Description of the agent-based model of evolution of copy number variation of MHC genes under sexual selection with varying mating rules

by **Piotr Bentkowski** (implementation, coding) & **Jacek Radwan** (conception, guidance)

The source code is available at www.github.com/pbentkowski/MHC_Evolution

---

This version of the model is an extension of the agent-based simulation developed by Piotr Bentkowski and Jacek Radwan for studying the impact of pathogen diversity on individual copy number variation of the major histocompatibility (MHC) complex genes [1]. It is a modification of the framework first proposed by Borghans et al. in 2004 [2] and developed further in our laboratory [3–6].

The main mechanism is based on the observation that only a short fragment of MHC protein, called **protein-binding region (PBR)**, binds the antigen. The antigen usually is 8-10 peptides long [7](pp. 47-52) and that PBR is affected with higher mutation rates than other regions of the MHC chains [8]. The binding mechanism in our model is a simplified version of the general mechanism seen in nature. Both, pathogens' antigens and hosts' MHC PBRs are represented as strings of zeros and ones (bit strings) of fixed length, where the antigen is two orders of magnitude longer than MHC's PBR. Binding is simulated as a match in all position of the bit strings representing the MHC PBR and an epitope (Fig. 1). For the sake of simplicity, biding affinities are also binary (only 'binding' or 'no binding'). Fitness in pathogen populations is evaluated in proportion to the number of hosts a pathogen infects and hosts fitness is proportional to the number of pathogens presented. To represent pathogen shorter generation times, we allow *K* generations of pathogens for one generation of hosts. This scheme constitutes a co-evolution system with the Red Queen dynamics [6].

Pathogens: Pathogen has a single antigen of length *a* bits representing all possible epitopes after degradation of pathogen's proteins by the host. The epitope length is equal to the length of *m*. The size of the pathogen population is fixed and divided into so-called species. Species are groups of pathogen shearing evolutionary origin and history. The number of individuals belonging to one pathogen species is fixed.

Hosts: Host's MHC protein-binding region (later referred simply as MHC) defines a gene and is represented by a bit string of length *m* (same as epitope) that is significantly shorter than the antigen's length *a*. The MHC genes are located on two chromosomes, and the number of genes in one individual is free to evolve under different selection scenarios. There are two numbers describing MHC genes on the chromosomes: the number of MHC copies ('loci') and the number of unique MHCs with distinct bit strings which we call **MHC variants**. In principle, a host can have fewer variants than MHC gene copies. The size of the host population is fixed. During reproduction two individuals, chosen by fitness proportionate selection method, are picked to mate contributing to offspring one chromosome each. The transferred chromosome is selected at random.

Mechanism of pathogen presentation and host infection: During the encounter with a pathogenic individual, a host tries to present the pathogen sliding each of its MHCs along the pathogen's antigen. It does so until it will find an identical bit sub-string (epitope) as the MHC bit string (triggering pathogen presentation) or until it will reach the end of the antigen string (Fig. 1). If all of the host's MHC will reach the antigen's end without finding a match, the host gets infected by this pathogen individual. To represent pathogen shorter generation times, we allow *K* generations of pathogens for one generation of hosts.

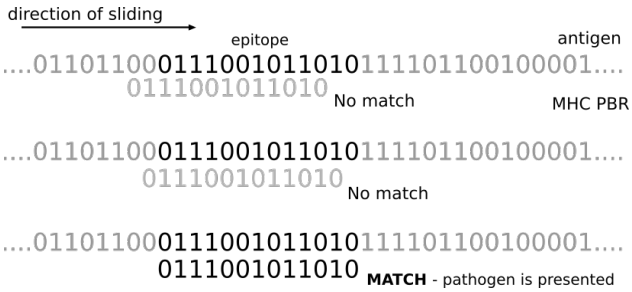

Fig 1: *A schema of pathogen presentation by a single MHC PBR (protein-binding region). A bit string representing MHC PBR slides along the bit string representing antigen until it will encounter an identical sub-string (epitope) in the antigen, which leads to immune response. If all of the host's MHCs will reach the end of the antigen without finding a matching sub-string, the host gets infected.*

Mutations: During reproduction and selection, both, hosts and pathogens may undergo some mutations. Mutations are represented by a flip of a single bit with a given probability (different for pathogens and hosts). In some previous theoretical studies [e.g. 2, 3, 4] mutation of MHC were given as the probability of change in MHC as a whole (replacement of old MHC with a new one). To be consistent with these studies,

we use the mutation rate per peptide, but in the model, to be more realistic, we calculate mutation rate per single bit (site) according to an equation:

$$\mu_{bit} = 1 - \left(1 - \mu_{MHC}\right)^{1/m} \quad (1)$$

where μ_{MHC} is the mutation probability per MHC peptide, μ_{bit} is the mutation probability per single bit in MHC and m is the number of bits in MHC. See also S1 Appendix in [6]. Another kind of mutations, affecting only the MHCs, are deletions and duplications. Duplications occur with probability μ_{dupl} by adding a new copy of an existing variant to the same chromosome. A gene can be deleted from the chromosome with probability μ_{del}. A chromosome is constrained to have at least one MHC gene.

Pathogen selection for reproduction: Pathogens are selected in proportion to the number of hosts they managed to infect. Selection works within each species separately by using fitness proportionate selection method (a.k.a. 'roulette wheel' selection).

Hosts selection: Host selection is conducted in two phases: (i) selection by pathogens and (ii) sexual selection phase.

<u>*Phase I: selection by pathogens:*</u> This phase is identical to used in our earlier work [1]. It uses fitness proportionate selection method, where the fitness is given by the question:

$$f_{host} = P \cdot e^{-(\alpha N)^2} \quad (2)$$

where P is the number of pathogen species a host presented (avoiding infection), N is the number of unique MHC variants (rather than loci, which may share identical variants) present in both host chromosomes, and α is the cost factor. The multiplier of P is a penalty function introduced to prevent the number of MHC from exploding under Red Queen dynamics [1]. This penalty reflects trade-offs associate with having too many MHC variants, .e.g. increased risk of autoimmunity or, as a consequence of a mechanism preventing autoimmunity, reduced repertoire of the T-cell receptors [9, 10]. Individuals are selected to the next phase with probability proportional to their f_{host} (eq. 2) until the population of N_H individuals is recreated. Each individual can be selected multiple times, increasing its probability of mating.

<u>*Phase II: sexual selection:*</u> The second phase of selection is conducted according to one of several mating rules we tested. For the sake of simplicity, all individuals could mate with any other (i.e. our hosts are equivalent to out-crossing hermaphrodites). The program selects one random individual (the chooser) and offers it p random mates from which it selects one (the partner) according to one of given mating rules (described below). The pair produces one offspring and the procedure in phase two repeats until recreating the host population of N_H individuals.

Mating rules: We tested a number of mating rules regarding partners MHC repertoire:

- *Radom* – no preference is given, a random partner is assigned.
- *MinShared* – selecting a partner that has the smallest number of the same MHC variants as the chooser.
- *PropShared* – selecting a partner that has a minimal proportion of shared MHC variants, i.e. the ratio of the number of shared MHC variants to the sum of variants in both partners.
- *MaxDiffer* – selecting a partner that has the maximal number of different MHC variants than the chooser.

To see if mating rules that could potentially suppress MHC expansion (*MinShared* and *PropShared*) can actually do so without extrinsic selection, we run two additional scenarios without the penalty function:

- *MinSharedUnc* – same as *MinShared*, only without constrains of eq. 2 (fitness is simply proportional to the number of presented pathogens *P*).
- *PropSharedUnc* – same as *PropShared*, only without constrains of eq. 2 (fitness is simply proportional to the number of presented pathogens *P)*.

After sexual selection, mutations are performed.

Parameter selection: The model does not aim at explaining observed MHC gene frequencies in real populations, but at capturing the general mechanisms of MHC diversity evolution under joint selection from pathogens and sexual selection. Thus, the parameters were chosen to first reproduce the Red-queen dynamics of host-parasite co-evolution [1, 6]. Parameters and their values are given in Tab. 1. The most crucial to the realism of the model is the selection of the bit strings length (antigen and MHC PBR). 6000 bits in an antigen means there are $\sim 1.5 \cdot 10^{1806}$ all possible antigens which is a vast number, for our purpose as good as infinity representing all potential peptides in pathogens. For comparison, a study that used full-length dengue virus polyprotein sequences retrieved from the National Center for Biotechnology Information Protein database generated 38 845 unique 9-mer peptide sequences for binding predictions [11]. 16-bit-long MHC PBR means there is 65 536 possible MHC PBR, and for comparison, there are over 10 000 human MHC variants of both class I and II known today [12]. A chance that a random 16-bit-long MHC will find a match in a random 6000-bit-long antigen is around 0.084, a value similar to the one used in previous modelling studies [2] and based on experimental estimates [13]. Bioinformatics analysis

showed a large variation of recognition potential in human MHCs (HLA, human leukocyte antigen). From an extensive collection of over 30 000 dengue virus-derived peptides, only 0.3% were predicted to bind HLA A*0101, whereas nearly 5% were predicted for A*0201 [11].

Tab. 1: *Parameters of the model. [1] number of individuals of one pathogen species, the total number of pathogens is $N_P \cdot S$; [2] probability per reproduction event; [3] probability per site (flip of a single bit); [4] probability of change of the whole MHC (change in any given site: see the paragraph 'Mutations').*

Parameter description	Symbol	Values
Hosts population size (number of individuals)	N_H	1000
Pathogen population size (number of individuals) [1]	N_P	1000
Number of pathogen species in the simulation	S	4, 8, 16
Antigen length (number of bits)	a	6000
MHC's protein-binding region length (number of bits)	m	16
Number of pathogens generation per one hosts generation	K	10
Total number of hosts generation (a.k.a. simulation time)	G	3000, 5000
Probability of mutation in antigen (per site) [2,1]	μ_A	10^{-5}, $5 \cdot 10^{-5}$
Probability of mutation in MHC PBR (per protein) [2,4]	μ_{MHC}	10^{-4}
Probability of deletion of MHC gene [2]	μ_{del}	10^{-3}
Probability of duplication of MHC gene [2]	μ_{dupl}	10^{-3}
Cost factor for the host selection function (eq. 2)	α	0.02
Fraction of antigen bits which get fixed and cannot mutate	Φ	0
Number of potential mating partners offered to the choosing host during sexual selection	p	10

Cost factor α for the host selection function was selected that so the number of unique MHC types in an individual host will not exceed numbers seen in known MHC-rich genomes [1].

Implementation: The model's program was written in C++14 language. It generates text files containing simulation results that later are analysed and plotted using Python scripts. General scheme of the program is shown in Fig. 2. The source code, Python scripts for analysis and documentation can be obtained from www.github.com/pbentkowski/MHC_Evolution.

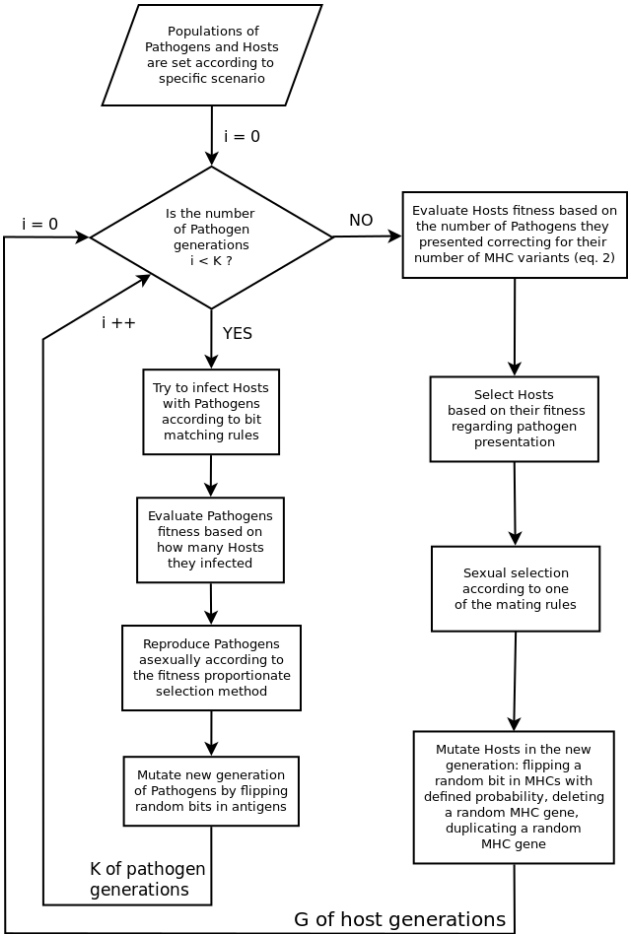

Fig 2: *General schema of the simulations. The program consists of two loops: more significant loop iterates trough number of host generations, embedded smaller loop corresponds to faster pathogen generation times.*

References:

1. Bentkowski P, Radwan J. Evolution of major histocompatibility complex gene copy number. *PLOS Comput Biol* 2019; **15**: e1007015.
2. Borghans JAM, Beltman JB, De Boer RJ. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* 2004; **55**: 732–9.
3. Ejsmond MJ, Radwan J. MHC diversity in bottlenecked populations: a simulation model. *Conserv Genet* 2009; **12**: 129–137.
4. Ejsmond M, Babik W, Radwan J. MHC allele frequency distributions under parasite-driven selection: A simulation model. *BMC Evol Biol* 2010; **10**: 332.
5. Ejsmond MJ, Radwan J, Wilson AB, Wilson AB. Sexual selection and the evolutionary dynamics of the

major histocompatibility complex. *Proc R Soc B* 2014; **281**: 20141662.

6. Ejsmond MJ, Radwan J. Red Queen processes drive positive selection on major histocompatibility complex (MHC) genes. *PLoS Comput Biol* 2015; **11**: e1004627.
7. Frank SA. Immunology and Evolution of Infectious Disease. 2002. Princeton University Press, Princeton, New Jersey.
8. Garrigan D, Hedrick PW. Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. *Evolution (N Y)* 2003; **57**: 1707–1722.
9. Nowak MA, Tarczy-Hornoch K, Austyn JM. The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci U S A* 1992; **89**: 10896–10899.
10. Migalska M, Sebastian A, Radwan J. Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. *Proc Natl Acad Sci* 2019; 201807864.
11. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide binding repertoires of different size, affinity and immunogenicity. *J Immunol* 2013; **191**: 5831–5839.
12. Marsh SGE. Nomenclature for factors of the HLA system, update February 2012. *Tissue Antigens* 2012; **80**: 72–77.
13. Kast WM, Brandt RMP, Sidney J, Drijfhout J, Kubo RT, Grey HM, et al. Role of HLA-A motifs in identification of potential CTL epitopes in human Papillomavirus type 16 E6 and E7 proteins. *J Immunol* 1994; **152**: 3905–3912.