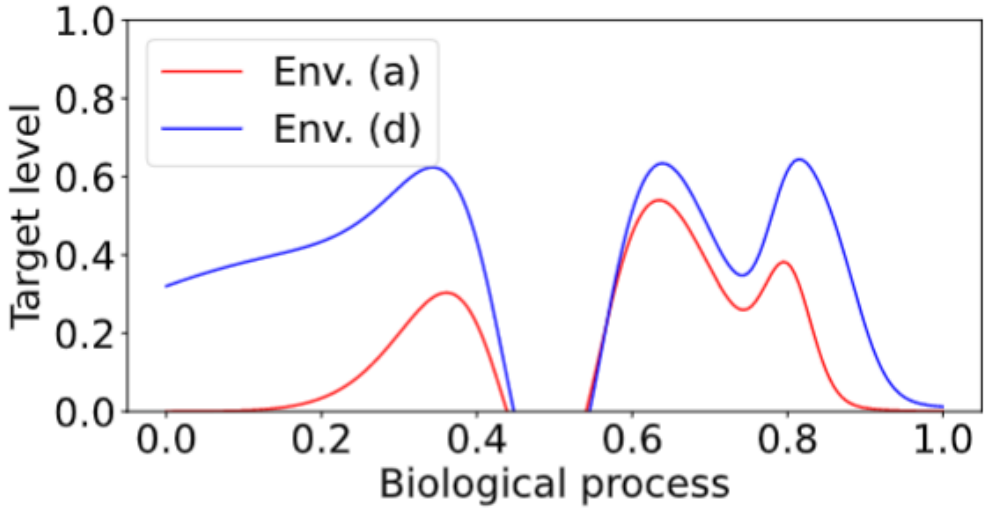

Environment (c)
Average fate proportion per paralogs
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P
N
C
DN
SF
SP
0.01
0.1
0.5
1
wmax (pleiotropy)
P N C DN SF SP

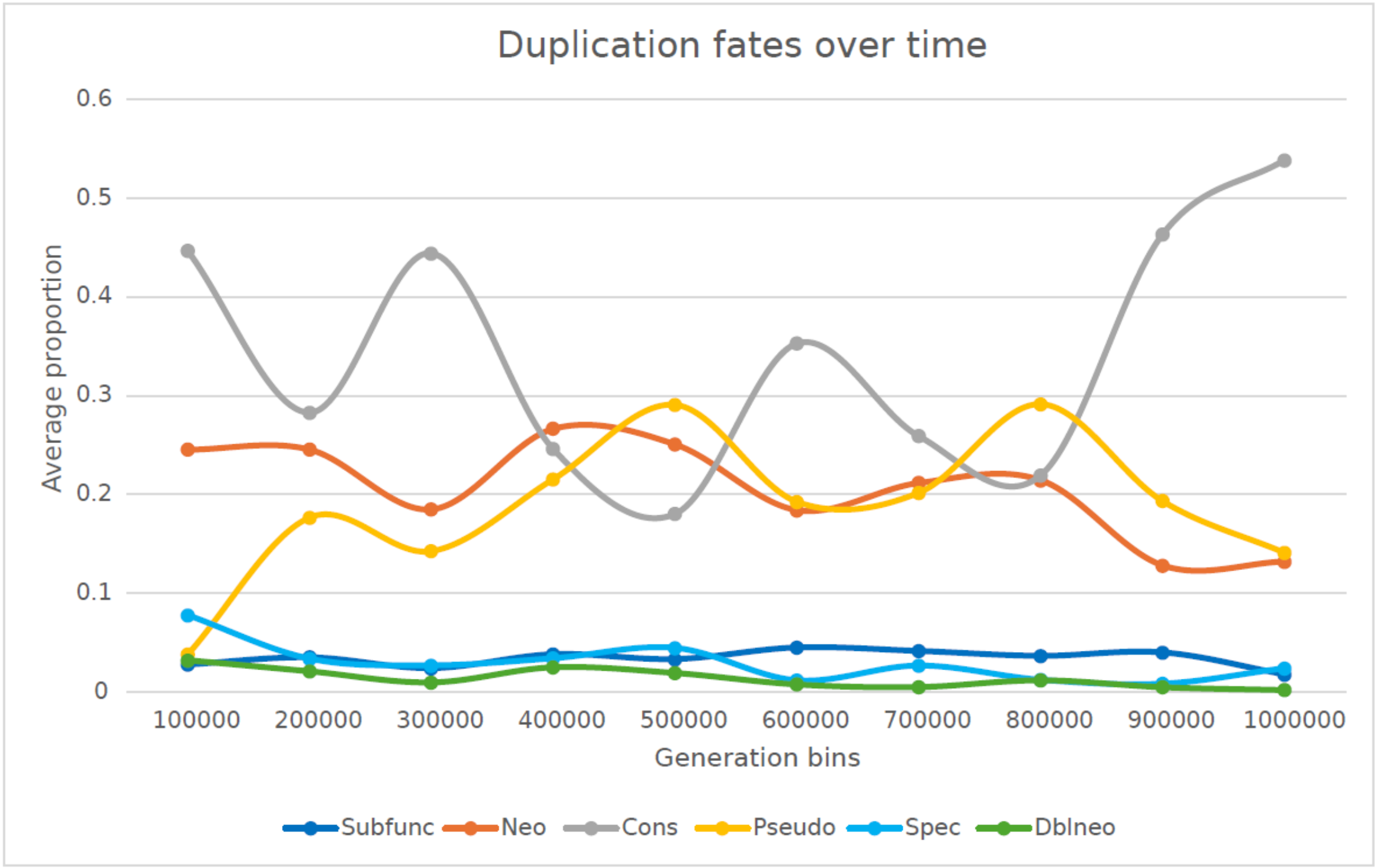

Environment (d)
Average fate proportion per paralogs
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P
N
C
DN
SF
SP
0.01
0.1
0.5
1
wmax (pleiotropy)
P N C DN SF SP


Env. (a)
Target level
Biological process
0.0
0.2
0.4
0.6
0.8
1.0


Env. (a)
Env. (b)
Target level
Biological process
0.0
0.2
0.4
0.6
0.8
1.0


Env. (a)
Env. (c)
Target level
Biological process
0.0
0.2
0.4
0.6
0.8
1.0


Env. (a)
Env. (d)
Target level
Biological process
0.0
0.2
0.4
0.6
0.8
1.0


Duplication fates over time
0.6
0.5
0.4
0.3
0.2
0.1
0
Average proportion
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
Generation bins
Subfunc
Neo
Cons
Pseudo
Spec
Dblneo

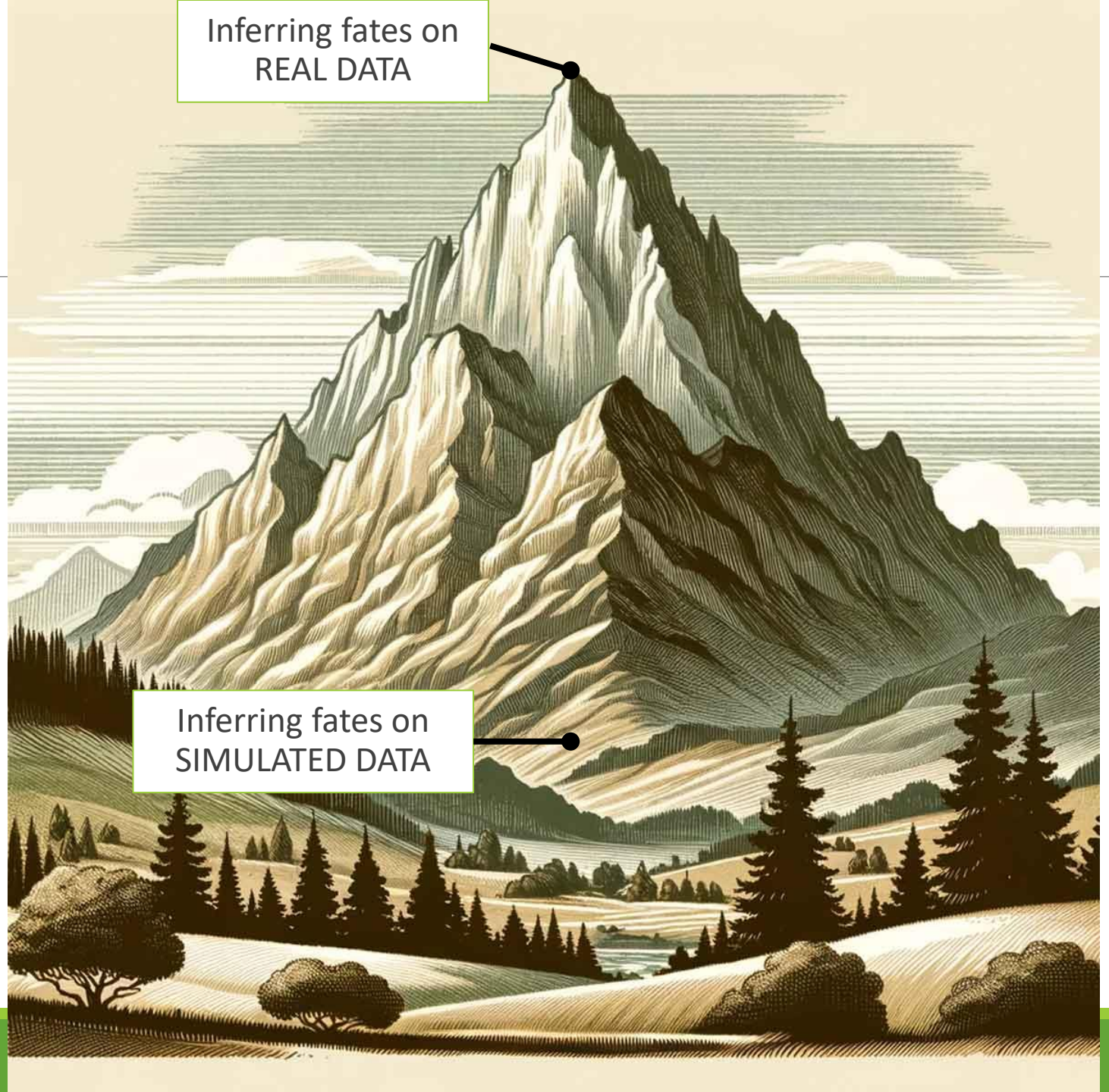
w_{max}	P	N	DN	C	SF	SP	Total	Dup. rate
Environment (a)								
0.01	0.147	**0.318**	0.003	0.227	0.032	0.007	0.733	4.703
0.1	**0.333**	0.181	0.000	0.093	0.004	0.045	0.658	1.875
0.5	**0.433**	0.135	0.005	0.146	0.022	0.0.19	0.760	4.143
1	**0.304**	0.212	0.002	0.156	0.019	0.024	0.717	3.844
Environment (b)								
0.01	0.049	0.224	0.036	**0.439**	0.031	0.071	0.849	48.403
0.1	0.051	0.250	0.010	**0.498**	0.022	0.044	0.875	17.800
0.5	0.188	0.174	0.005	**0.444**	0.033	0.029	0.872	11.563
1	0.117	0.101	0.006	**0.530**	0.057	0.027	0.837	13.025
Environment (c)								
0.01	0.038	0.261	0.034	**0.420**	0.027	0.083	0.862	113.428
0.1	0.031	0.252	0.007	**0.421**	0.016	0.055	0.782	35.700
0.5	0.088	0.156	0.008	**0.607**	0.034	0.016	0.909	26.33
1	0.080	0.143	0.000	**0.547**	0.075	0.017	0.863	24.900
Environment (d)								
0.01	0.037	0.296	0.050	**0.358**	0.022	0.105	0.868	159.668
0.1	0.041	0.209	0.006	**0.475**	0.024	0.054	0.809	60.000
0.5	0.075	0.140	0.008	**0.610**	0.036	0.024	0.892	33.300
1	0.070	0.141	0.003	**0.579**	0.047	0.035	0.876	29.300

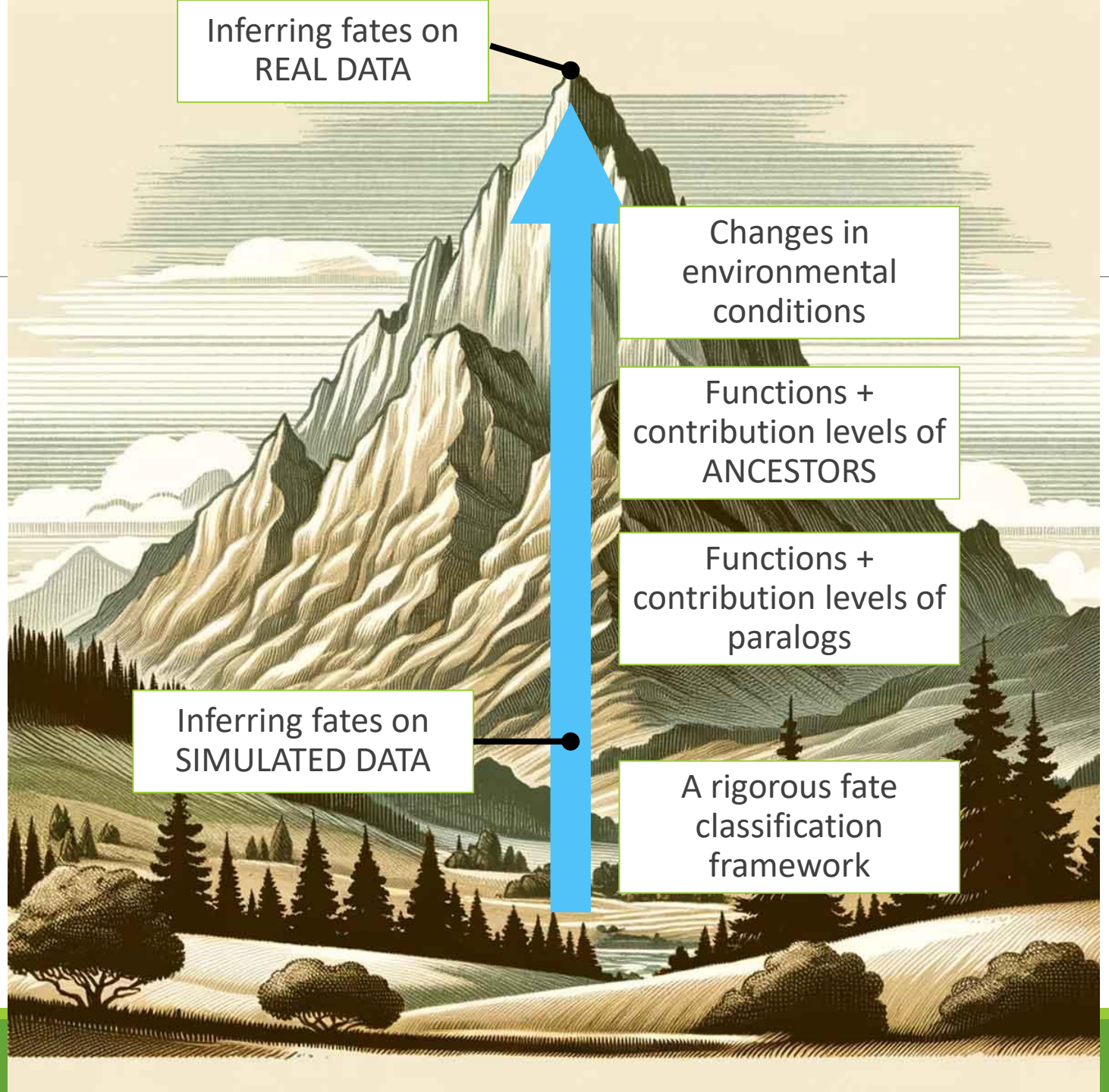
Some observations

- 10-20% of paralogs **did not fit any of the canonical fates** (sum-of-fates ≤ 0.7).
- 10-15% of **fates are hybrid** (two fates with ≥ 0.2 proportion)

- Conservation increases with w_{max}, Neofunc decreases with w_{max}.
- Subfunc is rare, which agrees with theory, and increases with w_{max}.
- In our simulations, the most frequent fate is Conservation, followed by Neofunc and Pseudo.
  - Probably because our environment asks for increased expression.

Bringing the framework to REAL DATA


Inferring fates on
REAL DATA
Inferring fates on
SIMULATED DATA


Inferring fates on REAL DATA
Changes in environmental conditions
Functions + contribution levels of ANCESTORS
Functions + contribution levels of paralogs
Inferring fates on SIMULATED DATA
A rigorous fate classification framework


shutterstock
sm.art
shutterstock
sm.art
shutterstock
shutterstock
shutterstock
IMAGE ID: 1460751842
www.shutterstock.com

Bringing the framework to REAL DATA

- **Question**: how does the environment affect the fate of duplications?
- Must get:
  - contribution levels of all paralogs to functions
  - contribution levels of all ancestors to functions
  - environmental conditions and required adaptation

Bringing the framework to REAL DATA

- **Question**: how does the environment affect the fate of duplications?
- Must get:
  - contribution levels of all paralogs to functions
  - contribution levels of all ancestors to functions
  - environmental conditions and required adaptation
- Anyone interested in these questions is more than welcome to work with us!!

Thank you!