Distributed computing with prokaryotic immune systems

Niall Murphy[1,2] and Alfonso Rodríguez-Patón[1]

[1] Facultad de Informática, Universidad Politécnica de Madrid, Spain.
[2] CEI Campus Moncloa, UCM-UPM, Madrid, Spain.

Summary: We propose the *in vivo/vitro* use of prokaryotic adaptive immune systems for distributed learning. In the coming years synthetic biologists will learn to control, program, and modify such systems. We design an enhancement to CRISPR/Cas immune systems and demonstrate the learning potential of the modified system by showing that it can approximate solutions to a computationally hard problem. To our knowledge this is the first proposed use of CRISPR/Cas systems for computational purposes.

Biological background: CRISPR/Cas systems are a family of adaptive immune systems found in prokaryotes [1, 2]. They allow cells to develop and inherit immunity to viruses and plasmids.

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) locus is a noncoding region of DNA with a "leader" section ($\sim$ 100–500 bp) followed immediately downstream by "spacers" ($\sim$ 30–40 bp) which are interspersed with identical, partly palindromic, repeat sequences. There can be between 20 to 300 spacers in a single CRISPR locus. Spacers typically correspond to sections of DNA from viruses and plasmids.

Once the CRISPR locus is transcribed, one of the Cas proteins separates the spacers by cutting the mRNA in the middle of the repeat sequences. Each individual spacer then forms part of a crRNA (CRISPR RNA). A crRNA then directs cleavage of any DNA (for example that of a virus or plasmid) with the specific target sequence.

The mechanism by which new spacers are inserted (immediately downstream of the leader) is not yet fully understood.

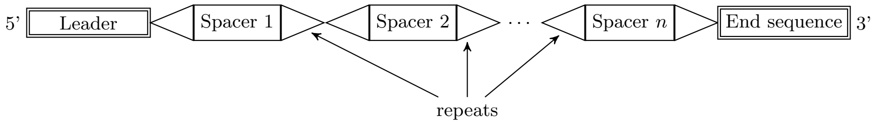

Fig. 1. A schematic diagram of a CRISPR locus showing the location of the leader sequence, the spacers, and the palindromic repeat sequences are represented by "◁" and "▷". We tag the end of the CRISPR locus with a known end sequence.

Motivation: CRISPR/Cas systems are adaptive and inheritable systems that detect and destroy foreign nucleic acids. CRISPR/Cas systems are not yet fully understood but biologists are already turning them into tools for use in synthetic biology [3, 4]. In the near future our ability to program the spacer list will improve and we will unlock the full potential of these systems for bio-technology, synthetic biology, and medicine [1]. CRISPR loci can also be transmitted from one bacterium to another by conjugation, this makes them a powerful tool to modify the plasmids circulating in a bacterial population.

Synthetic Biology: It is hypothesised that as a CRISPR locus gets longer, spacers are lost by homologous recombination between the repeat sequences [2]. However, if CRISPR spacers were constantly being removed by some synthetic system (for example, removing the last spacer in the locus using a restriction enzyme and ligase that match the "end sequence" in Figure 1) then the CRISPR locus would become a shorter, queue like, DNA memory.

We use the name "FORGETR" to refer to a hypothetical (re)engineered CRISPR/Cas system where the rates of spacer insertion and loss can be tuned. Also any arbitrary section of viral genome can be used as a spacer.

Let us consider a chemostat where the inflow and outflow rate keeps a heterogenetic set of bacteriophages mixed in equal proportions. Also in the chemostat is a colony of bacteria with a FORGETR system.

As each phage attacks the cell there is a certain probability that the FORGETR system will insert a section of viral DNA as a spacer in its CRISPR locus and destroy the phage. If this occurs, then the cell can divide and each daughter cell inherits the newly augmented CRISPR locus. If the cell fails to insert a new spacer, the cell will be killed by the page.

The spacers that offer protection against the most phages will allow those cells to survive and will appear in more CRISPR loci in the population.

Using simulations we have identified a ratio between the rates of spacer insertion and spacer loss such that the loci are as short as possible and the most common spacers in the population are those that provide protection against the most phages (assuming that each page has the same rate of infection).

If this experiment were carried out we predict that the most common spacers would match the most common or conserved sequences in the genomes of the viral population.

To prevent certain "forbidden" sections of virus DNA becoming spacers in a CRISPR locus we propose a toxin/anti-toxin system [5]. The forbidden viral DNA sequences are broken up and distributed around a plasmid. This plasmid codes for its own stability in the population and also the genes for a slowly decaying toxin and a fast decaying anti-toxin. If the FORGETR system learns a "forbidden" section of virus DNA then the plasmid is destroyed. The anti-toxin will no longer be produced and the cell will die.

Bacteria with FORGETR systems have many potential applications such as: a tool to study the mutations and genetic diversity in a viral population; to train a CRISPR locus to develop a "vaccine" plasmid that can be used to harden cells against a wide variety of viruses; and the development of bacteria that remove a particular trait from a bacterial population such as antibiotic resistance.

Computation: To demonstrate the distributed learning potential of population of bacteria combined with a FORGETR system, we outline how it can approximate minimum solutions to HittingSet, a classic NP-complete problem [6]. Let (D, P) be an instance of the HittingSet problem where $P = (P_1, \ldots, P_m)$ is a collection of subsets of a finite set $D = \{d_1, \ldots, d_n\}$. The minimum hitting set is the smallest set $C \subseteq D$ such that C contains at least one element from each subset in P.

We encode the set D as a set of unique DNA sequences. We encode each subset P_i in the set of subsets P as a virus whose DNA includes a sub-sequence that is the encoding of each $d \in D \cap P_i$.

We consider a well mixed experimental set-up with an infinite supply of each virus in P attacking a population of bacteria with FORGETR. The CRISPR locus, C_b, of each bacterium in the population contains a single potential hitting set solution.

The viruses attack the bacteria. If virus P_i attacks the cell with CRISPR locus C_b and there is an intersection between C_b and P_i then the virus is destroyed and the cell can divide (both children have a copy of the CRISPR locus).

If the CRISPR locus C_b does not have an intersection with P_i then with probability λ it inserts an element d in P_i to its CRISPR locus C_b. If this occurs the virus is destroyed and the cell can divide (both cells have a copy of the new CRISPR locus).

If C_b has no intersection with P_i and fails to learn an element of P_i then the cell is destroyed and the potential solution C_b lost.

Clearly cells whose CRISPR locus C_b has an intersection with all the viruses of P will survive. To force the system to approximate minimal solutions, we make key use of the forgetting aspect of FORGETR. With a probability ρ the last element of the CRISPR locus is removed.

This setup forces the CRISPR loci to be as short as possible while still retaining as much immunity to the viruses as possible. In simulations we find that the system rapidly reaches the state where the most common sets of spacers in the bacterial population are in fact minimal solutions to the input instance.

References

1. Horvath, P., Barrangou, R.: CRISPR/Cas, the immune system of bacteria and archaea. Science **327**(5962) (2010) 167–170
2. Terns, M.P., Terns, R.M.: CRISPR-based adaptive immune systems. Current Opinion in Microbiology **14**(3) (2011) 321–327
3. Hale, C., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A., Glover III, C., Graveley, B., Terns, R., Terns, M.: Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Molecular Cell **45**(3) (2012) 292–302

4. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E.: A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science (2012) (Epub ahead of print)
5. Hayes, F.: Toxins-antitoxins: Plasmid maintenance, programmed cell death, and cell cycle arrest. Science **301**(5639) (2003) 1496–1499
6. Garey, M.R., Johnson, D.S.: Computers and Intractability, A Guide to the Theory of NP–Completeness. W. H. Freeman and company, New York (1979)