Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 238 (2006) 157–166
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Evaluation of the current models for the evolution of bacterial DNA uptake signal sequences

Dominique Chu[a,*], Jonathan Rowe[b], Hoong-Chien Lee[c]

[a]Senter for Vitskapsteori, Universitetet i Bergen, 5020 Bergen, Norway
[b]School of Computer Science, University of Birmingham, B15 2TT, UK
[c]Department of Physics, National Central University, Jungli, Taiwan 320, ROC

Received 1 February 2005; received in revised form 4 May 2005; accepted 5 May 2005
Available online 14 July 2005

Abstract

Current opinion considers two main hypotheses for the evolutionary origin of uptake signal sequences in bacteria: one model regards the uptake signal sequence (USS) as the result of biased gene conversion, whereas the second model views the USS as a molecular tag that evolved as an adaptation. In this article, we present various computational models that implement specific versions of those hypotheses. Those models show that the two hypothesis are not necessarily as opposed to each other as may appear at first glance.

Keywords: Uptake signal sequence; Agent-based modelling; Natural competence

1. Introduction

Natural competence (Solomon and Grossman, 1996, 2000; Lorenz and Wackernagel, 1994; Macfadyen et al., 2001; Wang et al., 2002) is a genetically controlled form of horizontal gene transfer in some bacterial species. Given certain environmental conditions bacteria will take up free DNA fragments from (mainly dead) cells in the environment. Some bacterial species have a strong preference for DNA fragments that contain a specific sequence—the *uptake signal sequence* (USS). It has been found that in those species the USS is also highly over-represented on the respective genomes. Table 1 summarizes the USS core-sequences of some bacteria.

There have been thorough investigations of the statistical properties of USS, particularly in *Haemophilus influenzae* (HI); an exhaustive review of the literature (Smith et al., 1995, 1999; Karlin et al., 1996) would go beyond the scope of this article. Here, we only mention that the USS is highly over-represented. On a random string with the same base-composition and length as HI one would only expect about 8 copies of the USS, whereas the actual number is more than 1400. As Table 1 shows the USS and its reverse complement are approximately equally abundant on the genome; furthermore copies of the USS are remarkably equally spaced on the genome; in fact more equally than one would expect from random placement.

There have been various suggestions as to the function and the evolutionary origin of USS. For example, the USS has been proposed as a transcription terminator. It has also been suggested that the USS functions as a Chi sequence. However, more detailed investigations could not substantiate those conjectures (Bakkali et al., 2004). Currently, researchers consider two main hypotheses about the origin of USS (Redfield, 1988, 1993, 2001; Bakkali et al., 2004).

In one suggested scenario the USS is assumed to be a mate recognition signal. The basic idea of this *preference*

*Corresponding author. Tel.: +47 555 82978.
E-mail addresses: dominique.chu@svt.uib.no (D. Chu), j.e.rowe@cs.bham.ac.uk (J. Rowe), hclee@phy.ncu.edu.tw (H.-C. Lee).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.05.024

Table 1
The abundance of the USS sequence (and its reverse complement) in various species

Sequence	USS
Haemophilus influenzae	
AAGTGCGGT	737
ACCGCACTT	734
Streptococcus pneumoniae R6	
AAAATCAAA	182
TTTGATTTT	159
Pasteurella multocida	
AAGTGCGGT	468
ACCGCACTT	459
Neisseria meningiditis Z2491	
GCCGTCTGAA	958
TTCAGACGGC	934
Synechocystis sp. PCC 6803	
GGCGATCGCC	2823
GCGATCGCCA	1051

first hypothesis (PFH) is that uptake of conspecific DNA fragments is more beneficial than uptake of alien fragments. The USS evolved as a molecular signal or tag that allows bacteria to distinguish between conspecific and other fragments.

Somewhat opposed to this is the *molecular drive hypothesis* (MDH). According to this second main hypothesis, the USS serves no adaptive function as in the PFH, but is rather the result of a biased uptake mechanism. The idea (Bakkali et al., 2004) is that the receptor that binds to external DNA fragments has a strong preference for a specific sequence *s*. Hence, mutations producing DNA sequences that are similar to *s* will be fixed, thus creating a molecular drive towards *s*. Horizontal DNA transfer between bacterial cells allows the spreading of *s* in susceptible regions of the genome (i.e., non-functional and less-conserved genomic regions), while selection eliminates *s* from the functionally constrained regions. DNA uptake bias towards *s* combined with overrepresentation of *s* in the genome render this sequence a USS.

Researchers have investigated the evolutionary origin and current function of the USS by means of in vitro experimentation and extensive analysis of sequence data (Finkel and Kolter, 2001; Tomb et al., 1996). So far no definitive conclusion has been reached. The present contribution will add to previously employed methods by using computational *what-if* experiments to investigate the evolutionary dynamics of USS uptake. Specifically, this article has two main goals. Firstly to pin down *concrete* scenarios compatible with either the MDH or the PFH and secondly to investigate functional relationships between the parameters in those scenarios.

We do not claim that this article will close the book on the question about the evolutionary origin of the USS nor that the chosen level of abstraction of our computational experiments takes into account all biologically relevant processes. However, given finite computational resources, a situation of diminishing returns and the explosion of the parameter space we decided that the choice of models we present here is a useful and feasible selection.

2. Problems of the PFH

In this section, we will present objections against the PFH and some recent computational results investigating the evolution of USS in individual-based models.

2.1. Emergence of the USS under the PFH

The first objection against the PFH is that there is no biologically plausible evolutionary route leading to it. Bakkali et al. (2004), for example, suggest that in order to explain the origin of the USS under the PFH one would also need to assume biologically unrealistic group-selection mechanisms. The argument goes as follows. Before a USS sequence has become efficient as a molecular tag, it is not beneficial for a cell to be the first one to carry a USS. This is so because the benefit from such a USS only materializes after the death of the cell when its DNA can be taken up by others. Benefit from a USS can thus only occur at the group level, but not at the level of the individual cell. As such, the evolution of USS requires the assumption of group selection mechanisms, which in turn is not biologically realistic.

Closely connected to this is another problem of the evolutionary origin of the USS. A functioning USS-signal requires two components, a signal-sequence and a recognition system. In the present context the recognition system is the receptor that selects DNA fragments from the environment. The efficiency of a signal sequence mainly depends on two factors. The sequence is more efficient as a signal if it is present in more copies and if those copies are more equally spaced. Moreover, a condition sine qua non for the efficiency of a sequence as a signal is that it is identical to the receptor sequence.

The problem arises when one considers that during early stages of the evolution of the USS, the candidate signal sequence is present in a low number of copies, say N only. If N is low, then so is $N+1$; the improvement in going from N to $N+1$ will typically be very small and

not lead to a significant increase in fitness. Hence, as long as there are only few copies of the signal sequence on the DNA there will be little evolutionary drive to increase the number of copies or to correct point mutations of the receptor sequence. Therefore, as long as the efficiency of a signal is weak, so will be the evolutionary forces conserving (or indeed improving) it. Hence no USS will evolve.

There are at least two ways around this firstly, the genome might already contain highly repeated subsequences; one of those might then be used as a signal sequence, thus avoiding the difficulty of having to evolve one. Recent research has established that typical genomes do indeed contain highly over-represented oligomeres (Hsieh et al., 2003).

Another possibility is that the evolutionary process starts from an essentially unbiased DNA while there is a mechanism that, over time, creates a bias on the DNA. One possible (if perhaps not very realistic) process is segmental copying of subsequences of the DNA onto itself. This process, called "copy-mutation", works as follows:

- Randomly choose two stretches of DNA, s_1, s_2 of equal length m, where m is a random variable uniformly chosen from the interval $[m_{\min}, m_{\max}]$; the stretches s_1 and s_2 are allowed to overlap.
- Replace s_2 by s_1.

Application of a copy mutation very likely increases the number of copies of the source sequence s_1 by 1.

2.2. Conservation of the USS over evolutionary time-scales

Another set of arguments against the PFH scenario is based on the observation that the USS has been conserved over evolutionary time-scales. For example, the USS of HI and *Pasteurella multocida* are identical (see Table 1). This indicates two things. First, the USS might in fact not be an effective tag for conspecific DNA fragments. If a USS is shared between two or more species its presence on a fragment cannot be used any more to assume a conspecific origin of the fragment.

Second, it indicates that the specific USS sequence might be of adaptive significance. This also potentially undermines the credibility of the USS as a signal because for signals the actual sequence is unimportant; what is important is that the signal-sequence and the receptor-sequence are identical. One would then expect that, over evolutionary time-scales, the specific USS sequence would change (Bakkali et al., 2004). At the very least, the fact that it is conserved requires an explanation. However, note that similar arguments also apply in the case of a MDH scenario.

3. Definition of USS in a computational context

Before we can describe our simulation models, we need to clarify what we mean by a USS in the context of our computational experiments. In real systems, the notion of USS is fairly clear: a short subsequence of the genome that is highly repeated on the genome and identical to the receptor sequence.

In the context of the present computational experiments we will use the same criteria. We measure the rate of uptake of conspecific fragments containing the USS, the average length of the receptor sequence, and the number of copies of the receptor sequence on the genome. Altogether we say a USS has emerged if firstly, there is a high number of copies of the receptor sequence on the DNA, secondly, the lengths of the receptor sequences in the population are rather short on average and thirdly, the rate of uptake of conspecific fragments containing the USS is high in the population. Usually, this state is reached by a sudden transition from a less typical state at some point during the system's evolution.

4. The base model

In this section, we provide an informal overview of the basic elements of our computational models. This section is not strictly required for the understanding of the remainder of the article, but the reader might find it helpful in order to develop an informal understanding of the model.

In the models, bacteria ("agents") live in an computational environment[1] that contains strings of the letters `a, c, g, t` (the *DNA fragments*; all of which are of equal length). At every time step the environment is replenished with a certain number of randomly generated DNA fragments. We will refer to those randomly generated DNA fragments as *alien fragments*. Alien fragments are meant to model DNA fragments from non-conspecific cells in the environment of the bacteria.

The agents themselves are represented as simple strings of `a,c,g,t` of specified length longer than the fragments. We will refer to those strings as the agents' *DNA*. The main activity agents engage in is the uptake of DNA fragments from the environment. Before uptake, agents compare the candidate DNA fragment under consideration with (depending on the model) either a specific sub-sequence of their own DNA or with a globally fixed string of `a,c,g,t`.

[1]In all simulations presented here the environment is divided into two cells. Offspring might not be placed into the same cells as their parents; agents remain in the same cell during their lifetime. It has been shown previously (Chu et al., 2005) that the cell structure of the environment has negligible effect on the behaviours of the model.

Agents mainly die because of over-crowding. The environment in which agents live is assumed to have a limited carrying capacity. Once this upper limit is reached, for every new born agent, the oldest agent currently in the system will be killed. Another possible reason for the death of an agent is old age. Whenever an agent is killed, a randomly chosen piece of its DNA is placed into the environment. These pieces taken from agents constitute the second type of DNA, which we will refer to as *bacterial fragments*.

Agents can give rise to offspring if they have collected enough pay-off. Newly born offspring might be mutated. There are several ways in which mutations can affect the agent: Firstly, the DNA itself might be changed (either through point mutations or copy-mutations). Secondly, mutations might change the length or position of the sub-sequence that determines the receptor sequence.

In what follows we will introduce a number of computational models each of which implements a particular scenario for the evolutionary origin of the USS. The main distinction between the models below concerns the pay-off agents receive for bacterial and alien fragments, respectively. Furthermore, in models of the MDH scenario, the sequence agents use to compare fragments with is fixed for all agents and times during a simulation, whereas in models of the PFH the receptor sequence is a specific sub-sequence of their DNA.

5. Evolution of USS in PUREPFH

Chu et al. (2005) confirmed that a USS can indeed emerge in a PFH scenario with copy-mutations. This model used the following algorithm, PUREPFH.

1. Initialize the population of agents by endowing them with a randomly generated DNA of fixed length l.
2. Specify for each agent a random subsequence of the DNA with maximal length $u_{\max}$; this subsequence functions as receptor sequence s.
3. Seed the environment with randomly generated DNA fragments of length f.
4. Update all agents as follows:
   (a) Check k DNA fragments from the environment for the sequence s.
      (i) If no fragment contains s, then take the last fragment.
      (ii) If a match is found, then take the matching fragment and abort search.
   (b) Exchange the fragment for pay-off points. Randomly generated ("alien") fragments and fragments taken from dead agents ("bacterial") receive different amounts of pay-off.
   (c) If the accrued number of pay-off points exceeds a certain threshold then reproduce.
      (i) If the system has reached its carrying capacity, then kill off one of the oldest agents to create space for offspring.
      (ii) Offspring is mutated with probability *mut* in one of the following ways:
         - A point mutation of the DNA.
         - A copy-mutation of the DNA.
         - Change the subsequence of the DNA that is used to determine s in one of the following ways:
            - Shift it by one, either to the left or to the right.
            - Adjust its length by ± 1 either to the left or to the right.
   (d) If an agent has reached the maximum age, kill it.
5. For each agent killed in this time step choose a random piece of its DNA of length f and place it into the environment.
6. Refill the environment with randomly generated DNA fragments of length f.
7. Goto 4.

Fig. 1 summarizes results obtained from an implementation of PUREPFH. The main conclusion is that a USS does evolve provided that the pay-off for bacterial DNA is moderately higher than the pay-off for uptake of alien DNA fragments and provided that there is a mechanism that creates a bias on the DNA. The results indicate that the conclusions scale well with the length of the USS

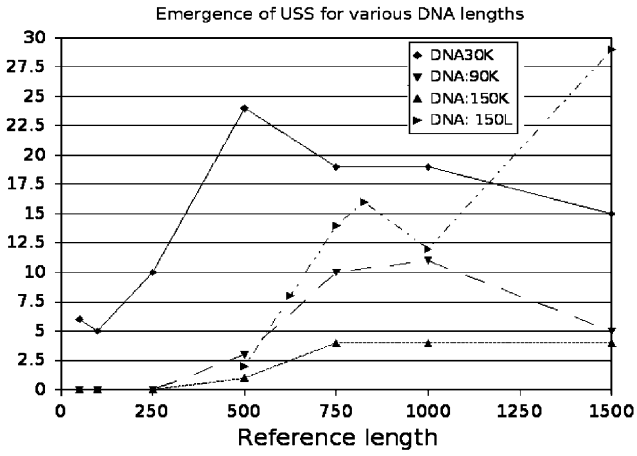

Fig. 1. Simulation results obtained with the PUREPFH algorithm. The y-axis records the number of simulations (out of 60 trials) where a USS emerged within the first 1 million time steps for simulations with various DNA-lengths. The x-axis shows the length of the DNA fragments in the environment. In the shown simulations, the upper limit for the length of the receptor $u_{\max}$ is equal to the fragment size f, except for the run labelled "DNA150L," where the receptor has a fixed upper limit of $u_{\max} = 500$. In the case where the receptor length is equal to the fragment length there tend to be fewer USS as the DNA length and fragment size increases. However, once the upper limit for the receptor length is fixed (as in the case of the DNA150L run) the number of instances where the USS emerges increases with the fragment size.

provided that the range of possible lengths for the receptor u_{max} is limited to not too long values. Chu et al. found that imposing an upper limit of 500 to possible lengths of the USS is sufficient to make the emergence of USS a probable event even for long DNA strings (see Fig. 1). Nowhere in PUREPFH is the assumption of group selection used, yet a USS emerges. Hence, group selection is not necessary to explain the evolutionary origin of the USS.

6. Models and results

The experiments by Chu et al. show that a USS certainly could have evolved in a PFH scenario. This does not mean that it actually has. In this section, we will introduce various scenarios/algorithms that formulate alternatives to PUREPFH. No claim is made to those proposed scenarios exhaust the space of all possible or relevant scenarios.

Unless specifically stated otherwise, all results reported in this section assume a DNA of length $l = 10\,000$, a maximum signal length of $u_{max} = 50$, a fragment length $f = 100$, and a maximum population size of 30. In all experiments we performed the system was initialized with 29 different agents and we ran simulations for 500 000 time steps; furthermore, the ratio of bacterial to alien fragments in the environment was no better than 1:20. Unless stated otherwise, in all following figures each data point corresponds to a time average over the population average of the corresponding quantity. In most cases the average has been taken over the last 400 000 time steps. Exceptions have only been made where no USS emerged within the first 100 000 time steps.

A possible criticism of the following results is the rather short DNA compared to real organisms. The drawback of such a short DNA is certainly that it substantially reduces the quantitative reliability of the results obtained. We think that this loss is a worthwhile price to pay for the decreased computational costs of the simulation runs which in turn allowed a more generous exploration of the parameter space of the system.

6.1. PUREMDH

The first scenario we considered was a minimal version of the MDH. We assume an initially random sequence and a receptor that takes up DNA fragments from the environment if they contain a pre-specified sequence of length 9 (in all MDH experiments the receptor sequence was "aagtgcggt"). We used the following algorithm PUREMDH:

1. Initialize all agents by endowing them with a randomly generated DNA of fixed length l.
2. For each agent set the receptor sequence to a fixed string s. This sequence is the same for all agents and does not change during runtime.
3. Check a fixed number of fragments in the environment.
   (a) If find a fragment containing s then recombine with it.
   (b) If have enough energy reproduce (reproduction/mutations as in PUREPFH).
4. Generate random fragments and place bacterial and random fragments in environment.
5. Goto step 3.

Simulations of PUREPFH show that a USS emerges for receptor sequences of length up to $u = 16$ (data not shown). For USS of lengths $16 < u < 19$ a USS still emerges but potentially only after a long time. We never observed the emergence of a USS for $u > 18$.

6.2. RECOMBPFH

The above algorithms PUREMDH and PUREPFH represent bare-bone versions of the MDH and PFH, respectively. A possible variation of the PFH replaces copy-mutations with recombination with DNA fragments. The algorithm RECOMBPFH is similar to PUREPFH but contains the following variations:

1. Agents perform only point mutations (no copy mutations).
2. Agents may recombine with DNA fragment they take up. The specificity can be varied as follows:
   - At every time step agents recombine with the fragment they take up ("recomb").
   - Agents recombine only with fragments containing the exact receptor sequence.
   - Agents recombine only with perfect matches and with single mismatches with probability $\frac{1}{2}$ ("mismatch = 1").
   - As before but they also recombine with double mismatches with probability $\frac{1}{3}$ ("mismatch = 2").
   - etc…

We performed a number of simulations of RECOMBPFH varying key parameters. Some results of simulations are summarized in Figs. 2–5. RECOMBPFH does lead to the emergence of USS. Furthermore, the quantitative and qualitative details of the behaviour of the model only weakly depend on the mutation rate (data not shown). Unless explicitly stated otherwise we have therefore chosen a fixed mutation rate of 0.3.

For non-zero mutation rates the main parameter determining the emergence of USS is the pay-off received for bacterial fragments. For low levels of pay-off (up to about 25) no USS emerges. Up to a pay-off of

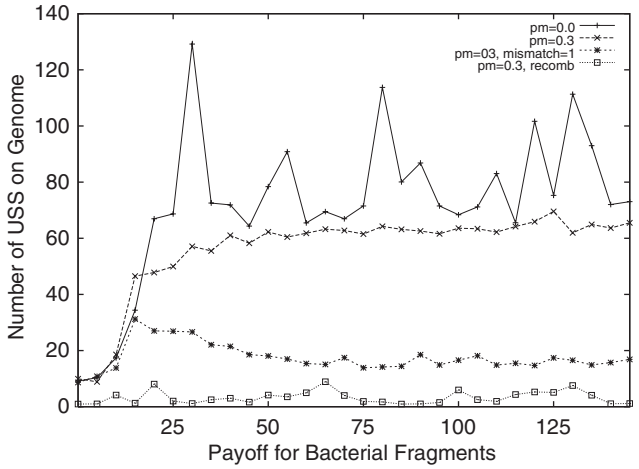

Fig. 2. RECOMPFH:The number of USS on the DNA as a function of the bacterial pay-off. The pay-off is increased in steps of 5 along the x-axis. Each data point corresponds to the average over the population over 400 000 time steps. For a point mutation rate $pm = 0.3$ the number of USS increases sharply initially but levels off once the bacterial pay-off reaches about 25. If DNA fragments are incorporated into the DNA irrespective of whether they contain the USS or not, then the number of USS will be considerably lower ($pm = 0.3$, recomb). Similarly, if the receptor mechanism is imperfect and also accepts single mismatches with a certain probability ($pm = 0.3$, mismatch = 1), then this leads to a much lower count of perfect USS.

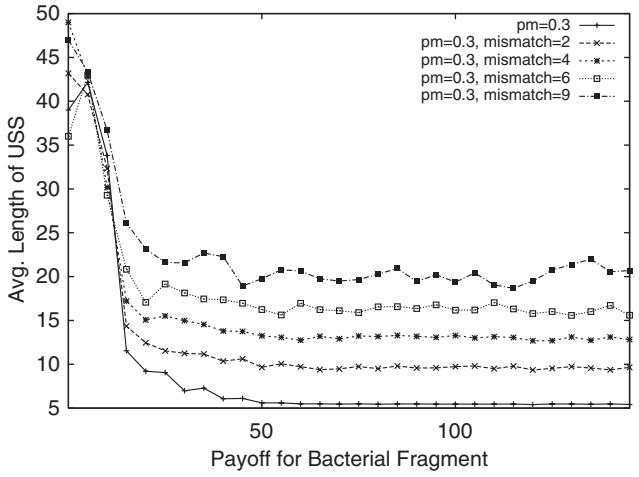

Fig. 3. RECOMPFH: Length of the USS as a function of bacterial pay-off for various uptake accuracies. The less accurate the uptake mechanism, the longer the USS.

about 50 an increase of the pay-off leads to a steady increase of the quality of the signal. Beyond this the quantitative behaviour of the model does not change significantly. Fig. 2 shows that a USS emerges even if mutations are turned off ($pm = 0.0$); this is a finite population effect: The initial population has a range of receptor sequences of different lengths. For each lineage the receptor sequence is fixed. However, by chance some will relatively quickly produce a bias on the DNA (by

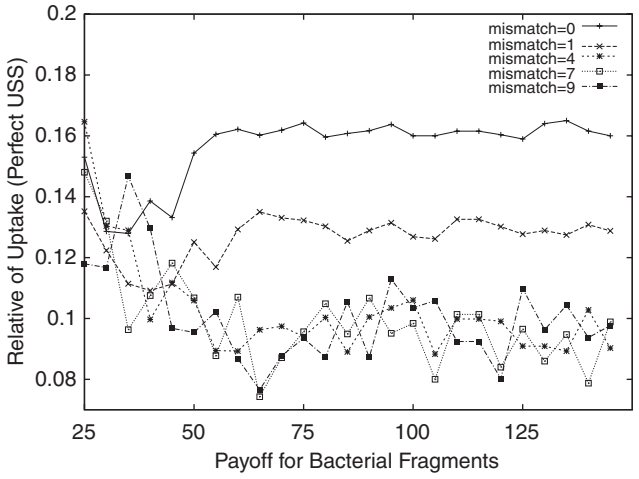

Fig. 4. RECOMBPFH: The uptake efficiency for perfect USS (ratio of DNA fragments that contained the signal sequence and were bacterial and fragments that contained the signal sequence and were alien). For the $pm = 0.3$ model only about 16 percent of the recognized fragments are not false recognitions. This number drops to under 10 percent if mismatches are allowed.

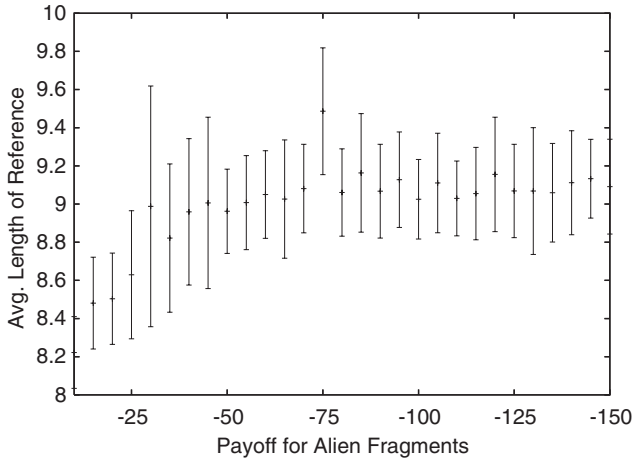

Fig. 5. RECOMBPFH: In these simulations, the pay-off for bacterial fragments has been kept constant at 50, whereas the pay-off for alien fragments was allowed to take various negative values.

recombination) and will consequently be better at recognising conspecific DNA. The corresponding lineage will thus outperform the rest of the population. This effect will only be relevant if there is sufficient extra pay-off to be gained from uptake of bacterial fragments. Fig. 2 shows that with $pm = 0$ a USS emerges only if the pay-off is higher than about 25.

Note that a USS will only emerge if the recombination is restricted to sequences that contain the receptor sequence. If fragments are taken up irrespective of whether or not they contain this sequence (curve labelled "recomb" in Fig. 2) then a USS will normally not emerge.

A USS also emerges if the receptor specificity is reduced. In this case a USS still evolves, but the efficiency of USS uptake (that is the ratio of true-positive recognitions over false-positive recognition) is drastically reduced (see Fig. 4). Furthermore, note that a reduction of the receptor specificity leads to an increase of the length of the USS (see Fig. 3).

Once a signal emerges the length of the USS itself is only weakly dependent on the absolute value of the bacterial pay-off. However, it is strongly dependent on the sign of the pay-off for alien fragments. Fig. 5 shows a set of results where the bacterial pay-off is kept fixed at 50, and the pay-off for alien fragments is decreased from -5 to -150. In this area the fragment length increases from about 8.5 to about 9 (however see error-bars). This is significantly higher than the fragment length of between 5 and 6 in the regime of positive pay-off for alien fragments.

6.3. FITMDH

FITMDH is a variation of the RECOMBPFH algorithm in that it keeps the receptor sequence fixed for all agents and all times. Note that FITMDH is identical to PUREMDH except for the fact that bacterial fragments give a higher fitness than alien fragments. This introduces an adaptive pressure to recognize bacterial fragments.

The number of USS on the genome and the uptake rate of perfect USS under the FITMDH shows a qualitatively similar dependence on the pay-off for bacterial fragments as under the RECOMBPFH. Both the number of USS on the genome and the uptake rate are strictly higher in the FITMDH case than in the RECOMBPFH case irrespective of the mutation rate. Fig. 6 shows that the relative rate of USS uptake of the FITMDH model is more than 100 times higher than the corresponding measure in the RECOMBPFH model (see Fig. 4), although the absolute rate of correct recognitions of fragments containing a USS is only slightly higher. This means that the FITMDH model has a substantially lower rate of false-positive recognitions (Table 2).

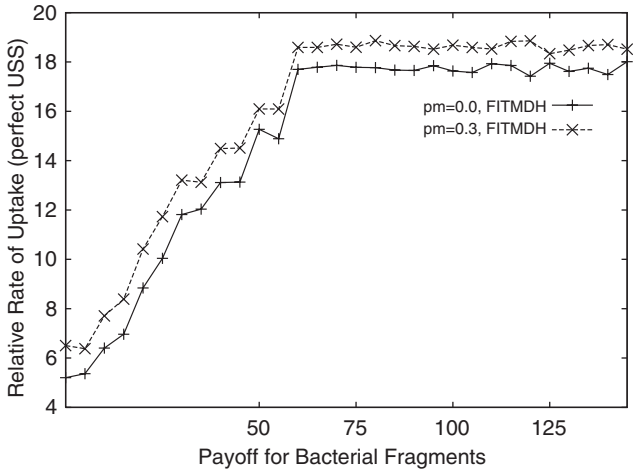

Fig. 6. FITMDH: The relative rate of USS uptake for the FITMDH model. The upper curve shows the relative uptake rate for the model with a mutation rate of $pm = 0.3$ and the lower one shows the model with mutations turned off.

7. Discussion

The simulations indicate a strong dependence of the RECOMBPFH scenario on the sign of the pay-off for alien DNA. The following discussion will therefore distinguish between the case where alien fragments are deleterious (carry negative pay-off) and the case where they are merely less advantageous than bacterial fragments. Currently, it is unknown what happens with DNA fragments that are taken into the cell. Consequently, it is not clear how to determine or estimate the fitness of fragments.

In the case of deleterious fragments there is an additional factor that needs to be taken into account: Can a DNA uptake machinery evolve at all? Consider the following: before a molecular tag evolves, bacteria will sample the fragments in their environment randomly. Hence, their uptake of alien fragments will be proportional to the concentration of alien fragments in their environment. If most fragments were alien and carried negative fitness, then those bacteria that avoided uptake altogether would have an adaptive edge over those that did take up DNA. As a result DNA-uptake would be a maladaptive trait and thus never evolve. Having an uptake machinery only pays if the expected pay-off (over some time) is positive. More specifically , if $n_a(n_c)$ is the number of alien (conspecific) fragments taken up over a lifetime and $g_a(g_c)$ the corresponding

Table 2
Features of the various algorithms

	PUREPFH	RECOMPFH	FITMDH	PUREMDH
Receptor	Variable	Variable	Fixed	Fixed
Recombination	No	Yes	Yes	Yes
Copy mutations	Yes	No	No	No
Differential pay-off	Yes	Yes	Yes	No

pay-off, then an initially unbiased uptake mechanism can be maintained if

$$n_a g_a + n_c g_c > 0. \quad (1)$$

Given this condition is fulfilled, we can now continue by estimating the optimal length for the USS in the case of alien fragments being deleterious. From the point of view of agents, the problem is to strike the right balance. The longer the USS, the fewer randomly generated alien fragments will contain the entire signal sequence; at the same time a longer USS will also reduce the probability that a bacterial fragment contains an entire USS. Consider[2] the respective probabilities to find a signal on a conspecific (P_c) and on a random fragment (P_r).

$$P_c = 1 - \left(1 - \frac{n}{l}\right)^{f-u+1},$$
$$P_r = \left(1 - \left(\frac{1}{4}\right)^u\right)^{yf-u+1}. \quad (2)$$

Here, as before, u, f, l are the lengths of the USS, the fragment and the DNA respectively, while n is the number of USS on the genome and y the fraction of alien DNA fragments in the environment. P_r is the probability that a random sequence of length yf contains a USS of length u. P_c is the probability that a sequence of length u that is present in n copies on a string of length l will be contained on a randomly chosen subsequence of length $l > f > u$. Given this, the probability of taking up a conspecific DNA fragment is:

$$P_{USS} = P_c(1 - P_r). \quad (3)$$

This can be used to calculate the optimal length of the USS in the case when the alien pay-off is negative. Setting $y = 20, l = 10\,000, f = 100$ and $n = 60$ (approximate values used in simulations) yields an average USS length of $u \approx 9.2$ which is within the range of error of the computational experiments (see Fig. 5). Similarly, the equations can also be used to predict a signal length of 14 for biologically realistic parameter settings ($l = 1\,800\,000$, $f = 2000, n = 1470, y = 20$). This is in reasonable good agreement with simulations of RECOMBPFH using the same parameter setting but does not correctly predict the observed signal length in HI for the particular choice of $y = 20$.

Setting in the known values for HI ($u = 9, l = 1\,800\,000, n = 1470$) one can determine the ratio of conspecific and alien fragments in the environment for which a USS of length 9 would be optimal. This ratio does depend on the fragment length. Fig. 7 shows various predictions for the content of alien DNA fragments in the environment for fragment lengths between 500 and 10 000. Taking all this into account, one would expect that there are considerably fewer alien than conspecific fragments in the environment of HI.

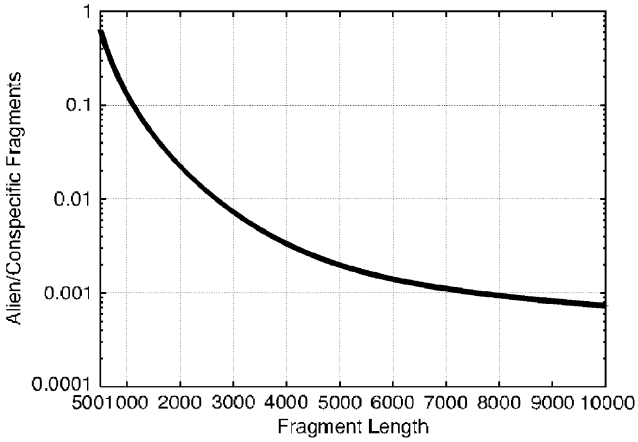

Fig. 7. This graph shows the fraction of alien DNA fragments in the environment as a function of fragment lengths for which a USS length of $u = 9$ would be optimal. In HI about every fragment of length 2k contains a copy of the USS. If this reflects the typical fragment length, then this suggests that fraction of alien DNA is about 0.02.

If one assumes that alien DNA fragments are not deleterious (carry negative pay-off), but merely less advantageous than conspecific DNA (positive pay-off), then the situation changes somewhat. The computational results with both PUREPFH and RECOMBPFH (for various DNA and fragment lengths; data not shown) indicate that in this case one would expect a USS of shorter length (for HI: 6–8 rather than 9; data not shown). If agents can collect pay-off from both alien and bacterial fragments, then a certain level of false positive recognitions does not do any harm, as long as it helps to increase the rate of correct recognitions. Hence, it is beneficial to have somewhat shorter signal sequences than in the case of deleterious alien fragments.

Fig. 4 indicates that the length of the USS also strongly depends on the specificity of the uptake mechanism. As the specificity decreases, the length of the USS will increase. This effect can be explained by considering the probability that a specific fragment contains the USS (up to mismatches) and is alien. In the case of a fault intolerant uptake mechanism the probability to find the fragment on a random string is proportional to $(\frac{1}{4})^u$, where u is the length of the USS. If one allows for i mismatches, then the same probability is proportional to finding any string of length n with $n - i$ letters matching and i mismatches. There are $\binom{n}{n-i}$ possibilities to arrange the mismatches. Thus the probability is proportional to

$$\binom{n}{n-i}\left(\frac{1}{4}\right)^{n-i}\left(\frac{3}{4}\right)^{i}.$$

[2]In order to simplify the analysis, this model assumes that bacterial fragments are as good as alien fragments are bad. This will most likely not be true. However, as Fig. 5 shows, the length of the USS is only weakly dependent on the amount of negative pay-off.

Note that this is only applicable in the case of an unbiased base alphabet. For strings with a biased base composition the probability of occurrence of a sequence s will depend on the base composition of the sequence. This complicates the mathematics considerably. We will therefore restrict our analysis to the case of an unbiased base alphabet.

Consider the following: If a signal sequence is of length 5 then one would expect to find it on a random string of length 1098, whereas a single mismatch of it is expected to be contained on a random string of length of only 72. Given that the simulations of the RECOMBPFH yielded a USS of length only slightly higher than 5 and a fragment length of 100, it is clear that a USS of length 5 would be useless as a discriminator between alien and bacterial DNA if one allowed single mismatches. Every fragment would be expected to contain such a sequence. Indeed the simulations (Fig. 3) showed that allowing single mismatches leads to an increase of the length of the USS sequence to about 9. Here one would expect to find a single mismatch on a random string of length 9700.

An increased length of the receptor sequence helps agents to prevent false-positive recognitions, yet it does not increase (but indeed decreases) the rate of correct recognitions. Unsurprisingly, then, the relative (Fig. 4) and the absolute (data not shown) rate of bacterial uptake decreases if mismatches are allowed.

Altogether, we conclude that PFH models can both explain the emergence of the USS and the observed length of the USS core sequence as observed in some species (assuming plausible parameter settings). In the case of the MDH no such explanation is required because the length of the USS sequence is not assumed to be an adaptive response to an external selective pressure. Essentially it can thus be regarded as a random variable. Our simulations with the rather short DNA sequences of 10k show that receptor sequences of length up to 16 will spread on the USS through recombination. Given those results one would thus expect to see a range of USS lengths in various species. In real bacteria, however, signal sequences are limited to lengths of 9 or 10 (see Table 1). In some, but not all, cases this can be explained by a common evolutionary origin of the USS. Identifying further constraints for the length of the USS under the MDH is a challenge for future research.

Closely connected to the common evolutionary origin of the USS is another set of questions. As mentioned above, there are instances of species that share the same USS. This is a problem for both the MDH and the PFH. On the one hand, the MDH cannot explain why the specific receptor sequence should be conserved over evolutionary time-scales. The signal (and the receptor) sequence are assumed to be parasitic serving no adaptive role (although not harmful either). Hence there should be no forces conserving them. On the other hand, once shared with other species in the same habitat, the USS loses its efficiency as a molecular tag; the conservation of the USS over evolutionary time-scales is thus a problem for the PFH as well.

There are two possibilities to save the PFH under those circumstances: Firstly, the species might be sufficiently closely related to each other to make genetic exchange between them viable. Secondly, if gene exchange is not viable, but the species rarely encounter each other's DNA fragments, then the shared USS will not reduce fitness. This second case though leads to another problem: Why was the USS conserved over evolutionary time-scales? Normally, one would expect millions of years of evolution to lead to a somewhat different USS.

A possible explanation for this is a lock-in effect. As discussed above, the emergence of a USS requires a highly repeated signal sequence and a receptor sequence identical to the signal sequence. A mutation of the receptor sequence from s to s' would lead to an uptake mechanism with a new signal sequence s'; s' will typically be much less abundant on the genome than s. Assuming a PFH model, those agents that did not undergo the mutation will be fitter than those who did, hence mutations from s to s' will not be evolutionary stable. Thus the suboptimal (because shared) receptor s will remain and the globally optimal solution will not be found by the system.

It should be noted though that there might be ways to get away from this local fitness maximum. For example, a gradual increase of the length of the USS coupled with a reduction in the specificity of the receptor might offer a way for agents to gradually change the USS sequence without suffering a decrease of fitness at any point. Furthermore, note that we could not find evidence for the conjectured lock-in effect in our computational experiments.

A final discussion point concerns a possible scenario for the emergence of the MDH. A core assumption of the MDH is that recombination events happen frequently enough to leave a permanent mark on the genome. Given that recombinations with alien DNA fragments are likely to have lethal consequences for the cell, those cells that take up a lower proportion of alien USS will do better in the long term. In this scenario, therefore, alien DNA fragments yield negative pay-off. As such, the MDH undermines its own assumption of neutral pay-off difference between alien and conspecific fragments. This does not mean that a USS could not evolve under an MDH-scenario, but rather that it leads to scenarios that are in-between the PFH and MDH (such as FITMDH).

8. Conclusion

We have investigated the functional dependence between variables in some scenarios for the evolutionary

origin of USS. Computational simulations indicated a number of relationships between the variables of the systems. Firstly, the length of the USS is only weakly dependent on the length of the fragment, the length of the DNA, and the pay-off for bacterial fragments. It is strongly dependent on the sign of the pay-off for alien fragments and the specificity of the receptor. Secondly, for realistic parameter values, RECOMBPFH is compatible with the following two scenarios:

- If the pay-off for alien DNA fragments is negative, then one would expect a low concentration of alien fragments and rather short fragment lengths.
- Alternatively, if the pay-off is positive, the receptor specificity can be expected to be high and fragments to be longer. (Specificity and fragment length depend on each other).

There are fundamental problems with both hypotheses, notably the conservation of the USS over evolutionary time-scales. It is likely that this conservation will require a separate explanation. Possibly new sequencing data will give more insight into the evolutionary relationship between USS in various species.

Another issue is that the MDH might undermine itself. Even if DNA uptake started in an MDH-like scenario, alien fragments might acquire negative pay-off thus exerting evolutionary pressures on the uptake system. Finally, a further weakness of the PFH is that it cannot explain the uniform distribution of USS-lengths across species.

The models presented in this paper are maximally simple models designed to create a basic understanding of the inter-relationships between variables. In order to settle the question of the evolutionary origin of USS, more biologically realistic models are needed. Also more empirical data is required. Particularly:

- What is the typical length of DNA fragments?
- What is the typical concentration of alien DNA fragments in the environment of competent bacteria?
- What is the recombination rate with alien/conspecific fragments?

Acknowledgements

DC and JR thank the "Paul & Yuanbi Ramsay Research Fund" for financial support of this work. DC also thanks the Norwegian Research Council.

References

Bakkali, M., Chen, T., Lee, H., Redfield, R., 2004. Evolutionary stability of DNA uptake signal sequences in the Pasteurelleceae. Proc. Natl. Acad. Sci. 101 (13), 4513–4518.

Chu, D., Lee, H., Lenaerts, T., 2005. Emergence of uptake signals in bacterial DNA. Artificial Life 11(3), in press.

Finkel, S., Kolter, R., 2001. DNA as nutrient: novel role for bacterial competence gene homologs. J. Bacteriol. 183 (3), 6288–6293.

Hsieh, L., Luo, L., Ji, F., Lee, H., 2003. Minimal model for genome evolution and growth. Phys. Rev. Lett. 90 (5), 101–104.

Karlin, S., Mrazek, J., Campell, M., 1996. Frequent oligonucleotides and peptides in the *Haemophilus influenzae* genome. Nucleic Acids Res. 24, 4263–4272.

Lorenz, M., Wackernagel, W., 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Mol. Biol. Rev. 58 (3), 563–602.

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. Mol. Microbiol. 40, 700–707.

Redfield, R., 1988. Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? Genetics 119, 213–221.

Redfield, R., 1993. Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. J. Hered. 84 (5), 400–404.

Redfield, R., 2001. Do bacteria have sex? Nat. Rev. Genet. 2, 634–639.

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.

Smith, H., Gwinn, M., Salzberg, S., 1999. DNA uptake signal sequences in naturally transformable bacteria. Res. Microbiol. 150, 603–616.

Solomon, J., Grossman, A., 1996. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12 (4), 150–155.

Solomon, J., Grossman, A., 2000. Who's competent when: regulation of natural genetic competence in bacteria. Proc. Natl. Acad. Sci. (PNAS) 97, 6981–6985.

Tomb, J., el Hajj, H., Smith, H., 1996. Nucleotide sequence if a cluster of genes involved in the transformation of *Haemophilus influenzae* Rd. Gene. Trends Genet. 12 (4), 150–155.

Wang, Y., Goodman, S., Redfield, R., Chen, C., 2002. Natural transformation and dna uptake signal sequences in actinobacillus actinoomycetemcomitans. J. Bacteriol. 184 (13), 3442–3449.