Evolution of DNA Uptake Signal Sequences

Dominique Chu
School of Computer Science
University of Birmingham
Birmingham B15 2TT
United Kingdom
and
National Center for Theoretical Sciences
Hsinchu, Taiwan 30043
Republic of China
Dominique.Chu@svt.uib.no

Hoong-Chien Lee
Department of Life Sciences and Department of Physics
National Central University
Chungli, Taiwan 320
Republic of China
hclee@phy.ncu.edu.tw

Tom Lenaerts
IRIDIA
(Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle)
Université Libre de Bruxelles
Brussels
Belgium
tlenaert@ulb.ac.be

Abstract The DNA of some naturally competent species of bacteria contains a large number of evenly distributed copies of a short sequence. This highly overrepresented sequence is believed to be an uptake signal sequence (USS) that helps bacteria to take up DNA selectively from (dead) members of their own species. For some time it has been assumed that the USS evolved in order to enable bacteria to distinguish between conspecific and non-conspecific DNA fragments (the *preference-first* hypothesis). Recently, Redfield suggested that this hypothesis is not in fact realistic, as it would require biologically implausible group selection. In this article we present a model designed to demonstrate the emergence of similar USSs in a population of simulated evolving agents. We use this model to examine the conditions under which a USS will emerge in a preference-first scenario.

Keywords
Uptake signal sequences, agent-based modeling, natural transformation, group selection

1 Introduction

1.1 Natural Competence and Uptake Signal Sequences

Many bacteria are, at least under certain conditions [11], *competent*, meaning they have the ability to take up (relatively short) DNA fragments from their environment through genetically programmed developmental pathways [8, 17]. Under certain circumstances the taken-up fragments may be incorporated into the bacterium's genome through recombination; this process is often referred to as *natural transformation*. Natural transformation is one type of horizontal gene transfer available to

bacteria. The others are *conjugation* and *transduction*. Those are not genetically controlled, but are rather mediated by direct cell contact and phages, respectively.

It is generally assumed that transduction and conjugation are mere side effects of other cellular processes and not in themselves an adaptation [7, 14], whereas biologists believe that natural competence is an adaptation. One indication in this direction is that it is genetically controlled. As to the precise benefit of natural competence, there is currently no consensus among biologists. External DNA certainly has a nutritional value as a source of nucleotides [10, 12]. Acquired homologous DNA fragments can also be used as patches for DNA repair or for recombination. It is furthermore hypothesized that DNA uptake plays a crucial role in the development of bacterial resistance to antibiotics.

While many of the competent bacterial species are indiscriminate as to what type of DNA they take up, others have a strong preference for *conspecific* DNA fragments, that is, those from dead members of their own species. Some of the species with this preference have a highly repeated uptake signal sequence (USS) on their DNA [15]. The USS mediates the uptake of DNA fragments from the environment. Examples of such species are *Haemophilus influenzae*, *Neisseria gonorrhoeae*, and *N. meningitidis*.

In *H. influenzae* [6, 16] the USS is 9 bp long (AAGTGCGGT) and rather evenly distributed throughout the DNA. Being repeated over 1,400 times (on both strands), the USS is statistically highly overrepresented on the DNA; on a random DNA with the same G/C content one would only expect approximately 8 copies. About 34% of the USSs are in noncoding parts of the genome (which makes up about 10.4% of the DNA). Bakkali et al. [1] suggested that the remaining 56% of USSs are contained in parts of genes that code for nonessential parts of the protein. USSs in other species have similar statistical properties [1].

The evolutionary forces leading to the USS are currently still unclear; researchers consider two scenarios [1, 5]:

1. *USS first:* Naturally competent bacteria had a certain preference to bind to USSs. The high USS content is a result of recombinational inclusion of bound DNA fragments containing the USS.
2. *Preference first:* Conspecific DNA is more beneficial than nonconspecific DNA. The USS evolved as a signal to allow bacteria to decide whether a piece of DNA stems from a member of the own species or not. Those that could effectively recognize conspecific DNA fragments had higher fitness.

Which scenario is the correct one is still an open question. A mathematical model described by Johnson and Redfield [6] suggests that the first one is at least a plausible route to the emergence of USS. On the other hand, an argument for the preference-first scenario is that there are naturally competent species that selectively take up conspecific DNA without having a USS; this suggests that uptake of conspecific DNA is beneficial in itself. This could be the case if the DNA fragments are used for recombination or DNA repair. However, the DNA taken up by competent bacteria stems from dead organisms and might be substantially degraded; this is a substantial risk with potentially lethal consequences for the recombining bacterium [1, 15].

Another argument that has been brought forward against the preference-first scenario directly concerns the existence of efficient selection mechanisms for USSs. Redfield pointed out that carrying a USS is not actually useful to the donor cell, because it will be dead by the time fragments of its DNA can be taken up. Given this, she hypothesized that USSs do not evolve unless we assume (strong) group selection pressure [5, 13]:

> However, selection for more USSs in the genome is problematic, because USSs only become useful after the genome they reside in has been released by lysis of the cell. Thus selection for genomes with many USSs might require biologically-improbable group selection to support the extreme altruism of cell death. [13]

The basic idea of this argument seems to be that a mutation that leads to USS-mediated competence will for the bearer be either detrimental or at best neutral. In either case there will be no selective advantage for agents carrying this mutation. On the other hand, if the population had efficient uptake signaling mechanisms, then by the assumption of the preference-first alternative, the population as a whole would be better off than without the mechanism. From this Redfield seems to conclude that only selection on the group as a whole can lead to USSs. Such selection forces in turn are biologically unrealistic; hence the preference-first scenario ought to be rejected.

We are not convinced that USSs can only emerge through a group selection mechanism. An alternative to the admittedly implausible group selection is a more plausible kin-selection mechanism [3]. If not individual agents, but genes, are selected for, then a mutation for a USS uptake mechanism will propagate through the population if it confers enough benefit to the offspring of the bearer of this gene (which is also a bearer). While the agent that carries the initial mutation will indeed not have increased fitness from having an uptake mechanism, its immediate offspring has at least a small advantage; once the respective gene is fixed in the population, there will be an advantage to all members of the species. The main problem is thus whether or not the initial emergence is possible. One assumption of this scenario is that there is no significant disadvantage in having an uptake mechanism.

A functioning USS system requires at least two separate components to be in place. Firstly, there needs to be a signal that is reasonably well spread over the DNA, and secondly, there needs to be a recognition (binding) mechanism to selectively take up DNA with the particular sequence signal. It would be biologically implausible to assume that such a system springs into existence in one piece. We should therefore assume that at least one component of a signal-mediated uptake system is already in place.

A possible scenario for the emergence of USSs under the preference-first hypothesis is as follows: This scenario rests on the observation that DNA sequences are not perfectly random sequences but contain biases, most notably highly repeated subsequences [4]. The origin of those biases must be thought of as being independent of the uptake signaling mechanism. Under the assumption that a bias in the genome is already present in the population, a mutation that leads to a recognition (binding) mechanism for one of the already present overrepresented subsequences will then—by assumption of the preference-first scenario—be of immediate benefit to the bearer during its lifetime; this is so because this mutation allows preferential uptake of conspecific DNA fragments. It is thus not necessary to evoke group or kin selection in order to explain the emergence of USSs.

Initially any evolved uptake mechanisms will probably not be particularly efficient. There are several ways to improve the recognition of conspecific DNA fragments. The obvious solution is to increase the number of copies of the signal sequence (i.e., the sequence bacteria bind to) on the DNA. Another possibility is to improve the distribution of the USSs. If the signal is evenly distributed over the genome, then the chances that it will be contained in a random fragment (and thus be recognized) are much greater than if all copies cluster in one small part of the genome. Finally, recognition efficiency can also be improved by optimizing the length of the signal sequences; if the sequences are too long, then it is very likely that they will not be contained entirely on a randomly chosen DNA fragment. On the other hand, if they are too short, then they cease to be efficient signals.

Increasing the abundance of signal sequences on the genome will normally not come without a cost to its bearer, particularly if the additional copies are in coding areas. Recent analysis of the location of USSs within genes showed that USSs are usually in parts of the gene that are coding for relatively nonessential parts of the corresponding proteins [1]. This suggests that the costs of having the USS there can be kept low. It remains a fact, though, that in the preference-first scenario there is an adaptive tradeoff between the benefits of an efficient uptake signaling system and the cost of additional copies of USSs.

This observation that additional USSs do come at a cost does not remain without consequences for the interpretation of the preference-first scenario. Any additional copy of the USS comes with a fitness penalty in the first generation; in order for this mutation to sustain itself in the population,

this must be balanced by a fitness advantage (from increased uptake specificity) in the following generations. One can assume that real bacteria will optimize the probability of finding the USS on a fragment while keeping the damage acceptable.

1.2 Aim of the Article

In this article we will explore the plausibility of the preference-first hypothesis, that is, we assume that the uptake of conspecific DNA is more beneficial than the uptake of other DNA fragments. Our aim is to investigate under which conditions a genetic uptake signaling system can emerge from an initially random DNA in the preference-first scenario.

A detailed analysis of specific statistical properties of the USSs in bacterial genomes—such as their absolute number, their relative abundance in coding versus noncoding areas, or other biological details—is outside the scope of this article. Also, we will not investigate the effects of costs of USSs on the genome. The investigation of those effects will be left to future contributions. The main purpose of this article is to contribute to the discussion of the evolutionary origin of the USS by examining the overall plausibility of the preference-first scenario.

1.3 Method

We will approach our questions by studying a simple and abstract agent-based model (ABM) of the emergence of uptake signals. The model is based on the following minimal, but biologically plausible, assumptions:

- Agents can choose between two types of food, namely, what we call *bacterial* (fragments of the DNA of dead agents, roughly corresponding to conspecific DNA in the real world) and randomly generated (henceforth *alien*) food fragments.
- The length of the fragments is small compared to the length of the whole DNA.
- Uptake of bacterial DNA fragments is more beneficial than uptake of other DNA.
- Agents can compare DNA fragments with genetically determined stretches of their own DNA and selectively take up (reject) matching (nonmatching) fragments. Henceforth we will refer to the part of the DNA against which a fragment is compared as the *reference sequence*, or simply the *reference*.
- Mutation events will over time change both the DNA of the agents and the details of how fragments are compared with the DNA.

Will mutation and selection lead to agents with USSs, that is, agents that carry a characteristic repeated subsequence and preferentially take up DNA fragments carrying this subsequence? As in real organisms, the emergence of uptake mechanisms requires two things to happen simultaneously. Firstly the random initial DNA has to acquire a statistically overrepresented subsequence. Such a sequence should be neither too short nor too long. Secondly, the agents need to recognize the repeated subsequence (i.e., preferentially take up fragments that contain it). If a suitable signal sequence already exists on the DNA of the agents, a successful strategy will be to set the reference sequence to one of the repeated subsequences. Depending on how many copies of the uptake signal are in the DNA, the agents will then more or less effectively recognize and specifically take up fragments containing the signal (because the agents compare the fragments with the reference; see assumptions). Of course, to set the reference sequence to an already existing highly repeated sequence is a rather unchallenging evolutionary task and of little interest to us.

Much more challenging is the case where no suitable signal sequence exists to begin with. If the initial DNA is random (i.e., has the statistical properties of a random sequence), then the hurdles on the way to the emergence of a USS are far greater. As we will indicate in more detail below, the major problem for the agents will be the initial establishment of a moderately effective uptake

signal: Differential fitness will then tend to improve the quality of the signal sequence, but significantly so only if the signal has already reached a certain quality. For this to happen it is necessary that the repeated sequence that is used as a signal be present in a sufficiently large number of copies that the bearer can profit from it. Once the initial hurdles are overcome, an evolutionary runaway process will ensure that USSs proliferate in the population and more and more copies are produced on the DNA.

The outline of this article is as follows: Section 2 describes our minimal agent-based model,[1] which demonstrates the emergence of uptake signals, and Section 3 discusses the relevant mechanisms that drive the dynamics of the model; Section 4 presents the results obtained from simulating the model using various parameters. We discuss the results in Section 5 and conclude in Section 6.

2 The Model

In this section we will describe a highly simplified ABM of the emergence of USSs. Following the agent-based paradigm, the model consists of adaptive agents embedded in an environment following simple, prespecified behavior rules. We will discuss the agents, their environment, and the rules.

2.1 The Agents

All agents in the model are of the same type and share a set of fixed parameter values. Agents live in a one-dimensional environment consisting of N discrete cells (N being user-defined). Throughout their lifetime they are immobile; however, an offspring is placed in one of the adjacent cells with probability 2/3. Agents only perform two actions, namely, uptake of food from the environment, and reproduction, that is, the placement of identical (or near identical) copies of themselves in the environment. There is no direct agent-agent interaction. If an agent reaches a certain age (defined in time steps), then it will die (i.e., be removed from the environment) with a certain probability per time step. There is also a limit age at which agents are killed with probability one.

The maximum overall number of agents in the environment is fixed (user-defined) throughout a run. Once the population number has reached this limit, then one agent must be removed from the environment (die) for each one born. The victim is always chosen among the currently oldest agents. Agents can only reproduce if they have collected a certain (user-defined) minimum number of energy units; once this limit is reached, they will at each time step reproduce with probability 1/2. The reproducing agent pays half of its energy units as reproduction tax, and the offspring obtains one energy unit at birth.

Energy units can be obtained by taking up food items from the environment. Agents take up exactly one food item[2] per time step, which is then converted into energy units. The exchange rate of food items into energy is user-defined and depends on the type of food (see below).

Besides its position, age, and energy holdings, each agent has two evolvable parts, which we call the GENOME and the DNA, respectively. The latter is a *dnalength*-long string of the letters {a, c, g, t} (*dnalength* is user-defined). At the start of a run each agent is initialized with its own, random DNA. Unless indicated otherwise, in all runs we seeded the initial population with the maximal number of agents possible in order to ensure optimal diversity.

The second evolvable component of the agent is the GENOME, which consists of three integers (*ORIG*, *OFF*−, *OFF*+) that define a subsequence of the DNA (the reference sequence). *ORIG* takes a value between 0 and *dnalength*, determining a position on the DNA (see Figure 2). The two integers *OFF*− and *OFF*+ respectively define a stretch of DNA to the left and to the right of the position

1 The C++ source code of the model can be obtained from the authors on request.

2 There is a fixed maintenance tax for the agents, but it was set to zero in all runs. No agent was therefore threatened by starvation (i.e., zero energy).

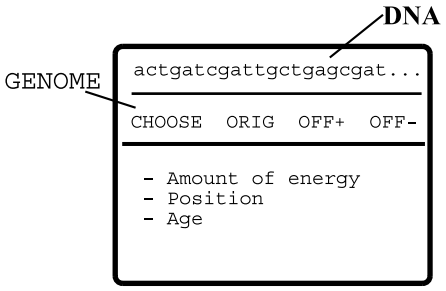

Figure 1. An agent consists of DNA, the GENOME, and variables.

defined by *ORIG*. Also a part of the GENOME is the Boolean variable *CHOOSE*; its function will be described below.

The DNA and the GENOME of an agent are identical to those of its parents, but with a user-defined probability they are subjected to slight changes (mutations) at birth. If an offspring agent mutates, then a fair coin is tossed to decide whether the GENOME or the DNA will be mutated. If the former is chosen, then either the parameter *CHOOSE* is flipped, or one of the integer parameters (*ORIG*, *OFF*–, *OFF*+) is changed by ± 1 (subject to the constraint that the parameters must retain meaningful values, i.e., continue to define a subsequence of the DNA).

There are two possible ways to mutate the DNA, *point mutations* and *copy mutations*. Point mutations change the letter at a randomly chosen position of the DNA. Parent and offspring DNA will disagree in only one position after the point mutation. Copy mutations are the replacement of a target substring of the DNA by a source substring of the same length. Note that target and source may partly or entirely overlap. Source and target are randomly chosen (both length and position), but there is a (user-defined) upper limit for their length. A process very similar to copy mutations has recently been identified as an important force in genome evolution [5]. While our implementation of copy mutations is not entirely biologically realistic, it is intended to be a source of bias rendering the DNA nonrandom (a similar source in real organisms would be for example gene duplication [9]). Also user-defined is the probability that a given DNA mutation is a point mutation.

2.2 The Environment

Each agent lives in one of the N cells that together form the environment. Every cell is adjacent to two other cells. Beside the agents, the cells also hold *DNA fragments* that can be taken up by the agents. Whenever an agent dies, a randomly chosen, fixed-length subsequence of its DNA is chosen and added to the DNA fragment collection of the cell in which the agent lived (the rest of the agent being discarded). In what follows we will refer to those fragments as *bacterial* DNA fragments. The number of bacterial DNA fragments in every cell is limited (user-defined). Once the limit is reached, for every new bacterial fragment added to the collection, the fragment that has been in the collection

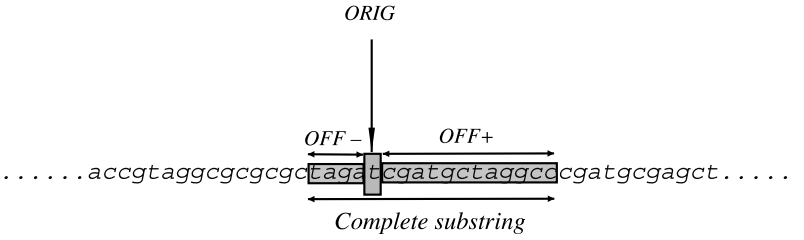

Figure 2. *ORIG*, *OFF*+, and *OFF*– uniquely determine a substring of the DNA—the reference sequence. In this particular case we have *OFF*+ = 12 and *OFF*– = 4.

for the longest period of time will be removed. This simulates the degeneration of DNA over time. In addition to the bacterial DNA fragments there is also a collection of randomly generated DNA fragments (henceforth called *alien* DNA fragments). The length of a fragment is determined by the user.

2.3 Rules for Uptake of Food

At every time step, each agent is offered a number of DNA fragments for uptake. Most of the presented fragments are alien, but with a certain (user-definable) probability each fragment is drawn from the bacterial DNA collection of the grid cell. (Throughout all simulations we report here, this probability is set to a value that ensures that on average each agent is presented one bacterial food item per time step.) If the *CHOOSE* parameter of the GENOME takes the value *false*, then the first of the presented food items is chosen and converted into energy. *CHOOSE* thus determines[3] whether an agent attempts to find a bacterial fragment or not. Otherwise, the agent compares the offered food items with its reference sequence. If the reference is a subsequence of the food fragment, then the agent takes up the fragment; otherwise the next food fragment is inspected, and so on, until a match is found or the last one has been unsuccessfully inspected. If an agent did not find any matching food item, then it will take up the last one presented. Once a food item has been taken up, the agent leaves the feeding mode.

For example, let us assume that the reference sequence is "aactt" and the agent is offered the three fragments "acttggtata," "acacgaactt," and "aacttaactt," (in that order). It will then reject the first sequence, take up the second (because it contains the reference), and never inspect the third.

If an agent accepts a bacterial DNA, then the corresponding fragment is removed from the system; chosen alien fragments are returned to the fragment pool. This, together with the uptake algorithm, guarantees the continued availability of DNA fragments. The contents of the alien fragment pool is renewed at certain user-defined intervals.

The amount of energy that alien and bacterial fragments can be converted into is a user-defined parameter. Note that it is necessary for bookkeeping purposes that the fragments carry tags that—though invisible to the agents—allow the program to determine whether a chosen food item yields high or low energy; however, it is important to understand that the agents are prevented from reading this tag.

3 Discussion of Model Design

While there is no explicit external fitness function in the current model, its design implicitly specifies a criterion for fitness: Efficiency in distinguishing between bacterial and alien fragments. Evolutionary success is thus largely determined by the way in which the agents compare food fragments with their own DNA. This drastically limits the potential richness of the evolutionary dynamics to be observed, but increases conceptual clarity and facilitates the interpretation of the results.

3.1 Adaptive Pressure

All agents are guaranteed to take up exactly one food item per time step, providing them with enough energy to survive and reproduce. Food per se is thus not a limiting resource in the model. The real limiting resource that drives the evolution of the agents is the assumed carrying capacity of the environment (maximal population size). The removal of the oldest agents in the model whenever an agent sends offspring into an overcrowded world is the driving force of evolution in the model. If an agent does not reproduce fast enough, then it may be removed from the system before getting the chance to do so. At the beginning of the simulation it is mostly a matter of luck whether or not an agent is able to reproduce before being removed: All agents collect the same amount of energy

3 The *CHOOSE* parameter is not strictly necessary in the current model framework. Besides extendability of the model, another reason to have it anyway is that it helps to speed up simulations before the USS emerges: If no USS is present in the population, then on average only half of the agents will then engage in the computationally expensive string matching.

per time step and will thus reproduce at roughly the same age (agents cannot discriminate alien and bacterial fragments); some (approximately one in twenty), though, will be lucky enough to find bacterial fragments. Altogether, at early stages of the simulation no agent or lineage will have a sustained advantage over any other.

Once, however, some agents start to collect energy units at a higher rate than others, they will effectively decrease the probability of their being removed before reproducing (because they will reproduce earlier). There is only one way to significantly and sustainably increase the collection rate of energy units: Agents must find ways to tell apart bacterial and alien fragments. Those that can do this more efficiently will fulfill the minimum energy requirements for reproduction at a younger age and thus have a smaller chance of being killed before they reproduce. Depending on the parameter settings, some may even reproduce twice.

How can agents improve their chances to correctly recognize bacterial fragments (and thus increase their fitness)? In order to develop an intuition for the challenges agents are facing at the beginning of a simulation run, let us assume that an agent is presented with fragments taken from an ancestor identical to itself:

- The length of the DNA is L.
- The length of the fragment is f.
- The length of the reference sequence of the agent (as defined by her GENOME) is u.
- Altogether there are n copies of the reference sequence on the DNA.

It is then possible to write down the probability that an agent correctly identifies a bacterial fragment; this is essentially the probability that a randomly chosen piece of length f (taken from a string of length L) contains a *specified* subsequence (of length u) of the L-long string. Since there are altogether n copies of the reference sequence on the DNA, exactly n of the L letters of the DNA are last letters of one of those copies. Thus, the probability that any particular letter of the DNA is *not* such a last letter is $1- n/L$. The sought probability is equal to the probability that the fragment contains at least one copy of the reference, which in turn is the probability that at least one of the last $f - u$ letters of the fragment is a last letter of a copy of the reference sequence:

$$P_1 := P(\text{fragment contains reference}) = 1 - \left(1 - \frac{n}{L}\right)^{f-u+1} \tag{1}$$

At the beginning of a simulation there will typically be only one copy of the reference on the DNA (thus $n = 1$); if $L = 10{,}000$ and $f = 150$, then a typical initial length of the reference sequence will be $u = 50$; with those values the probability of actually correctly identifying a bacterial fragment is then $P_1 \approx 0.005$, that is, approximately one in 200 presented sequences will be correctly recognized as a bacterial fragment. This compares rather unfavorably to the probability that an agent finds a bacterial fragment by randomly choosing one of the food items presented to it at each time step (which, in the simulations we will present here, is 1/20). This indicates that at early stages of the simulation, when DNA sequences still resemble random sequences, agents cannot effectively distinguish alien from bacterial fragments.

There are two ways in which the agents can improve their ability to recognize bacterial fragments. First of all, they can reduce the length f of their reference sequence. This however is a double-edged sword: If the reference length becomes too short, then the number of incidents where alien fragments are falsely recognized as bacterial will increase; false positive recognitions are detrimental (but only slightly so) in that they effectively decrease the number of fragments an agent gets to see.

The probability that an alien fragment (which is a random sequence of {a,c,g,t}) will contain the reference sequence can be obtained by a sliding window method. Consider the first u letters of the fragment. The probability that they are not identical to the reference sequence is given by $1 - (1/4)^u$; this is also the probability that the second window covering the second letter of the fragment to the $u + 1$st letter does not contain the reference sequence. We can thus write down the probability that the reference sequence contains at least one copy of the reference sequence:

$$P_2 := P(\text{alien fragment contains reference}) = 1 - \left(1 - \left(\frac{1}{4}\right)^u\right)^{f-u+1} \tag{2}$$

In order to get an estimate of the probability of a false positive recognition, this probability has to be multiplied by the number of alien food items an agent is presented with per time step. Assuming the above values for L and f, we obtain $19P_2 \approx 0.1$ (taking a reference length of $u = 7$). The corresponding probability of correctly recognizing a bacterial DNA is, according to Equation 1, approximately 0.01.

This shows that lowering the reference length, at least by itself, will not lead to better discrimination between bacterial and agent fragments. Note, however, that the above probabilities are only valid under the assumption that both the alien and the bacterial fragments have the statistical properties of random sequences. Mutation (and selection) will render this assumption invalid in the course of the simulation. The main mechanism responsible for this is copy mutations.

3.2 Copy Mutations

One possibly successful strategy for the agents is to distribute copies of not too long subsequences throughout the DNA and to simultaneously have GENOME pointing to one of them (i.e., the reference sequence should be equal to this repeated sequence). The probability of finding a specific repeated subsequence in a bacterial fragment is of course proportional to the number of repetitions of this subsequence. More specifically: If a sequence of length u is repeated on the DNA n times, then the probability for a bacterial fragment taken from this DNA to contain the signal at least once is again given by Equation 1 with $n > 1$. Setting $L = 10{,}000$, $f = 150$, and $u = 7$, we see that for about 12 repetitions we will for the first time get a probability of positive correct recognition of bacterial DNA slightly higher than 0.1, that is, higher than the probability for false positive recognitions. An even higher n will then further increase the probability of correct recognitions. This result suggests that the signal must be present in more than 11 copies to be effective. Below this threshold the signal will be of very little adaptive use for the agent; there will thus also be very little adaptive pressure to increase the number of repetitions of the signal as long as the number of repetitions is low. As pointed out above, creating an initial fitness differentiation in the population is thus a major problem.

Once there is an efficient signaling system in place, a further increase of the density of signals on the DNA will continue to have positive effects on agent fitness. Agents will then find themselves under continued pressure to improve their fitness. At least in the absence of counteracting forces, we thus have to expect that the number of copies of the signal will continue to grow until eventually all of the DNA consists of signals; at this stage agents will recognize bacterial DNA with a probability of (nearly) 1. This is possible only if the DNA is a string of identical letters (for example, . . . aaaaa . . .). Homogeneous DNA thus corresponds to a global fitness maximum. In real organisms this point will of course never be reached, as conflicting fitness constraints will become dominant long before a homogeneous state can be reached.

A factor that counteracts the proliferation of a signal on the DNA is point mutations. It is enough for a single point mutation to take place within a copy of the signal sequence, to destroy this particular copy; the probability of such an event can easily be seen to be $P(\text{destroy}) = nu/L$ and is therefore directly proportional to the number of sequences on the DNA.

Somewhat counterintuitively, it should also be noted that copy mutations also tend to reduce the number of copies of the reference sequence (if only slightly so): The reason for this is that the creation of a new copy requires that the *complete* signal be contained in the source of the copy mutation. At the same time a copy can be destroyed if the target area of a copy mutation contains only one letter of the signal. We can obtain the probability that the source of the copy mutation contains an entire signal by considering the probability that the last $c - u + 1$ letters of the source contain the last letter of a copy of the reference:

$$P_3 := P(\text{source contains at least 1 entire signal}) = 1 - \left(1 - \frac{n}{L}\right)^{c-u+1} \tag{3}$$

where c is the length of the copy mutation. We can also write down the probability that the target area of a copy mutation contains at least one letter of a signal:

$$P_4 := P(\text{target contains at least 1 letter signal}) = 1 - \left(1 - \frac{n}{L}\right)^{c-u+1} \tag{4}$$

Given this, we can now calculate the probabilities that the source and target areas do or do not contain copies of the reference sequence. Of real interest in the current case are the probability that the source contains a copy of the reference sequence but the target does not (creation of a new copy) and the probability that the target contains a copy but the source does not (destruction of a copy). It can be easily seen from the definition of the probabilities that the probability of the destruction of a copy is greater than the probability of the creation of a copy; Figure 3 illustrates this by using a numerical example.

The observations of this sections can be summarized as follows:

- At the outset (i.e., with an unbiased DNA), the agent has no effective means of distinguishing between bacterial and alien DNA.
- Agents have to decrease the size of their reference sequence by mutation of the GENOME.
- Agents have to distribute copies of a subsequence of their DNA throughout the string.
- Agents have to surmount a threshold before the signal is effective, but once they have reached this point, an evolutionary runaway effect will tend to produce ever fitter agents.

4 Results

In this section we will present results obtained from simulations of the model. In Section 4.1 we discuss in some detail three example runs. This will illustrate the main features of the model and provide the reader with an intuition about the behavior of the model and its main features. In Section 4.2 we will summarize results obtained from many simulations. This will illustrate the stability of the qualitative behavior of the model with respect to changes of some crucial parameters.

4.1 Three Example Runs

In this section we will present results from three typical single simulation runs of the model. These results mainly serve as examples to illustrate a typical behavior of the model. We found that the

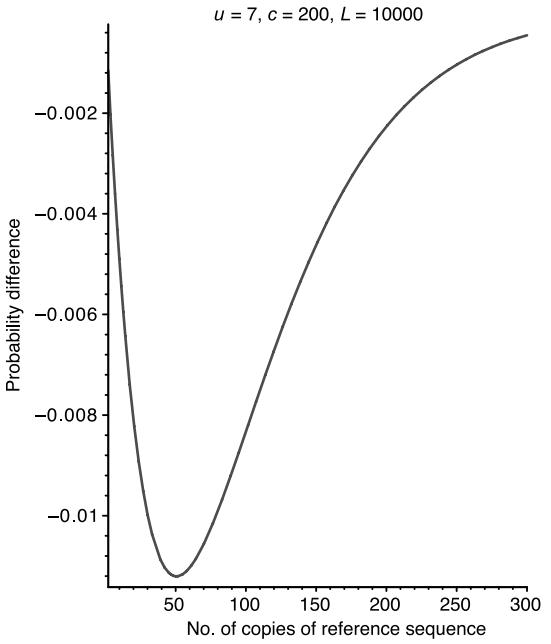

Figure 3. This graph shows the difference between the probability for a copy mutation to create a new copy of the reference sequence and to destroy an existing one, $P_3(1 - P_4) - P_4(1 - P_3)$; the chosen values are $L = 10{,}000$, $c = 200$, and $u = 7$. We see that irrespective of the number of copies of the USS on the DNA, the probability that a *specific* USS is destroyed is always slightly higher than the probability that a new copy of this specific subsequence is created. Note, however, that this is only true for every specific sequence. Nearly every application of the copy mutation will create a new copy of the source of the copy mutation.

behavior of the model is best understood by looking at the behavior of the following variables over time:

- *Length of the reference:* The length of all the references in the population averaged over the population.
- *Repetitions of the reference:* The number of repetitions of the reference sequence of an agent on its DNA, averaged over the population.
- *Entropy:* The entropy of the DNA, averaged over the population.
- *Signal entropy:* The entropy of the concatenation of all references in the population.
- *Accepted food:* Whenever an agent accepts a bacterial (alien) item as a consequence of a comparison of the fragment with the reference, the counter for positive (false positive) recognitions is increased by one. If *CHOOSE* = *false* or if the agent accepted the last food item because it failed to find a match with its reference, then these counters are not increased.

Generally we can assume that an efficient signal is emerging if the average length of the reference suddenly falls to a value not too long and not too short (as discussed earlier) and if the repeats of the reference on the DNA and the rate for correct recognition of bacterial DNA sharply increase; the latter indicates that the population is capable of selectively taking up bacterial fragments.

4.1.1 Run 1

In the first simulation run we present here (see Figure 4), the energy gain from alien (bacterial) fragments is 1 (2); all other parameters are listed in Table 1. Note that point mutations are turned off.

In this and all subsequent reported runs, we updated the model one million times before aborting the simulation.

If bacterial and alien fragments yield the same amount of energy, then there will be no adaptive pressure in the model. Under these circumstances we found the model to behave qualitatively like the neutral model (i.e., bacterial and alien fragments yield equal energy; data not shown). The average length of the reference sequence performs a random walk around some mean value. The observed steady decline of the entropy is a result of the continuous operation of the copy mutations in the absence of the counteracting force of random point mutations and does not indicate adaptation by the agents. Rather discouraging is the record of the recognized food particles: Most of the recognized particles are in fact false positive recognitions, that is, alien DNA fragments that by coincidence contained the reference sequence of the inspecting agent; even those false positive recognitions happen on average only about once in 500 time steps per agent; the rate for correct identifications is even lower, with an average of about 5000 time steps between correct recognitions by an agent. Understanding that this is less than once every 500 generations, we conclude that there are no indications of an emerged uptake signal.

It seems that at early stages when the quality of the signals is still low, the payoff for bacterial DNA is not sufficient to confer enough fitness advantage to agents to initiate evolutionary runaway.

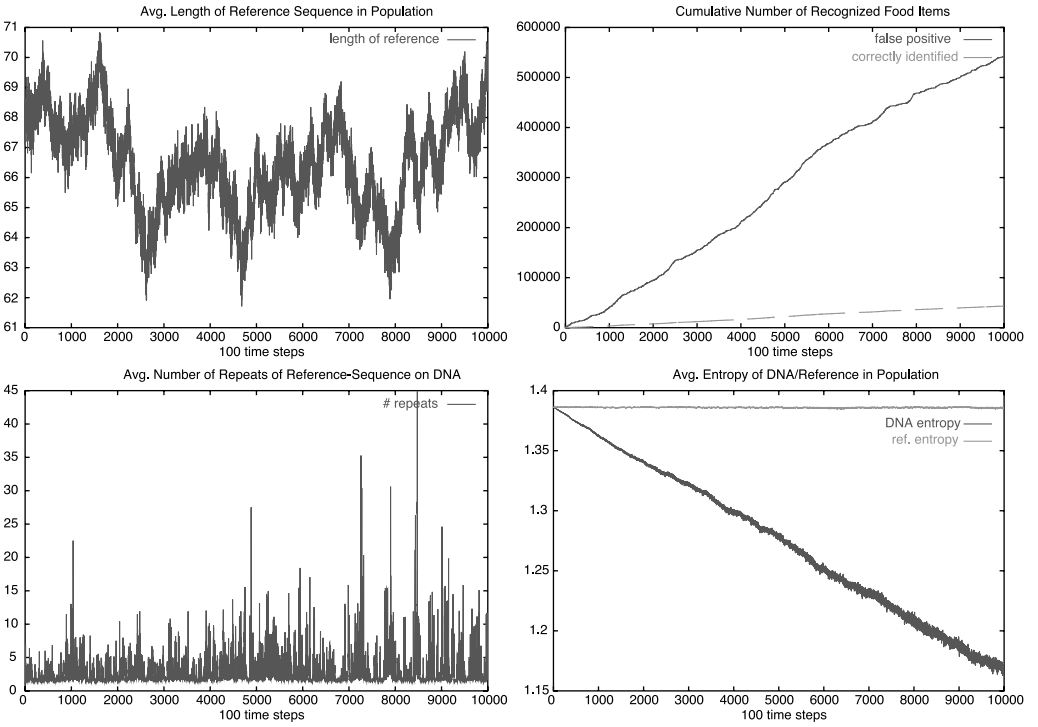

Figure 4. Run I. Top left: Time series of the length of the reference averaged over the population. One measurement is taken every hundred time steps. This variable clearly performs a random walk. Top right: Count of the number of bacterial (dashed) and alien (solid) fragments as a result of inspection by the agents. On average an agent has to wait 5,000 time steps for a correct identification and about 500 time steps for a false positive identification. Bottom left: The number of repeats of the reference sequence on the DNA. The maximum value this variable can take is equal to the size of the DNA. If that were the case, then the agent would have reached maximal fitness. In the present case however, there does not seem to be a consistent trend toward higher repeats. Bottom right: The entropy ($-\sum i \in \Gamma p_i \log p_i$, $\Gamma = \{a, c, g, t\}$) of the DNA averaged over the whole population (dark gray) and the entropy of the concatenation of all references (light gray). The latter is consistently higher because of the diversity of the population: As the falling entropy suggests, the DNA sequences of the agents are skewed with respect to the base composition but not in the same way; thus in sum the concatenation of the references has no bias.

Table 1. The values of the parameters used in the simulation of Section 4.1.

DNA length	10,000
World size	30
Max. population size	300
Mutation rate	0.9
Point-mutation rate	0
Max. no. of fragments presented to agents	20
Size of fragments	100
Min. energy to reproduce	6
Max. lifetime	10
Payoff for alien fragments	1
Max. no. of bacteria per site	200

4.1.2 Run II

This run is identical to Run I except that bacterial fragments now yield 3 energy units; thus agents experience a stronger adaptive pressure, that is, good agents are relatively better than when the bacterial payoff was 2. This is sufficient to alter the evolutionary dynamics drastically. The diagrams in Figure 5 clearly show a discontinuity between time steps 500,000 and 600,000: Approximately simultaneously we observe a sharp drop of the average reference length to below 10, an increase of the rate with which fragments are recognized (both false positive and correct), and an increase of the average number of repetitions of the reference sequence on the DNA. All this indicates that an efficient uptake signal has indeed emerged.

A short calculation shows just how effective the signal recognition of the agents is: If we take the reference to be of length 7, and assume it to be repeated 600 times on the DNA (those values roughly correspond to the ones shown in the figures), then 4,200 of the 10,000 letters of the DNA will be contained in copies of the signal. If evenly spaced, then there will be about 97 letters between successive copies of the sequence. The probability that a random fragment (which is of length 100) contains the reference will then be greater than 0.9. This shows that an agent that is confronted with a bacterial fragment that is taken from one of its recent ancestors has a very good chance to actually find an uptake signal on it. It is sufficient that a bacterial fragment yield 3 energy units in order for the agents to evolve a USS.

4.1.3 Run III

This run is the same as the previous two in all respects except for the number of energy units for bacterial DNA fragments, which is now set to 18. Three different regimes of system behavior can be distinguished (see Figure 6). The first regime is the initial phase where the agents have not yet recognized or imprinted a signal. This phase is extremely short and barely visible on the graph. It is followed by a regime characterized by short reference sequences and high levels of repetition of the reference sequence. Similarly to Run II, it can be argued that an efficient uptake signal is present in the population. Continued adaptive pressure to even higher recognition rates of the population

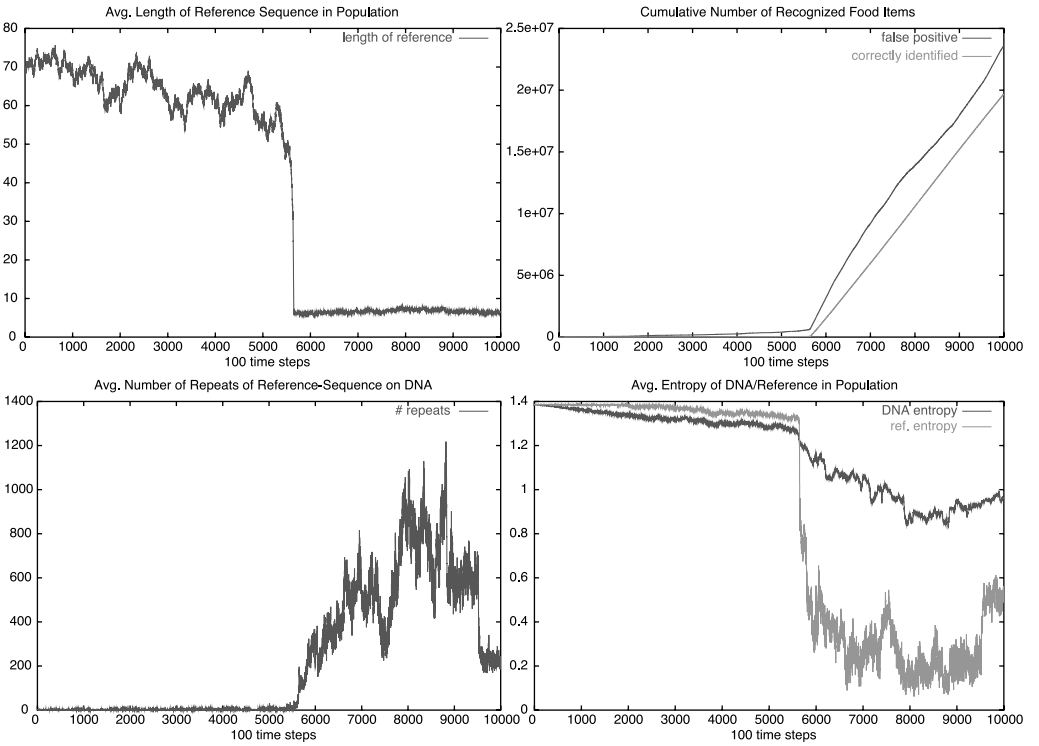

Figure 5. Run II; cf. Figure 4. We observe a clear discontinuity between time steps 500,000 and 600,000. From then on, the agents have a shorter reference sequence and many copies of the reference on their DNA. In consequence they are dramatically more efficient in recognizing bacterial fragments. The decrease of the entropy of the concatenation of the reference sequences indicates that the population has been homogenized by selection, so that agents are very similar to each other. Note that the rate of false positive recognitions is in fact higher than that of correct recognitions.

drives the model to the final phase, where the DNA of the agents has an entropy of zero. This corresponds to a completely homogeneous DNA. The signal is now repeated nearly 10,000 times on average. Since the concatenation of all reference sequences also has zero entropy, we see that the population is homogeneous as well (i.e., all agents have the same DNA).

4.2 Stability of the Qualitative Behavior

The results of the three illustrative runs in the previous section are obtained using a specific set of parameters; such simulations are not necessarily indicative of the typical behavior of the model under arbitrary conditions and certainly do not allow conclusions about the robustness of the observed phenomena with respect to changes of the parameters. It is thus necessary to test the behavior of the model under various conditions.

A DNA length of 10,000 is short compared to that of real DNA. One might thus suspect that the emergence of the USS depends on the use of such relatively short DNA sequences. We have performed a series of simulations in order to test the dependence of the emergence of USS on the DNA length. In the following simulations we kept all parameters constant except for the DNA length and the size of the DNA fragment (the piece of DNA to be taken up); the key parameters of the simulations can be found in Table 2. In this section we ignore the details of the simulation results and are only interested in whether or not a USS emerged in a simulation. In a specific run we decided that a USS had emerged if the average reference length suddenly dropped to a short value and if the average number of repetitions of the reference length on the DNA increased simultaneously.

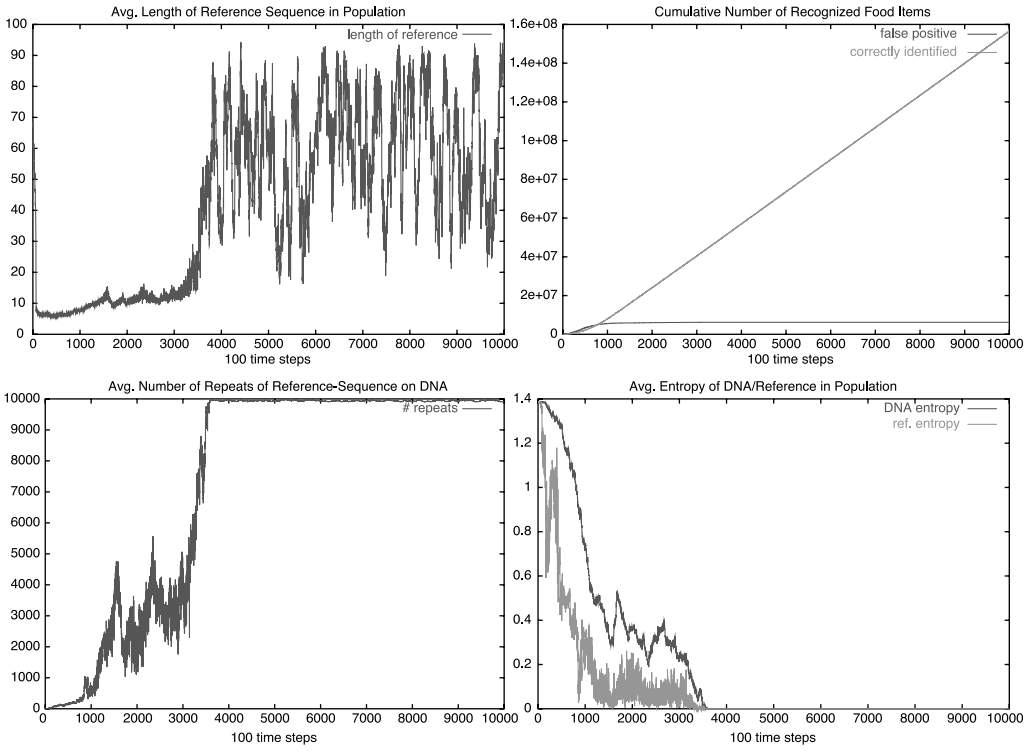

Figure 6. Run III; cf. Figure 4. The high relative benefit from bacterial fragments leads to the rapid initial development of a signaling system. After that, the population increases its fitness by having more and more repetitions of the signal on the DNA. This process leads to an area of maximal fitness characterized by a completely homogeneous DNA. The reference length is not fitness-relevant any more in these areas and can thus resume high values. Recognition of bacterial fragments after time step 400,000 happens with probability close to one, and there are no more false positive recognitions.

We tested DNA lengths of 30,000, 90,000, and 150,000 time steps (henceforth referred to as 30K, 90K, and 150K, respectively). Furthermore, for every DNA length we also performed separate series of simulations for various fragment sizes. We chose the following values: 50, 100, 250, 500, 750, 1000, and 1500. For each setting of the parameters, we performed 60 simulations; thus the whole series consists of 1080 simulations. In addition we performed another 60 simulations with a DNA of length 150K, but with the length of the reference string restricted to a maximum of 500 (or the length of the fragment; whichever was shorter); we will henceforth refer to this series as 150Kl. Thus the difference between the 150Kl and the 150K series is that in the former the reference lengths are always equal to or smaller than 500, whereas in the latter the upper bound for this value is always equal to the fragment length. Figure 7 summarizes the results. The figure clearly demonstrates a strong dependence of the emergence of USSs on the length of the DNA and the fragment. In particular, for fragment sizes below 250 the USS emerges in the shortest DNA (30K) only. Increasing the fragment length to 500, we see nearly 25 occurrences in 60 runs for the 30K DNA, but very small numbers for longer DNA. Using a 90K DNA we get a somewhat larger number of emerged USSs for longer DNA fragments, but never more than 12. Likewise, we never get more than 5 emergences for the longest DNA (150K).

Very different behavior is shown by the 150Kl series. This series of simulations is identical to the 150K series, but the DNA fragment has a maximum length of 500, whereas the maximum length in all other sequences is determined by the fragment size. Limiting the length of the reference sequence in such a way leads to dramatic improvements of the rate of the emergence of USSs. Figure 7 shows that with a fragment size of 1500 the 150Kl series shows more emerged USSs (29) than any other

Table 2. Overview of the most important parameters.

World Size	30
Max. population size	300
Mutation rate	0.1
Point-mutation rate	0
Max. no. of fragments presented to agents	20
Min. energy to reproduce	60
Max. lifetime	10
Lifetime limit	20
Max. no. of copy mutations	200
Payoff for alien fragments	10
Max. no. of bacteria per site	200

series. In order to test the statistical significance of this difference, we performed a binomial logistic regression analysis on the data (150K and 150Kl). The fitted model shows that the success rate in the 150Kl series is 6.09 times greater on average than that of the 150K series. With a standard error of 0.3329, this gives a highly significant result on a z test. In turn, this tells us that the 150Kl series gives a significantly higher proportion of successes than the 150K condition (beyond a P value of 0.001).

There are two reservations about this data. Firstly, if the model (with the parameter settings of the 150Kl series) is seeded with only one initial agent (so all agents have only one ancestor), then the number of emerged USSs is somewhat reduced. We again performed 60 runs with only one initial

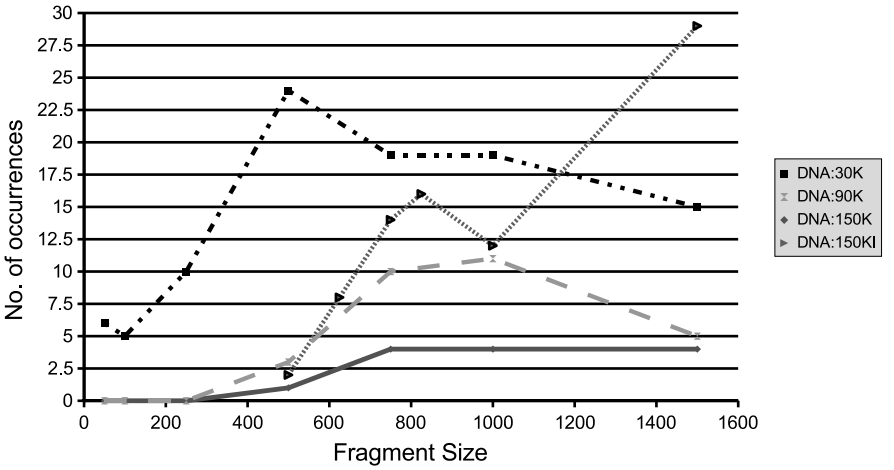

Figure 7. The number of emerged USSs in 60 runs for various DNA lengths and reference sizes. The label "DNA:30K" stands for a DNA of length 30 thousand. The 150Kl series is like the 150K series, but the reference length is limited to 500 (and not to the fragment size).

agent and found that a USS had emerged in 15 of those simulations (data not shown). Secondly, the success rate seems to depend on the absolute carrying capacity. If we set the carrying capacity to 1,000 (and seed the model with 1,000) agents and the number of sites in the world to 1 (so all agents are in one site and share the same pool of bacterial fragments), then no USS emerges. The emergence of USSs thus seems to depend on how many agents share a pool of bacterial DNA fragments (see discussion in Section 5).

In the above simulations the point-mutation rate was turned off. Point mutations are a counterforce to the spread of uptake signals in the DNA. We tested the influence of point mutations on the emergence of USSs by conducting 30 identical runs for five different point-mutation rates. All other parameters were equal, as in the 150Kl series.

The results (summarized in Table 3) indicate that up to a point-mutation rate of 0.5 the occurrence of USSs is not greatly reduced, but it falls dramatically for higher point-mutation rates. If 90% of all mutations were point mutations, the USS still emerged in two out of 30 runs. At least for the current parameters, no USS emerged if all mutations were point mutations. It seems therefore that copy mutations are necessary in order to produce the phenomenon.

Another question concerns the speed of uptake of the bacterial DNA, once the USS has emerged. A possible measure for this is the slope of the curve recording the cumulative number of successfully recognized bacterial fragments (for example, the top right graph in Figure 6). In order to estimate the slope, we performed a linear regression of this curve once a USS was established in the population (the cumulative uptake of the bacterial (alien) DNA versus time behaves nearly linearly); we then recorded all slopes obtained from one set of identical simulations and took the average. The data in Table 4 suggests that setting a lower limit of 9 for the reference sequence substantially decreases the number of false positive recognitions without, however, decreasing the correct recognitions. A further increase of the minimum reference length substantially decreases the false positive recognitions, but at the same time, the correct recognitions are reduced.

In the unrestricted case the graph of correct recognitions has a slope of about 2,000; since there are 300 agents in the system and each recorded data value corresponds to 100 times steps in the simulation, a slope of 2000 corresponds to an update frequency of $2000/100 \times 300 = 1/15$; that is, each agent takes up a bacterial fragment once every 15 time steps on average.

Finally, we attempted to find the minimum bacterial payoff that still leads to the emergence of a USS. In all the above results bacterial fragments yielded 3.7 times more payoff than alien food. In order to determine the minimum bacterial payoff required, we performed a series of 30 simulations for various values of the bacterial payoff. In order to save computational time, we performed the

Table 3. Sensitivity of the emergence of USSs to variation of the probability that a mutation of the DNA is a point mutation. The right column gives the number of simulations that showed the emergence of USS in 30 simulations (thus the numbers here need to be multiplied by two in order to be comparable to the results in Figure 7). Apart from the point-mutation rate, the parameter settings for those simulations were identical to the 150Kl series.

Point-mutation probability	No. of occurrences
0.3	13
0.5	12
0.7	10
0.8	6
0.9	2
1.0	0

Table 4. The slopes of the graphs showing the cumulative positive and false-positive recognitions of bacterial fragments. The values shown are the mean slopes of all runs with emergent USSs in a series of 30 simulations. In the first row we used the parameters of the 150KI series; in the second and third rows, the same parameters but with lower limits of 9 and 11 respectively set for the reference length. The values in parentheses are the standard deviations of the data sets. All entries are rounded to integers.

	Slope	
Limit	Correct	False pos.
None	2130 (159)	3540 (361)
9	2125 (274)	1577 (81)
11	1554 (202)	28 (3)

experiments on a DNA of length 10,000 and a fragment size of 150. We found that if alien fragments yield 10 energy units, then bacterial fragments need to yield at least 25 energy units in order for a USS to emerge.

5 Discussion

The example run in Section 4.1 (Figure 5) shows that when the USS emerges, agents suddenly start to take up bacterial fragments at a dramatically increased rate (see Figure 5). Apparently, though, not only the rate of correct recognitions increases, but also the rate of false positive recognitions; in fact, the rate of false positive recognitions is higher than the rate of correct recognitions. Given that agents do not collect negative payoff from false recognition in the current model, this is not surprising in itself, yet from the data it is unclear what the exact nature of this effect is. The fact that setting a minimum length of 9 for the reference sequence reduces the rate of false positive recognitions while keeping the rate of correct recognitions at about the same value (see Table 4) suggests that the high rate of false positive recognitions was largely due to a few (unfit) agents with very small fragments; because of their short reference sequence, those agents will make many false positive recognitions, but few correct ones. A further increase of the minimum reference length, however, leads to a deterioration of the rate of correct recognitions. This suggests that 9 is close to the ideal length for the reference sequence.

This result is interesting in that the length of the core uptake signal in real bacteria is also about 9 bp long. Furthermore, calculations [2] indicate that 9 is a good balance between minimizing the uptake of alien DNA and maximizing the probability of correct recognition of bacterial DNA fragments. However, the significance of this is not clear, as this tradeoff is likely to depend on the negative payoff received from false positive recognitions; in the present model we do not consider such a negative payoff. We therefore leave the investigation of this effect to future work.

The summarized simulation results in Figure 7 seem to indicate that the emergence of USSs strongly depends on the length of the DNA. This would be bad news for the preference-first scenario, because even the longest DNA in the current simulations is orders of magnitude shorter than real bacterial DNA. It seems that increasing the size of the fragments somewhat dampens this tendency. The data also indicates that an increase of the fragment size beyond a certain limit actually makes it more difficult for a USS to emerge.

An explanation of this effect has to do with the way in which the reference length is determined: The allowed reference lengths of the agents are constrained to be between 0 and the fragment size. Since the length of the reference sequence in the seed population is determined randomly, increasing the length of the fragment will also lead to agents having initially longer reference sequences. In

Section 3 we noted that all genomic configurations with long reference sequences can be considered as giving approximately equal fitness, that is, decreasing or increasing a long reference sequence by just a little bit will not have a considerable effect on the fitness of the agents. Thus, at early stages of the simulation, agents perform a random walk in genomic space (for an extensive discussion of the fitness landscape in the USS model see [2]), and there will be no fitness differentiation between them.

To make things worse, as long as the reference sequence is long (that is, >12 or so, see [2]), the probability that there will be a copy of the reference sequence on a bacterial fragment (see Equation 1) and also a copy mutation will increase the number of copies of the reference sequence on the DNA will be very low (this can be easily seen from Equation 3). Before an effective adaptive pressure can drive the evolution of USSs, agents will have to decrease the length of the reference sequence. In the absence of any adaptive pressure, this is done by a random walk. The time required for random walkers to reach the area in which effective adaptive pressure is exerted grows quadratically with the initial distance to this area (one-dimensional random walk). Realistically, a USS can thus emerge only if the initial reference length is sufficiently short; the probability for a sufficiently short reference sequence in turn depends on the fragment size. In this perspective, thus, smaller fragment sizes have a positive effect on the emergence of USSs. On the other hand, if the fragment is too small, then again no USS will emerge, because then the probability for a fragment to contain a copy of the reference sequence is too low.

If the initial reference lengths are decoupled from the fragment sizes, then these conflicting requirements are reconciled. This interpretation is corroborated by the above simulations: While in the 150K series the maximum allowed length of the reference sequence is equal to the actual length, in the 150Kl series the maximum allowed length is set to 500; this increases the likelihood of initially short reference sequences. As expected from the arguments in the previous paragraph, this increases the number of instances where a USS emerges.

While we do not simulate DNA sequences longer than 150K, we conjecture that a further increase of the length of the DNA will not negatively affect the emergence of USSs if it can be offset by a corresponding increase of the fragment length (while keeping the maximum reference length fixed). The general problem with increasing the DNA length is that this reduces the probability of finding a copy of the signal on a fragment taken from the DNA (assuming n remains constant; see Equation 1); this effect can be offset by a corresponding increase of the fragment size, which will make it much more probable for a given fragment to contain one of the n copies of the reference on the DNA. If we increase the length of the DNA to L' and fragment length to f', then according to a first order approximation of Equation 1 we will find one of the n copies of the signal with the same probability as on the original sequence of length L and fragment size f if we keep the ratios between the DNA length and the fragment size constant (assuming n remains constant):

$$\frac{L'}{f' - u} = \frac{L}{f - u}$$

Thus we conclude that it is the length of the DNA in relation to the fragment size that is important, not the absolute length of the DNA.

We will now discuss the importance of the absolute number of agents sharing a pool of bacterial DNA fragments. All simulations reported in this article were performed with a maximum population size of 300. If the maximal carrying capacity is substantially increased (to 1,000), then we found that no USS will emerge. This effect can be readily understood from the internal construction of the model. The assumption of the model is that most of the DNA fragments in the agents' environment are alien; this is represented in the model through the low probability of an agent being presented with a bacterial fragment: On average an agent sees only one bacterial fragment per time step. Ultimately, because of this low probability, a USS can only emerge and have adaptive value if the bacterial fragments an agent will be presented with are, at least with reasonable probability, taken from a DNA sequence that is very similar to the agent's own. If the world contains a large number of

agents, then the population of living agents and thus also the pool of DNA fragments in the environment will be very heterogeneous. In this case, most bacterial fragments an agent will be presented with will then be very dissimilar to its own DNA. Thus, agents have no effective way to gain sustained advantage from a USS, and thus none will emerge.

An important question is whether this sensitivity to large population sizes is relevant to the emergence of USSs in real bacteria. Typical population sizes in the real world will be larger than 300. However, the population size in the model does not translate 1 : 1 into population size in the real world. What is important in the above argument is not so much the absolute population size in terms of individual bacteria, but the concentration of conspecific fragments in the environment of bacteria. This in turn depends on the concentration of conspecific bacteria in the environment, not on their absolute numbers.

Note that the required density of bacterial food in the environment need not be very high: In the simulations at every time step (on average) agents are presented with 19 alien fragments and one bacterial fragment. Even in the case of 300 agents, many of those bacterial fragments might be taken from very dissimilar agents (particularly at early stages of the simulation) and thus essentially look random as well. This effect is somewhat dampened by the compartmentalization of the world in which the agents live; however, our experiments (data not given here) show that even if all agents share the same compartment, the number of experiments where the USS emerges is only slightly reduced. Altogether, we thus conclude that a carrying capacity of 300 in the model is not a particularly strong assumption, as it implies a very low probability for an agent to actually see a fragment that is taken from one of its ancestors. Future research will need to explain in detail how realistic this assumption is with respect to real bacteria.

Most experiments reported in the results section were initialized with 300 initial agents, thus providing a high initial diversity for evolution to act on. As expected, if the world is seeded with only one initial agent, the overall number of emerging USSs is smaller. The reason for the smaller number, however, is not the lack of diversity of DNA sequences, but the lack of diversity of GENOMEs. The fitness of an agent, that is, its ability to distinguish between bacterial and alien DNA, does not depend on the specific DNA sequence, but only on the number of repeats of the reference sequence on the DNA; this is also reflected in the fact that Equations 1–3 do not depend on the details of the DNA sequence. The probability of the emergence of a USS in the case of only one initial agent is lower mainly because then there is no spread in the length of the initial reference sequences. If the initial agent has a very long reference length, then it is very unlikely that an agent will sufficiently lower its reference length by a random walk in genetic space. Thus the emergence of a USS will depend on the initial agent having a sufficiently short reference length. Experimental results (data not shown) verified this interpretation.

Building upon this insight into the importance of the reference length for the emergence of USSs on bacterial DNA, we also conjecture that there are two effects that contribute to the reduction of the instances of the emergence of a USS when point mutations are turned on (see Table 4). One effect certainly is that point mutations counteract the spread of copies of the reference sequence on the DNA. The second effect is that increasing the number of point mutations will decrease the number of copy mutations and, most importantly, the number of mutations of the reference length. Consequently, random walkers will move slower, and it will take longer to reach a short reference length from a given initial state. Thus, altogether the agents will need shorter initial reference sequences in order to have a chance of evolving a USS within the simulated time.

6 Conclusion and Outlook

The present model is highly simplified and abstracts away most features of real bacteria, but allows in return a clear understanding of the various factors contributing to the emergence of the USS. The model can probably not be used to answer all questions about the evolution of USSs. Its main benefit is that it helps to sharpen our intuition about the important parameters and conditions that

favor the evolution of USSs in a preference-first scenario. Further work is necessary to develop insights gained from the model into a better understanding of the evolution of USSs in real bacteria and to derive testable hypotheses from it.

Can USSs emerge in a preference-first scenario? At least under the assumptions of the model, they can. Real organisms operate of course on a much more complicated (time-dependent) fitness landscape than the simulated agents in the present model. Many important issues, such as the cost of having a USS and other conflicting adaptive pressures, have been neglected here. In order to be able to make solid predictions, it will be necessary to consider the influence of those factors in future contributions.

Despite those reservations, our results have clearly indicated that the emergence of USSs is possible as long as uptake of conspecific DNA fragments is moderately more beneficial than the uptake of non-conspecific fragments. Most of all, nowhere in the model did we explicitly implement any group-selection mechanisms. We thus conclude that a USS can emerge in the preference-first scenario without assuming group selection.

Acknowledgments

We thank Jon Rowe for discussions and comments on the draft, Allan White for statistical advice, and the Norwegian High Performance Consortium (NOTUR) for generous allocation of CPU time. D.C. is particularly grateful for financial support received from the Paul & Yuanbi Ramsay Research Fund at the School of Computer Science, University of Birmingham.

References

1. Bakkali, M., Chen, T., Lee, H., & Redfield, R. (2004). Evolutionary stability of DNA uptake signal sequences in the *Pasteurelleceae*. *Proceedings of the National Academy of Science*, *101*(13), 4513–4518.
2. Chu, D. and Rowe, J. (2005). A fitness landscape for the evolution of uptake signal sequences on bacterial DNA. 9th European Conference on Artificial Life (submitted).
3. Hamilton, W. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*(1), 1–16.
4. Hsieh, L., Luo, L., Ji, F., & Lee, H. (2003). Minimal model for genome evolution and growth. *Physical Review Letters*, *90*(5), 101–104.
5. Johnson, T., and Redfield, R. (2003). The molecular evolution of DNA uptake signal sequences (in preparation).
6. Karlin, S., Mrazek, J., & Campell, M. (1996). Frequent oligonucleotides and peptides in the *Haemophilus influenzae* genome. *Nucleic Acids Research*, *24*, 4263–4272.
7. Lawrence, J., & Ochman, H. (1998). Molecular archeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the U.S.A.*, *95*, 9413–9417.
8. Lorenz, M., & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiology and Molecular Biology Reviews*, *58*(3), 563–602.
9. Lynch, M. (2002). Gene duplication and evolution. *Science*, *297*, 945–947.
10. Macfadyen, L., Chen, D., Vo, H., Liao, D., Sinotte, R., & Redfield, R. (2001). Competence development by *Haemophilus influenzae* is regulated by the availability of nucleic acid precursors. *Molecular Microbiology*, *40*, 700–707.
11. Redfield, R. (1988). Evolution of bacterial transformation: Is sex with dead cells ever better than no sex at all? *Genetics*, *119*, 213–221.
12. Redfield, R. (1993). Genes for breakfast: The have-your-cake-and-eat-it-too of bacterial transformation. *Journal of Heredity*, *84*(5), 400–404.
13. Redfield, R. (2001a). NIH proposal to study the evolution and function of uptake signal sequences in naturally competent bacteria, available at http://www.zoology.ubc.ca/redfield/research/USS/USSprpsl.pdf.

14. Redfield, R. (2001b). Do bacteria have sex? *Nature Review Genetics*, *2*, 634–639.

15. Smith, H., Gwinn, M., & Salzberg, S. (1999). DNA uptake signal sequences in naturally transformable bacteria. *Research in Microbiology*, *150*, 603–616.

16. Smith, H., Tomb, J., Dougherty, B., Fleischmann, R., & Venter, J. (1995). Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science*, *269*, 538–540.

17. Solomon, J., & Grossman, A. (2000). Who's competent when: Regulation of natural genetic competence in bacteria. *Proceedings of the National Academy of Sciences of the U.S.A.*, *97*, 6981–6985.