Deciphering the language of fungal pathogen recognition receptors

Witold Dyrka, Pascal Durrens, Mathieu Paoletti, Sven J Saupe, David J Sherman

Deciphering the language of fungal pathogen recognition receptors

Witold Dyrka[1], Pascal Durrens[1], Mathieu Paoletti[2], Sven J. Saupe[2], and David J. Sherman[1]

[1] INRIA - Université Bordeaux - CNRS, Team MAGNOME, Talence, France
{witold.dyrka, david.sherman}@inria.fr, durrens@labri.fr
[2] Institut de Biochimie et Génétique Cellulaires, CNRS, Bordeaux, France
{mathieu.paoletti, sven.saupe}@ibgc.cnrs.fr

Abstract. The NLR family of receptors plays a key role in the innate immune system of animals, plants and fungi. In the latter two phyla NLRs adapt quickly to ever-changing pathogen-specific invasion markers thanks to their repeat-based architecture, which can produce diversity of recognition epitopes through unequal crossing-over and mutation. Characterizing computationally the language of these pathogen recognition receptors can provide insight into the molecular mechanisms of immune response and describe the limits of the pathogen targets that can be recognized. In this work, we model generation and selection of the recognition epitopes as a stochastic string rewriting system with constraints, tuned by analysis of observed evolutionary processes and validated with regard to a large dataset of fungal NLRs. Among others, analyzing the feasible set of solutions revealed that the model explained the $i/i+2$ periodicity observed in the repeat number distribution of a family of receptors. In addition, in exploring discrepancies between real and simulated data we discovered an overrepresented pattern which potentially has functional importance. The methodology developed in this work is general and therefore can be applied to any class of amino acid repeats generated by unequal crossing-over for which an equivalent high quality dataset is available.

Keywords: innate immune system, amino acid repeats, unequal crossing-over, string rewriting system, formal languages

1 Introduction

The immune system searches for pathogen invasion markers, which include pathogen proteins and host proteins modified in the course of the invasion. While some pathogen-associated molecular patterns, such as bacterial flagellins [37], are relatively invariant, numerous pathogen-specific markers change quickly [20]. Therefore, to win the arms race with pathogens, host recognition receptors must adapt quickly to varied and modulating markers. To achieve this goal, the recognition domain of the receptor requires the capacity to recognize diverse possible pathogen molecule epitopes, and ability to quickly learn new epitopes. In plants, which lack an adaptive immune system, this key role in the immunity is played by so-called NBS-LRR and NOD-like receptors termed collectively NLRs [13,5,15], which act as molecular switches, transforming from a closed inactive conformation to an open active conformation upon binding of the proper ligand. In fungi, proteins homologous to plant NLRs have been long known for playing a central role in heterokaryon incompatibility, a form of conspecific non-self recognition [26,12]. It has been only recently postulated that they are also involved in the innate immune response [20]. NLRs consist of three domains: the N-terminal effector domain, the C-terminal recognition domain and the central nucleotide binding domain (NBD). The NLR repertoire in fungi appears greatly expanded and diversified with up to 273 NLR candidates in certain species [6]. Characterizing computationally the language of these pathogen recognition receptors not only provides insight into the molecular mechanisms of immune response, it also describes the limits of the pathogen targets that can be recognized. The extensive genomic coverage in the fungal phylum allows us to build large data sets of NLR-like proteins in that branch of the eukaryotic tree, making it possible to derive generative models for NLR evolution.

In order to practically derive models of the language of pathogen recognition receptors, it is necessary to understand various evolutionary processes in detail. In the vast majority of cases, the NLR recognition domain is built from repeats forming a solenoid-like structure [5]. Amino

acid repeats have gained substantial interest of researchers because of their abundance, functional and evolutionary significance [14], and also because of the difficulties they impose on sequence the alignment [28]. Moreover, engineered repeat proteins have been developed to act as diagnostic and therapeutic tools due to their specific binding affinity [30,18]. Nevertheless, describing the evolution and function of amino acid repeats remains mostly elusive [14]. In the context of the immune system, a benefit of the repeat-based architecture is the ease of producing diversity by shuffling of repeats, a process which is 10,000 to 100,000 times quicker than the standard point mutation [24]. According to the recent study [19], the most essential mechanism of the repeat shuffling is unequal crossing-over [31]. This recombination process both requires and promotes high sequence similarity between consecutive repeat units. Simultaneously, recombination locally induces point mutations at the rate 100-1000 times more quickly than in reference regions [24]. This is complemented by the process of repeat-induced point mutation (RIP) aimed at mutating repeat fragments to neutralize transposons, whose rate differs between species [29]. The requirement of high homology between repeat units defines a subclass of amino acid repeats which we devise as Highly internally Conserved (HiC). While they are only a small fraction of eukaryotic repeat regions (around 1%), majority of them - at least in fungi - can be found in the NLRs [6,27]. The recognition domain of fungal NLRs is usually built from repeats belonging to three Pfam clans [23]: Ankyrin (Ank), Tetratrico Peptide Repeat (TPR) and Beta propeller including WD40 instead of the Leucine-Rich Repeats (LRR) typically found in vascular plant NLRs. Repeat units of the three types are about 30-40 amino acids long. It has been recently shown that certain positions in HiC repeats in NLR are under positive diversifying selection [21,6]. Interestingly, these positions are placed at the surface of the repeat-based structure. Therefore, it is likely that these sites form a recognition epitope which serves as a structural complement to epitopes of pathogen invasion markers.

This subclass of HiC repeats provides a relatively well defined mechanisms of their generation, which can be used to build a computational model to simulate their evolution. However, to our knowledge, this opportunity has never been exploited. In this work we model the generation process of highly conserved repeats as a string rewriting system, specifically a stochastic string rewriting system with constraints, tuned to model observed evolutionary processes. We validate the model with regard to the fungal NLRs recognition receptors, identify possible applications of our approach, and identify paths for future improvement.

2 Materials and Methods

2.1 Materials

NLR proteins were extracted from completely annotated genomes in the "nr" database as of June 27, 2013 as described in [6]. C-terminals domains of STAND proteins were scanned using PfamA [23] for matches to three PfamA signatures, PF00023 (Ank), PF00400 (WD40) and PF13374 (TPR_10). Sequences mixing repeats from CL0020 (TPR), CL0186 (Beta propeller) and CL0465 (Ankyrin) clans were removed. Regions containing HiC repeats were detected using T-reks [9] with customized parameters (PSIM=0.85, kmeans=10, overlapfilter on, external MSA). PfamA hits from the three repeat signatures which were contained within HiC repeat regions (plus 20-amino acid envelope) were extracted and aligned, for each repeat type separately, using Muscle 3.8.31 [7]. After visual inspection of the TPR_10 alignment, 26 repeats were removed as they exhibited different pattern of alignment and the others were re-aligned. This way we obtained alignments of 866 Ank repeats from 190 sequences, 1284 WD40 repeats from 251 sequences and 965 TPR_10 repeats from 196 sequences. The alignments were used to identify highly variable sites. For analysis of patterns of the highly variable sites, these sets were further pruned from sequences where any two repeats were separated by more than 20 amino acids. This resulted in 161 Ank, 207 WD40 and 185 TPR_10 sequences. The substitution matrix for modeling single point mutations consisted of probabilities that a given amino acid would mutate to another one as the consequence of a single mutation to its codon. Transitions to the stop codons were not taken into account.

2.2 Methods

Model of evolution of repeat sequence In this paper we propose to model generation of the highly conserved repeat as a stochastic String Rewriting System with Constraints (sSRSwC), a variation of the Semi-Thue system [32,22]. A sSRSwC as a quadruple $\langle \Sigma, R, P, Q \rangle$, where Σ is an alphabet, R is a set of binary relations (rewriting rules) defining possible transitions from string (word) u to another string v, such that $u, v \in \Sigma^*$, P is a set of probabilities assigned to rewriting rules, and Q is a set of constraints. Intuitively, R and P model repeat generation, while Q models selection. In the case of repeat modeling, the alphabet consists of symbols representing repeats. The set of rules $(u \to v)$ contains all transitions in which u can be transformed into v through a single event of unequal crossing-over followed by point mutations. Unequal crossing-over $X_{p,l}$ can occur to string u at position p over length l with probability $\Pr_X(N, p, l)$, where L is length of u and $p + l \leq L$. Crossing-over can produce two kinds of output $u' = X_{p,l}(u)$ from the string $u = u_1 \dots u_N$: a *copy offspring*,

$$u^C = u_1 \dots u_p u_{p+1} \dots u_{p+l-1} u_p u_{p+1} \dots u_{p+l-1} u_{p+l} \dots u_N$$

or a *delete offspring*,

$$u^D = u_1 \dots u_{p+l-1} u_{p+l} \dots u_N.$$

Single point mutation $SNP_r(\mathtt{X}, \mathtt{Y})$ occurs with probability p_{mut} at each position r of the string u' after the unequal crossing-over. It changes residue $\mathtt{X}$ into $\mathtt{Y}$ with probability $\Pr_{SNP}(\mathtt{X} = u'_i,\ \mathtt{Y} = v_i)$ based on the level of degeneration of the genetic code. Therefore, probability $\Pr_M(u, v)$ that u' mutates into v is given by the following formula:

$$\Pr_M(u', v) = \prod_{i=1..N \pm l} \left(p_{mut} \cdot \Pr_{SNP}(u'_i, v_i) + (1 - p_{mut}) \cdot [u'_i = v_i] \right),$$

where $[u'_i = v_i]$ is the Iverson bracket. Eventually, probability of rewriting $\Pr(u \to v)$ is given by the formula:

$$\Pr(u \to v) = \Pr_M(X_{p,l}(u), v) \cdot \Pr_X(\mathit{length}(u), p, l)$$

In addition Q is a set of constraints in the arbitrary form imposed on strings that model selection. E.g., a structural constraint: $Q_1 :\ \mathit{length}(u) \in [a, b]$ implies that probability of the copy offspring $\Pr_{\mathrm{Q_1}}(\mathrm{u^C}) = \Pr_C(L, l)$ and delete offsprings $\Pr_{\mathrm{Q_1}}(\mathrm{u^D}) = \Pr_D(L, l)$ is given as follows:

$$\begin{cases} \Pr_C(L,l) = \Pr_D(L,l) = 0.5 & L - l \geq a,\ L + l \leq b \\ \Pr_C(L,l) = 1,\ \Pr_D(L,l) = 0 & L - l < a,\ L + l \leq b \\ \Pr_C(L,l) = 0,\ \Pr_D(L,l) = 1 & L - l \geq a,\ L + l > b \\ \Pr_C(L,l) = \Pr_D(L,l) = 0.0 & \mathit{otherwise} \end{cases}$$

Other constraints can be implemented e.g. in the form of a formal grammar (see 3).

Informally, in the first step each repeat sequence first undergoes the unequal crossing-over. The position and length of the event are randomly chosen based on predefined distribution (e.g. uniform). The crossing-over produces two offspring: a Copy-offspring and a Delete-offspring. Then each position in the offspring sequence is subjected to mutation with probability p_{mut}. The mutation operator can substitute only to amino acid which is reachable by changing just one letter in the codon. Then constraints that model selection are applied. First are the *structural constraints*, which impose minimal and maximal numbers of repeats. If both offspring pass throughout this filter, each of them have 50% chance of getting selected to the next population. If one does not fit in the length limits, then the other is promoted to the next generation. Finally if both are not compatible with the constraints, the generation process restarts from the unequal crossing-over. After the structural filtering, other constraints are applied. They can be given in the form of a probabilistic formal grammar (with not too high complexity, e.g. regular). The better offspring fitness to the grammar relative to the parent fitness, the better chances the former has to get to the next generation (through *the roll of dice*). If it is not selected, the process restarts from the crossing-over.

The model was implemented as a standalone command-line application in C++11.

Characterization of repeats population We propose 49 features which describes amino acid repeat space and enable monitoring of the population:

- General:
  - number of repeats (or sequence length in repeat units, also called *exponent* or *order* in the combinatorics),
  - number of unique amino acids in the repeat sequence
  - average number of unique amino acids in the repeat sequence per length;
- Composition:
  - fraction of each amino acid in the sequence,
  - average Miyazawa-Jernigan hydrophobicity [17], van der Waals volume [8], and net charge [11,8] per sequence according to the AAindex database [10],
  - total charge in sequence,
  - total and average absolute charge;
- Recurrence Quantification Analysis (RQA):
  - REC: recurrence of the sequence, i.e. fraction of reoccurring amino acids per sequence,
  - DETpar: sequence (parallel) determinism, i.e. fraction of reoccurring parallel stretches ..AB..AB.. of amino acids of length 2 or more relative to REC,
  - DETant: sequence antiparallel determinism (as above but concerns antiparallel stretches ..AB..BA..),
  - RATIOpar and RATIOant: ratio of DETpar and DETant to REC,
  - REC, DET and RATIO for the physicochemical properties using the similarity threshold of 1.5 standard deviation.

  RQA is a technique developed for analysis of dynamic systems [35] which has been subsequently applied to protein sequences [34,4]. In this work we use its variables in an atypical setting of very short sequences as simple quantitative measures of sequence structurization. Because of the short length of analyzed objects, all RQA variables are calculated at embedding level of 1. In addition, we propose to use DETant in addition to standard parallel DET, which we rename therefore to DETpar.

3 Results

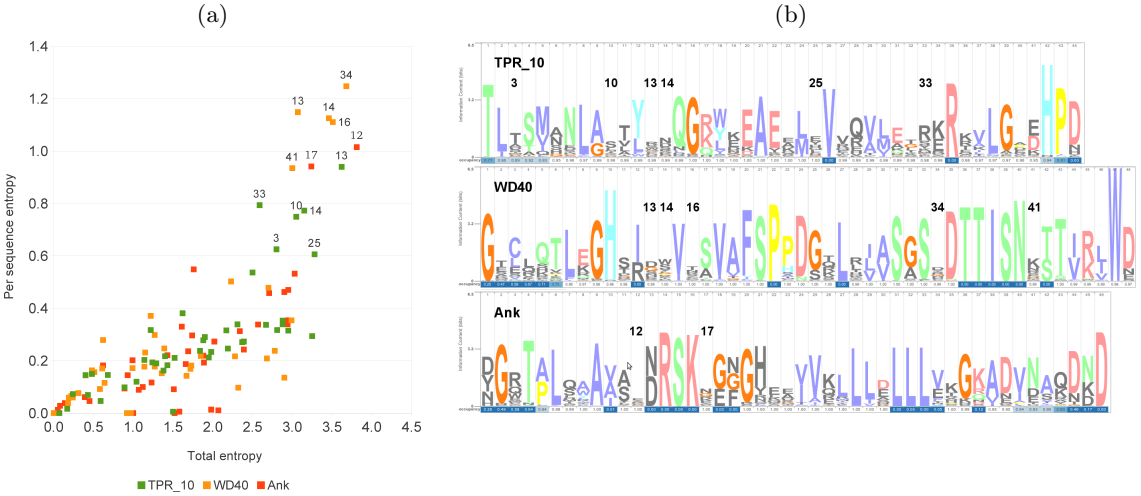
3.1 Identification of highly variable positions

HiC repeats can be represented by their highly variable sites because of almost perfect internal conservation of the other parts of repeat units. The highly variable sites were identified based on the sequence alignments of HiC repeats in PfamA Ank, WD40 and TPR_10 families. Entropy was calculated for each position in the alignments. A scatter plot of inter- and intra-sequence entropy for each position in the alignments revealed distinct groups of positions for Ank (2 sites) and WD40 (5 sites), and less clearly for TPR_10 (6 sites), see Figure 1a and Figure 1b. Out of these sites, four WD40, two Ank and two TPR_10 positions were recently found to be under positive selection [21,6].

3.2 Cross check with experimental data

A series of spontaneous mutants in the WD40 repeat domain of the HET-E1 NLR protein from *P. anserina* were recently isolated [3]. These mutants were suggested to represent a proxy for the natural evolution of the repeats. We thus chose this set of 9 mutants to test whether these mutants, represented by their highly variable sites, could be explained by a combination of unequal crossing and point mutation. Indeed, three same-length mutants can be explained by subsequent Copy and Delete events plus a Mutate event. Three out of four elongated mutants can be explained just by a Copy event, while two shrunk mutants are reachable by Delete plus Mutate events. One elongated mutant cannot be explained by a reasonably short sequence of unequal crossing-over and mutation events. Despite this last negative result, we conclude that unequal crossing-over and mutation operator are sufficient for the first approximation of the generative part of the repeat evolution model.

Fig. 1: Identification of Highly internally Conserved repeats. **(a)** scatter plot of total and per sequence entropy at each position in multiple sequence alignments of HiC repeats from Ank, WD40 and TPR_10 families found in fungal NLRs **(b)** logo representation [33] of the sequence alignments. Columns with many gaps are marked in blue. In both panels, highly variably sites are annotated by their positions in the alignments



3.3 General characteristics of the system

A full theoretical analysis of the system is beyond the scope of this paper. Here we only provide a few general experimental characteristics which are relevant for interpretation of further results. For the purpose of modeling, repeats were decomposed to their highly variable sites, i.e. sequences of residues from only one highly variable site are used at once. The basic setup assumed minimal and maximal sequence lengths of 1 and 14, respectively, and the "epsilon" (technically zero) overlap of crossed fragments. Note that intuitively the length of at least two repeats and the overlap over at least 1 repeat would be needed for the unequal crossing-over. However, repeats considered in this research are spanning over 30-40 residues, while just few identical residues is enough to initiate the unequal crossing-over [25]. Thus, one repeat and a small (*epsilon*) part of another one may enable the process. Moreover, this epsilon part is unlikely to contain any highly variable site. Therefore, in terms of representation by the highly variable positions, only one repeat and the zero overlap are required. The crossing-over position and length were randomly chosen using the uniform distribution, whereas probability of mutation was set to 0.1. We chose the uniform distribution for the crossing-over position because of experimental data for HiC WD40 repeats showing that there was no bias in the position of recombination events which occurred along the repeat array with similar frequency [3]. Similar results were obtained computationally for WD40 and TPR, whereas Ank repeats tended to expand in the middle [1].

Given the crossing-over parameters, the substitution matrix and the structural constraints, the system converged to the same solution space whatever was the original input population S. To demonstrate this, the system was initiated with 13 repeat sets (independently for each highly variable sites) and 49 sequence features were recorded every 100 steps during 100,000 generations. Means of features calculated over generations from 1000 onwards, were compared to means obtained with a system initiated with a dummy set of 100 identical `GGGGG` sequences, using the Welch t-test. Only two (11) out of 637 set-feature cases obtained p-value below 0.01 (0.05).

The pace of the convergence depended on the mutation probability. The number of unique sequences N visited after x generations closely obeyed the power law $N = ax^k$. For example, given $|S| = 100$, $p_{mut} = 0.1$, the coefficients were $a = 90.61$ and $k = 0.8838$.

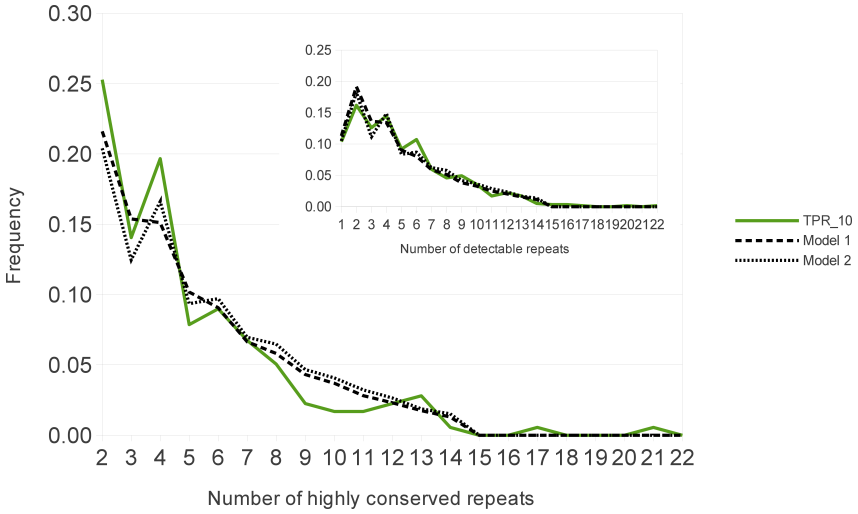
It is illustrative to relate the number of generations in our model to the timeline of evolutionary history. Assuming a 1 week reproduction cycle (as observed for *P. anserina*) and one unequal crossing-over per 1000 reproduction cycles, 100,000 crossing-overs would correspond to 2 million

years. For comparison, the common ancestor of *Asco-* and *Basidiomycetes* is believed to live 500 millions years ago (corresponding to 25 million crossing-overs).

3.4 Repeat number distribution

In [6], we examined the distribution of repeat number in Ank, WD40 and TPR_10 subsets of NLRs. The Ank and TPR_10 distributions resembled the negative binomial distribution. Moreover, an interesting peculiarity consisting of imperfect $i/i+2$ periodicity was observed in the case of ascomycetal and basidiomycetal TPR_10 and basidiomycetal Ank. This periodicity was however absent in the most populous ascomycetal Ank case. WD40 distributions were more complex, which probably reflected more stringent structural constraints imposed by the double beta propeller structure (in comparison to the solenoid structure formed by the other repeat types). Distributions of HiC repeat number fit a power-law behavior and were relatively similar in all three classes of repeats except that TPR_10 exhibited $i/i+2$ periodicity for small numbers of repeats (2-7) while the other for the large number of repeats (7-13). The repeat number distribution predicted by the model was calculated from all sequences from generation 101, at which point it converged to 10,000. We found that the $i/i+2$ periodicity with the first peak at number 2, observed in TPR_10, can be easily obtained by our model with minimal and maximal number of repeats set to 1 and 14, respectively, with minimal crossing-over overlap set to 0, and with uniform distributions of the crossing-over position and length (Model 1 in Figure 2). The peaks were even more pronounced if the probability of a single repeat copy/deletion was reduced to 1/10th in respect to other lengths (Model 2). This model is more consistent with the computational study by Björklund *et al.*, who determined that most repeat families (including Ank) expanded through duplication of several units at once [1]. Both models cannot be rejected by the chi-squared goodness-of-fit test when compared to observed distribution of HiC repeats or all detectable repeats of TPR_10 family (Model 1: p-values 0.50 and 0.27, respectively, Model 2: p-values 0.35 and 0.39, respectively). Therefore, we continue with the simpler Model 1.

Fig. 2: Comparison of simulated and observed distributions of repeat number in TPR_10. Simulated distributions (black) were calculated under Model 1 (dashed line) and Model 2 (dotted line) over high number generations after convergence and compared to the real distribution of the HiC repeats number (green). Note the minimum number of 2 repeat units. **Inset:** The same model distributions compared to the real distribution of all repeats detected by the PfamA TPR_10 profile. This comparison includes also single repeats.



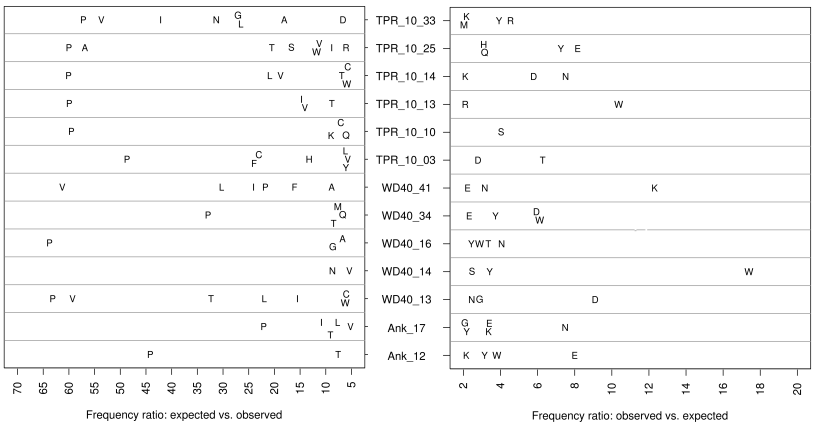
The $i/i+2$ periodicity vanishes when the difference between the minimal number of repeats and their overlap increases. Interestingly, if two subsequent events of unequal crossing-over were allowed before checking the length constraints (or selection), it was easy to obtain a repeat number

distribution with similar frequencies at lengths 1 and 2, which would correspond well to the Ank distribution for HiC repeats, extrapolated to length 1 based on the distribution of all detectable repeats. One possible difficulty in our approach is that in the real world single point mutations accumulating outside highly variable positions can result in the break down of high repeat conservation. This cannot be modeled in our approach. While it is possible that the most appropriate reference distributions for our model would be intermediary between observed repeat number distributions for all and for HiC repeats, this does not invalidate positive outcome of this test as we showed that the proposed model can generate observed repeat number distribution for all detectable and for HiC repeats.

3.5 Selective pressure

We showed that the system converged to the same solution space independently from the initial population. Therefore all differences in observed distributions at various highly variable sites must be a consequence of the selection (constraints) and the ratio of the mutation and the crossing-over rates. Moreover, the mutation rate and the crossing-over parameters must be identical at all positions that belong to the same repeat family. In order to model selection, we will use amino acid composition at the highly variable sites as the first approximation. Selective pressure can be simply described as amino acid frequency ratio between original data and the model (Figure 3). A +1 distribution smoothing was used to account for the limited real data available. With a notable exception at the WD40_14 site, there was universally a strong pressure against proline, while non-aromatic aliphatic amino acids (especially valine) were also under negative selection. On the other hand, tryptophan, as well as charged amino acids (except arginine), asparagine and tyrosine were positively selected, depending on the particular site. This picture is consistent with the presumed role of the highly variable positions as the recognition epitope. Indeed, in WD40 the variable sites are located in connecting loops which would be rigidified by proline, and in TPR_10 and Ank, where the variable sites are in helices, proline would also affect helix geometry. The counter selection of valine could be explained by its high beta-sheet propensity, so it should be disfavored in helical Ank and TPR_10 and in WD40 loops. Tryptophan is frequently found in protein-protein interfaces and also is often involved in glycolipid binding [36,16]. Tyrosine is also overrepresented in binding interfaces, while hydrophobic non-aromatic amino acids like leucine are rare [2].

Fig. 3: Selective pressure at highly variable sites. Selective pressure was computed as amino acid frequency ratio between original data and the model without selection. The +1 distribution smoothing was used to account for the limited real data available. **(left)** Amino acids at negative selective pressure (frequency ratio below 1:5). **(right)** Amino acids at positive selective pressure (frequency ratio above 2:1)



3.6 Modeling composition-based selection

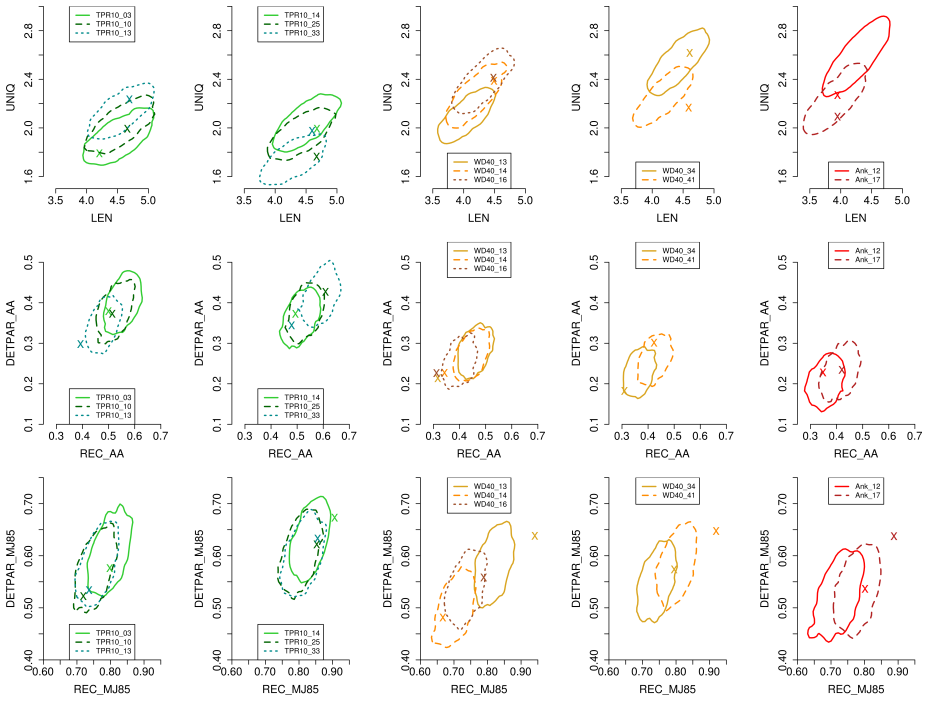
In the system, compositional selective pressure was accounted for by a simple weighted regular grammar with rules in the form: $\mathtt{S} \rightarrow \mathtt{a}\,\mathtt{S} \mid \epsilon.$, where $\mathtt{a}$ is an amino acid and ϵ is the empty symbol (sequence termination). This grammar is equivalent to a weighted unigram. Weights of the amino-acid-generating rules were assigned by an adaptive relaxation procedure such that the system converged, under given mutation probability, to the original composition (sum of deviations below 0.01, maximum relative deviation below 0.10). The sequence terminating rule $\mathtt{S} \rightarrow \epsilon$ had weight one (and therefore the grammar is not probabilistic). Parsing assigned each sequence a score equal to the geometric average of weights of applied rules (the average was taken to avoid a bias on the length). The proportion of the parent and child scores determined the probability that the new sequence was included in the next population.

Then, for each repeat family, a uniform single point mutation probability at different variable sites was sought on the basis of agreement between RQA parameters in the original and model dataset. Specifically, we counted cases where RQA feature means in the original populations were out of the range of 95% RQA feature means in generated populations (collected from 1001th to 100,000th iteration, modulo 100). The least number of discording cases was found for mutation probabilities of 0.150 for TPR_10, 0.275 for WD40, and 0.325 and 0.350 for Ank. For 8 out of 13 highly variable sites, the original means of sequence lengths (LEN) and of number of unique amino acids per sequence (UNIQ) were within the range given by the 95% quantile of joint distributions of model means of these features (Figure 4 top). Moreover, differences in UNIQ between different sites were usually well accounted. Notable exceptions were TPR10_25 at which real mean number of unique amino acids was smaller than in the model, and TPR10_33 and WD40_13 sites at which the opposite was true. In the case of 5 sites (TPR10_03,10,25, WD40_34 and Ank_12) all mean recurrences of amino acid identities and three physicochemical properties in the real populations were in the model ranges (see Figure 4 middle for amino acid identity and bottom for hydrophobicity). On the other hand, at the WD40_13 site, recurrence of amino acid identities $\mathrm{REC_{AA}}$was overestimated by the model at the same time as recurrence of hydrophobicity $\mathrm{REC_{MJ85}}$ was underestimated. This clearly indicates that, whatever the p_{mut} parameter, the current model, based on simple compositional pressure, is insufficient to explain recurrence observed in the real population at this site. Finally, distribution of average parallel determinism DETpar calculated for the model usually properly accounted for differences observed in the original datasets.

3.7 Exploration of solution space

The entire set of simulated sequences from a single run defines a solution space for repeats in terms of the feature vector describing the sequences. The location of the real sequences in this solution space can provide useful information regarding missing constraints. In principle this analysis can be done in an automated or in an interactive way. It is useful, especially in the case of interactive analysis, to reduce dimensionality of the repeat feature space. It can be either than by a simple projection onto 2 dimensions (features) of interest or by means of Principle Component Analysis, e.g. in the case of 20-dimensional space given by fractions of amino acids. Here, we apply the former approach to two highly variable site which least fitted the composition-only selection model, TPR10_33 and WD40_13 (Figure 5). The solution space defined in terms of $\mathrm{REC_{AA}}$ and $\mathrm{DETpar_{AA}}$ was divided into 81 square cells in which original and simulated sequences were counted. By applying the Fisher test (p=0.05) we found some significant deviations from the expected ratio of original and generated sequences. At both sites, sequences made of just one amino acid type were depopulated in the real populations (far right in both panels) in comparison to the model. On the contrary, sequences with both $\mathrm{REC_{AA}}$ and $\mathrm{DETpar_{AA}}$ around 0.5 were overrepresented in the original dataset. Moreover, in the case of TPR10_33, all 15 real sequences from the two significant regions in the middle (Figure 5 left) consisted of positively charged residues mixed with just one of the four: tyrosine, serine, glutamine and glutamic acid. Interestingly, 5 of them had a pattern $\mathtt{R}-[\mathtt{SYQ}](3)-\mathtt{R}$, according to the PROSITE-like notation. In addition 6 occurrences of a similar pattern $\mathtt{R}-[\mathtt{SYFW}](1,2)-\mathtt{R}$ were found in the subset with $\mathrm{REC_{AA}}$ around 0.3 and $\mathrm{DETpar_{AA}}$ of zero. In general, the combined pattern $\mathtt{R}-\{\mathtt{KR}\}(1,3)-\mathtt{R}$ occurred 10 times more often than expected from the model, which can suggest biological advantage of the motif. The sequences belong

Fig. 4: Density plots of pairs of selected feature means in model population. X marks position of the real population mean in the feature space. Mutation probability was 0.150 for TPR_10, 0.275 for WD40 and 0.350 for Ank. AA is amino acid identity, MJ85 is Miyazawa-Jernigan hydrophobicity, LEN is repeat number, UNIQ is number of unique amino acids per repeat sequence

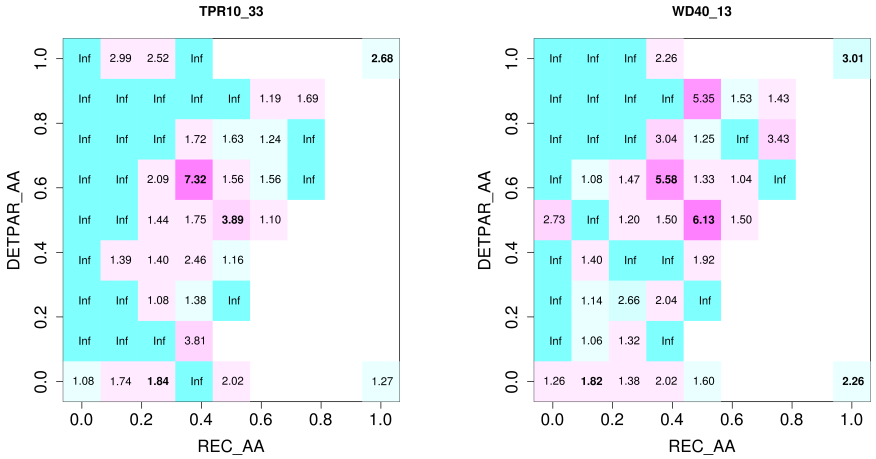


to species *T.stipitatus* (5 cases), *T.virens* (2 cases), *T.atroviride*, *P.tritici-repentis*, *B.maydis and A.bisporus*. No clear pattern was however found for WD40_13.

4 Conclusions

We have presented here a stochastic string rewriting system that models the generation process of highly internally conserved repeats. The system is grounded in the biology of the process as it models transformation of repeats through the events of unequal crossing-over and mutation, which are believed to be main mechanisms that produce diversity in repeats [19,24]. Extension of the model to account for other recombination processes, such as gene conversion and internal recombination through loops, is left to future work. We postulated that the solution space of repeats could be characterized by feature vector combing multiple quantitative parameters describing composition and determinism of the repeat sequence at levels of amino acid identity and physicochemical properties. We showed that in this solution space it converged to the same solution space independently from the initial population and therefore differences in observed distributions at various highly variable sites had to be a consequence of the continuous selective pressure rather than a legacy of the original population. The system was validated with regard to HiC repeats from fungal NLRs that belonged to three PfamA families. We showed that our model could explain the $i/i+2$ periodicity in the repeat number distribution observed for some classes of repeats in real populations. By comparing amino acid content in the model and original populations, we confirmed that highly variable sites identified on the basis of entropy, were subject to selective pressure towards composition typical for binding sites, which was coherent with the suggested role of recognition epitopes. Further simulations singled out sites at which variability can be explained by composition-only selection (e.g. position 3 in TPR_10) and those where it was not possible (e.g. position 33 in TPR_10 and position 12 in WD40). Moreover, they suggested different mutation rates for TPR_10 and two other families (this result held also when feature means weighted by repeat number were used). The difference corresponded to smaller intra-sequence entropy at TPR10 sites (Figure 1a).

Fig. 5: Comparison of real and model solution space. Magenta indicates overrepresentation of real sequences, while cyan - generated sequences. Exact ratio relative the expected ratio is given in each cell. Bold fonts marks significant deviations according to the Fisher's exact test.



We proposed an interactive approach to exploring the solution space of amino acid repeats by means of 2-d projections and significance tests in the sectors. In a sample case, we found an overrepresented pattern $\mathtt{R}-[\mathtt{SYQFW}](1,3)-\mathtt{R}$ at position 33 in TPR_10 family which potentially has functional importance. This kind of analysis could be supported by automatic inference of simple grammars or HMM profiles from sequences found in the sectors significantly enriched or depleted from original sequences. These grammars can be then used to further constrain the solution space in an iterative process towards accurate description of the plausible set of recognition epitopes. Another application of our model would be testing of hypotheses regarding repeat origin or function encoded as constraints on the string rewriting system.

Importantly, the model defined in this work is not specific to fungal NLRs. The methodology is general and therefore can be applied to any class of highly internally conserved repeats generated by unequal crossing-over (including NLRs from plants or animals) for which an equivalent high quality dataset is available.

References

1. Bjorklund, A., Ekman, D., Elofsson, A.: Expansion of protein domain repeats. PLoS Computational Biology 2, 114 (2006)
2. Bogan, A.A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. Journal of Molecular Biology 280, 1–9 (1998)
3. Chevanne, D., Saupe, S., Clave, C., Paoletti, M.: Wd-repeat instability and diversification of the podospora anserina hnwd non-self recognition gene family. BMC Evolutionary Biology 10(1), 134 (2010)
4. Colafranceschi, M., Colosimo, A., Zbilut, J.P., Uversky, V.N., Giuliani, A.: Structure-related statistical singularities along protein sequences: A correlation study. Journal of Chemical Information and Modeling 45, 183–189 (2005)
5. Danot, O., Marquenet, E., Vidal-Ingigliardi, D., Richet, E.: Wheel of life, wheel of death: A mechanistic insight into signaling by STAND proteins. Structure 17(2), 172 – 182 (2009)
6. Dyrka, W., Lamacchia, M., Durrens, P., Kobe, B., Daskalov, A., Paoletti, M., Sherman, D.J., Saupe, S.J.: Diversity and plasticity of nlrs in fungi (2014), under revision
7. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5), 1792–7 (2004)
8. Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A., Pliska, V.: Amino acid side chain parameters for correlation studies in biology and pharmacology. International Journal of Peptide and Protein Research 32, 269–278 (1988)
9. Jorda, J., Kajava, A.V.: T-reks: identification of tandem repeats in sequences with a k-means based algorithm. Bioinformatics 25(20), 2632–8 (2009)

10. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: AAindex: amino acid index database. Nucleic Acids Research 36, D202–5 (2008)
11. Klein, P., Kanehisa, M., DeLisi, C.: Prediction of protein function from sequence properties: Discriminant analysis of a data base. Biochimica et Biophysica Acta 787, 221–226 (1984)
12. Koonin, E., Aravind, L.: The nacht family - a new group of predicted ntpases implicated in apoptosis and mhc transcription activation. Trends in Biochemical Sciences 25, 223–224 (2000)
13. Leipe, D., Koonin, E., Aravind, L.: Stand, a class of p-loop ntpases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. Journal of Molecular Biology 343, 1–28 (2004)
14. Luo, H., Nijveen, H.: Understanding and identifying amino acid repeats. Briefings in Bioinformatics 15, 582–591 (2014)
15. Maekawa, T., Kufer, T.A., Schulze-Lefert, P.: Nlr functions in plant and animal immune systems: so far and yet so close. Nature Immunology 12,, 817–826 (2011)
16. Matsuzawa, T., Saito, Y., Yaoi, K.: Key amino acid residues for the endo-processive activity of gh74 xyloglucanase. FEBS Letters 588, 1731–1738 (2014)
17. Miyazawa, S., Jernigan, R.L.: Estimation of effective interresidue contact energies from protein crystal structure: quasi-chemical approximations. Macromolecules 18, 534–552 (1985)
18. Monroe, N., Sennhauser, G., Seeger, M., Briand, C., GrÃŒtter, M.: Designed ankyrin repeat protein binders for the crystallization of acrb: plasticity of the dominant interface. Journal of Structural Biology 174, 269–81 (2011)
19. Naidoo, K., Steenkamp, E., Coetzee, M., Wingfield, M., Wingfield, B.: Concerted evolution in the ribosomal rna cistron. PLoS One 8, e59355 (2013)
20. Paoletti, M., Saupe, S.J.: Fungal incompatibility: Evolutionary origin in pathogen defense? BioEssays 31(11), 1201–1210 (2009)
21. Paoletti, M., Saupe, S.J., Clave, C.: Genesis of a fungal non-self recognition repertoire. PLoS ONE 2(3), e283 (03 2007)
22. Post, E.L.: Recursive unsolvability of a problem of thue. The Journal of Symbolic Logic 12, 1–11 (1947)
23. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D.: The pfam protein families database. Nucleic Acids Research 42, D222–30 (2014)
24. Rando, O.J., Verstrepen, K.J.: Timescales of genetic and epigenetic inheritance. Cell 128, 655–668 (2007)
25. Razanamparany, V., Begueret, J.: Non-homologous integration of transforming vectors in the fungus podospora anserina: sequences of junctions at the integration sites. Gene 74, 399–409 (1988)
26. Saupe, S., Turcq, B., Begueret, J.: A gene responsible for vegetative incompatibility in the fungus podospora anserina encodes a protein with a gtp-binding motif and g beta homologous domain. Gene 162, 135–139 (1995)
27. Schaper, E., Gascuel, O., Anisimova, M.: Deep conservation of human protein tandem repeats within the eukaryotes. Molecular Biology and Evolution 31, 1132–1148. (2014)
28. Schaper, E., Kajava, A.V., Hauser, A., Anisimova, M.: Repeat or not repeat? - statistical validation of tandem repeat prediction in genomic sequences. Nucleic Acids Research 40, 10005–10017 (2012)
29. Selker, E.U.: Repeat-induced gene silencing in fungi. Advances in Genetics 46, 439–450 (2002)
30. Stumpp, M.T., Binz, H.K., Amstutz, P.: Darpins: A new generation of protein therapeutics. Drug Discovery Today 13, 695–701 (2008)
31. Szostak, J.W., Wu, R.: Unequal crossing-over in the ribosomal dna of saccharomyces cervisiae. Nature 284, 426–430 (1980)
32. Thue, A.: Probleme uber veriinderungen von zeichenreihen nach gegebenen regeln. Skriffer utgit av Videnskapsselskapet i Kristiania, I . Matematisk-natu idenskabelig klasse 10, 34 (1914)
33. Wheeler, T., Clements, J., Finn, R.: Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden markov models. BMC Bioinformatics 15, 7 (2014)
34. Zbilut, J.P., Sirabella, P., Giuliani, A., Manetti, C., Colosimo, A., Webber, C.L.: Review of nonlinear analysis of proteins through recurrence quantification. Cell Biochemistry and Biophysics 36, 67–87 (2002)
35. Zbilut, J., Webber, C.: Embeddings and delays as derived from quantification of recurrence plots. Physics Letters A 171, 199–203 (1992)
36. Zielezinski, A., Karlowski, W.M.: Integrative data analysis indicates an intrinsic disordered domain character of argonaute-binding motifs. Bioinformatics in press (2014)
37. Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E.J., Jones, J.D.G., Felix, G., Boller, T.: Bacterial disease resistance in arabidopsis through flagellin perception. Nature 428, 764–767 (2004)