Laboratory Rotation Report

Aims

The aims of the project were to produce an NK-type fitness landscape model in Python 3, and to examine the changes in average fitness and sequence diversity when the mutation rate and K (epistatic coupling parameter) were increased, starting from an initial population of random DNA sequences.

Background

A fitness landscape is a mapping of genotype to fitness. Adaptive processes on fitness landscapes are random walks across "peaks" and "valleys" of fitness to arrive at a genotype of higher fitness than the starting point. Since the space of possible genotypes is very large, often it is impossible to sample every genotype in order to determine the global fitness maximum. For example, a DNA sequence of length 60 has 4^{60} possible genotypes, too large to sample the entire space. Such a problem is a class of combinatorial optimisation, where only a sample of genotype space is utilised to find a fitness maximum. A simple optimisation strategy is the random walk, where the fitness of nearest neighbours to the starting genotype is assessed. Improvement in the average fitness of a population then proceeds through a cycle of reproduction of high-fitness members, mutation of offspring to generate new members, and the elimination of members with low-fitness (Kauffman & Levin, 1987).

If there are no interactions between the sites of a sequence, then an adaptive walk can proceed to the global maximum smoothly akin to a hill-climbing exercise. However, the existence of such interactions, or epistasis, contributes to the ruggedness of a fitness landscape, where multiple local maximums exist. The NK-model was introduced to study the dynamics of such landscapes, where N is the length of the sequence and K a "tuning parameter" which controls the degree of site interaction. For example, if K = 0, this resolves to a landscape with no epistasis, while K = 1 and K = 2 entails landscapes in which 1 and 2 other respective sites interact (Kauffman & Weinberger, 1989).

The emergence of new function in *de novo* genes is garnering interest (e.g Schlötterer, 2015). In contrast to the evolution of gain of function after gene duplication, the starting point of *de novo* gene evolution is essentially a random sequence (Andersson *et al*, 2015). However, the dynamics of evolution of *de novo* genes are little studied in NK-models. Unlike in traditional NK-models which only examines a single fitness-function mapping, *de novo* genes could potentially evolve towards a function from among a number of possible functions, depending on the selective pressures on the population and the interactions with the gene network. Thus, pertinent questions of the dynamics of evolution from a population of random sequences remain. These include: 1) How does the length of a sequence and population size affect the average fitness, 2) How does epistasis and mutation rate affect average fitness, 3) Is the trajectory of evolution similar from different starting sequences, and 4) Do the evolved, simulated sequences "look" similar to *de novo* genes found in nature?

Methods

A "greedy algorithm" (Kauffman & Levin, 1987) was implemented in Python 3 (**Fig 1a**). Epistasis was modelled up to K = 2 (**Fig 1b**). It should be noted that assigning fitness to DNA sequences was explicit (**Fig1b**); although this entails that only a low degree of epistasis can be modelled without running out of computer memory, it allows for the comparison of the computer evolved sequences to natural, extant ones. Sequence diversity in a population was calculated using a Hamming distance measure over the global population (**Fig 1c**). Selection of a population member to reproduce was implemented by fitting a Gaussian distribution centred on the maximum fitness (**Fig 1d**) with a variance of 1; this ensured that sequences with high fitness would have the highest probability of being selected to reproduce, although it is not impossible for low-fitness sequences to be chosen. Five simulations were run for varying parameters; sequence length (N = 60, 99, 150 or 300 nucleotides), mutation rate (1 or 10 substitutions per reproduction), population size (R = 10 or 100 members), and epistasis (K = 0, 1, or 2). All simulations used the same dictionary for fitness assignment.

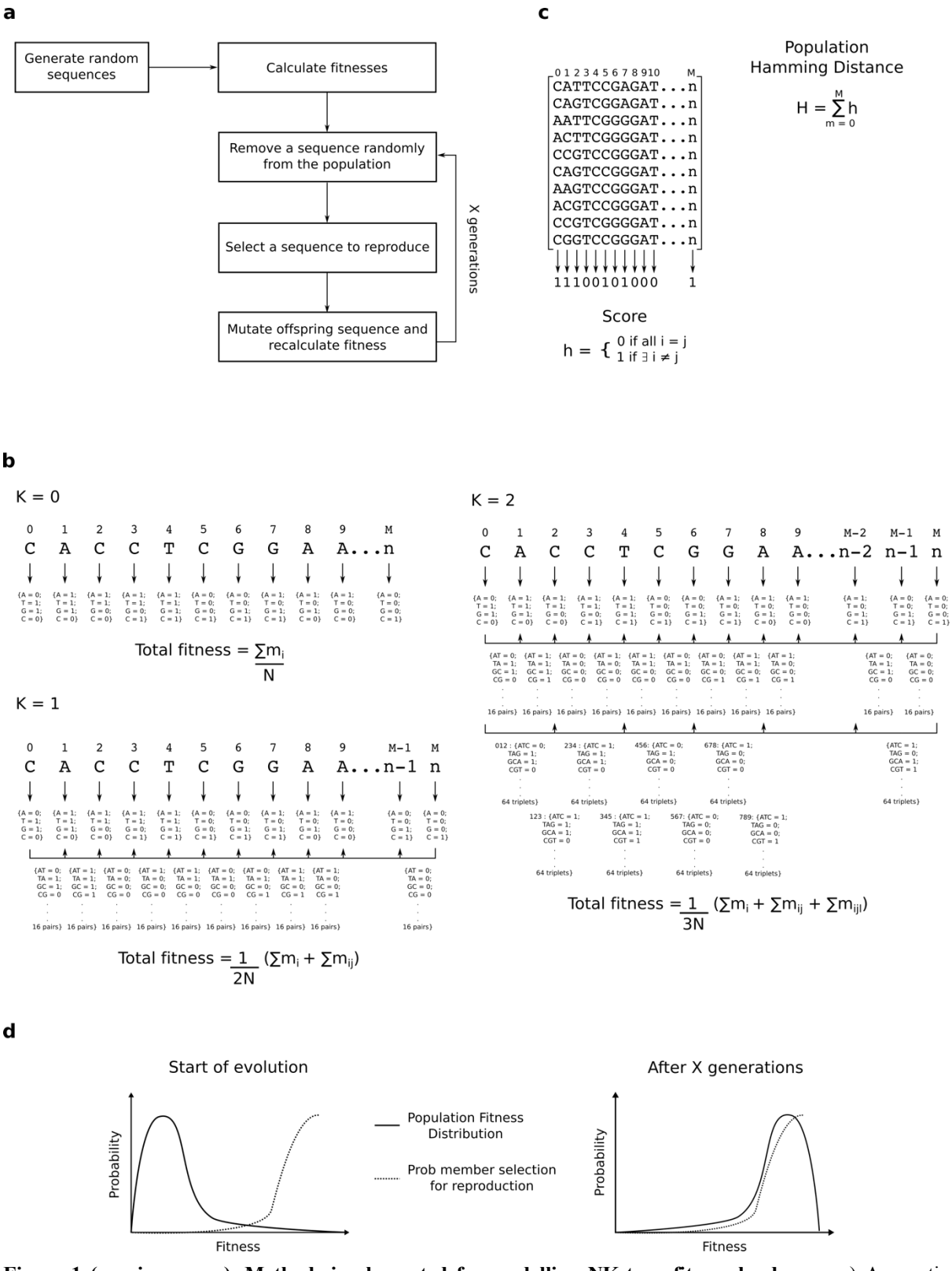


Figure 1 (previous page). Methods implemented for modelling NK-type fitness landscape. a) A genetic algorithm was applied to an initial population (R = 10 or 100) of randomly generated sequences, not necessarily protein encoding. The population size was kept constant throughout. b) Fitness was calculated using a dictionary with nucleotides randomly assigned as either 0 (no fitness contribution) or 1 (fitness contribution). When no epistasis is present (K = 0), each position in the sequence is assigned fitness independently of the other. When each site interacts with one other site (K = 1), an additional fitness contribution between the nearest neighbour pairs are considered. When each site interacts with two nearest neighbours (K = 2), an additional fitness contribution of the triplets is included. For epistatic fitness contributions, the sequences is considered circular,

and hence the n^{th} nucleotide would pairwise interact with the 1^{st} nucleotide; in a triplet interaction, the n-1^{th}, n^{th} and 1^{st} nucleotide interact, and the n^{th}, 1^{st}, and 2^{nd} nucleotide interact. c) The population sequence diversity is calculated by assigning a 0 for invariant sites and 1 otherwise, and then summing the scores over the sequence. d) Selection of a population member to reproduce is completed by calculating a theoretical Gaussian distribution centred on the maximum fitness for that generation with a variance of 1; a random number is drawn from this distribution; the first sequence in the population to be greater than this number is chosen to reproduce. At the start of the evolutionary process, few sequences have high fitness, but by X generations, the fitness distribution of the population is skewed towards high-fitness.

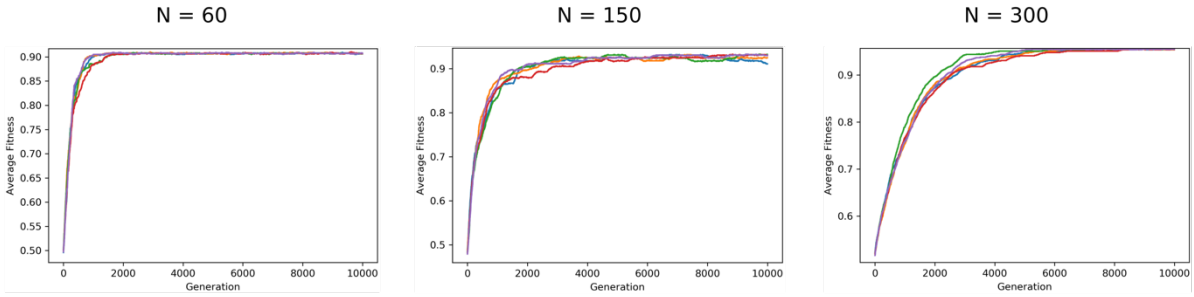
Results

In the absence of epistasis, all sequences converge to highest possible fitness, with shorter sequences converging in fewer generations than longer ones (**Fig 2a**). Smaller populations show larger fluctuations in average population fitness, most likely due to less sequence diversity than larger populations (**Fig 2b**). It is well-known that small populations are less robust than larger ones, since most mutations are usually deleterious rather than beneficial and hence in a small population these non-beneficial mutations are over-represented, leading to stochastic fluctuations in population fitness (Whitlock & Bürger, 2004). Unfortunately, the method for calculating sequence diversity is not equipped to detect more than one isoform.

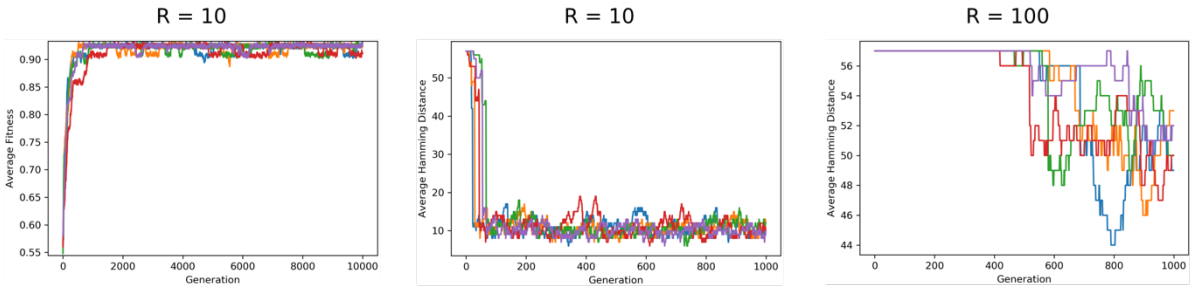
When epistasis is introduced, there is unlikely to be a solution which satisfies all constraints; however, many more "second-best" solutions are available than is the case for no epistasis (data not shown). When epistasis is present, the starting point (initial random sequences) affects the outcome of evolution (**Fig 2c**); different fitness plateaus are reached even though all simulations used the same fitness assignments, in contrast to the absence of epistasis, where all simulations converged to the same fitness. High mutation rates lead to suboptimal population fitness, creating large fluctuations in average fitness during evolution, which is more pronounced for smaller populations. The effect of mutation rate is proportional to sequence length, with shorter sequences showing larger fluctuations in fitness during evolution even in the absence of epistasis with the mutation rate used (**Fig 2d**).

In general, the computer evolved sequences show low-complexity (many repeating sequences) with a Shannon entropy near the expected value of 2 if all bases were present in equimolar concentrations (Adami, 2004) (**Table 1**). However, the Shannon entropy is not a measure of information-richness in a sequence, and other methods are required (Adami, 2004). Since no biochemical constraints were imposed on the model, the majority of the computer-evolved sequences have multiple stop codons in all three open-reading frames (data not shown). Other methods for assessing gene complexity and information-richness are targeted at genome-wide analysis (e.g Adami, 2004). Given the sparsity of information on the evolution of *de novo* sequences, and since the model is too simple to analyse complex gene-environment interactions, it is of limited utility in comparing the outcome of computer evolution to *de novo* genes found in nature.

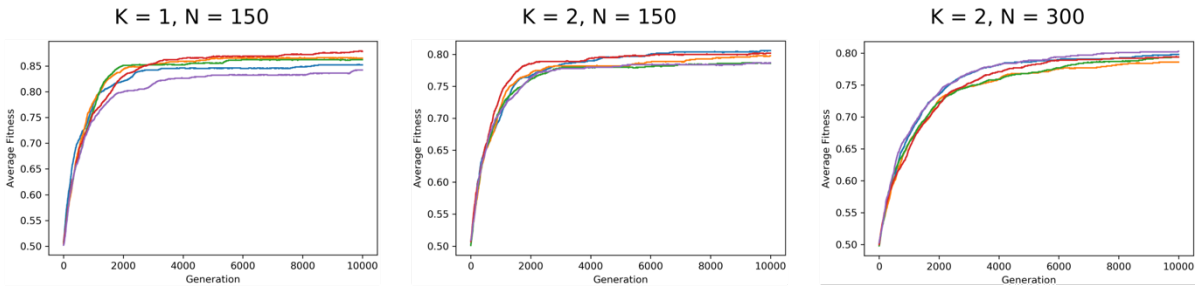
a Fitness increases based on sequence length



b Effect of population size



c Effect of epistasis



d Effect of mutation rate

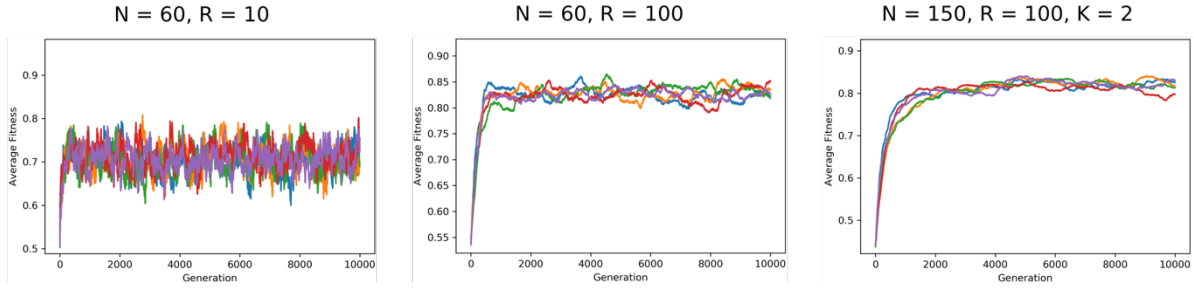


Figure 2. Results of evolution of random DNA sequences. **a)** Shorter sequences evolve faster towards higher fitness. **b)** Smaller populations show larger fluctuations in average fitness, as well as containing less sequence diversity. Graphs shown for N = 60 nucleotides. Note that population Hamming distance was calculated for 1000

generations. **c)** Higher epistasis leads to lower average fitness ("fitness plateau") reached by the population. In addition, different random sequences reach different fitness plateaus, in contrast to the case of no epistasis, where all starting points converge to the highest fitness. **d)** A high mutation rate (10 substitutions per reproduction) leads to loss of fitness, and fluctuations in average population fitness is more pronounced in smaller populations.

Table 1. Sample computer evolved sequences. N = number nucleotides, K = epistasis value, M = mutation rate, and R = population size. Shannon entropy given in brackets.

N	K	M	R	Sequence at Generation 10,000
60	0	1	10	ATGGACGTGTGTCTTGAAAAGATGGAAGAGTGGACTGAGGCGATGAAGTAGGGGTACCTG (1.855)
60	0	1	100	ATGTAAAACCAGTCCGTCTGAGGGACTAGATATCTGTAACAACTAGTCGTTAACTCCAGC (1.980)
99	0	1	10	ATGCGTTAGGCCCAGTGTAGGGGGCCTACTCTACATATTTGTTAACGGCAAGGAACCGAGGTGGGCTTACAGAACTAAATTGCTGGCGGTTTCCTGTGC (1.985)
99	0	1	100	ATGTCATTTATGCTCTGTAACATGTATAACGCTTGGGTCCTTACCTAATATTATGTGGTTCTTTCGTTCATTGATTGGCGTATGTCGCCGGACCTCACC (1.936)
150	0	1	10	ATGAAATCAATTACGAGCACCCTGACTCTCATGAGTAGATGCAACCCTGTTTGTAAGACGAAGGGAGCAACTGTCTTAAGTACATATGCAGTGTTCCCGGACGACTATCGGTGCACTACACGAGGTACTTAGACAAAGAGGACGAGTTTG (1.986)
150	0	1	100	ATGGTCTAGAGAAACTTAGTGGTCAAATGCAGTATAGGTCTTTAACTTAGCTGCGTAACTGGGCCTCCACGTTCACTCCATCGAGTCAGGTGTGCAACCCCCCGTGATACCGCCTTTAAGTATCTGTCTATCGCAAATTCAAGTTACTCG (1.992)
300	0	1	10	ATGCCACGTTAAGCCGACAGCACACGCCTCCTCGTAGAATCGTGCATTGATTCGCAAGTACTATGGTACAAGTAGTATCACAAGGGAAGAAGTGTCCCAACTCTGCTCTGATCCGGACGGGTTTCGAATCACCTGGTTACGTCTTTAGTCGAGTGGTCCTGAACACTGGCTCTAGATTTGGATGAAGAAATTAGCAAAATTGTTAAAATGGATTCCATATCATGAGGTACCAAGTTGCTCCATCGACACGAGGCATACTATCTAACCCTATTCGATTGCGCACTTGGACAAGTGTACGAC (1.995)
300	0	1	100	ATGGTATGGTGACGGGCGCCCCTACGAAGCGAGTTCCAATTTGCGTCCGGCGGATACTAGTCGCGGCGTGCCGAAGAAACGATGCGGTTTAGATTGATGTCTAATTATTCGATTTTAGGAGACTTTGATTATTGGTCCTCATAAATGCCTCCCTTCGCAAACTCTGGGAGCTAAGTCCTAAATCGCGAATAACTCAGAAAAAAGCACTCTTTGATACTGGATAACGAGAGCGGTCTCATGAAATCGAATAAAGTTCAATCGTAGACCGACCTGGGAGTACCCAGCCCCATTGAAGGACGG (1.996)
60	0	10	10	AGGATGGCGGAGACGTTAATGACCTCCTTCCTCCCAGTCTTTCTATACAAGAGAGGCAGG (1.998)
60	0	10	100	ACGCCTGTGGGTTGCGTGCAGATTAGCAAGTTCTTCCGGCGAACGACAATCCTCCTTGTG (1.981)
60	1	1	100	ATGTCTCGAGAAAAGTCAATTTCGGGCAGTTCCATTGGGTGACCAATACTCGCGATCTAG (1.995)
99	1	1	100	ATGGGGTAGCTTCGGTTTCGAGAATGATAGATTATGCGGAGGCCGTAAATTAAACTGGGCGTGCGATATTCAGTCGGACTGCTTGCTCACTCGGCCTAG (1.975)
150	1	1	100	ATGTAATCCTACATGGCTTTCAGACATGGTCGCCCCCCACAATAGTTCAATTCATGCAGCATGCGATTTGGATATGAGTTGCGTCTAAGTCGTTACAGCGGAGCAGGTGTAATTCTGAGGGTTGGGTTGAGTACTATTATTGTTGGATAG (1.973)
300	1	1	100	ATGTGAACTGCAAGGACAAGAGAGTCATAACTGTGCGGTAAGGGTCCTCACTGGCAGCCCAGTTTCATGGTTTCGCTAATTCGTTTTCAGACATAAGTCTTTCTCAGACTAGTGTCTCAACCTATAATGGGGGAGCGATCAGTGTTAACGGGGTCAACGGATAGTGCATCAATGACCTTCTGGACCGGTGTTTTCAATTCATTATTTCTAACATGGTCCTATGCTGGGCAGAGGAGGAGGGGATAACCGATTAAGTACGAGTACAGTGCTCAGGGAGGGGTGTTGGACTCCATGTGGGTAA (1.982)
60	2	1	100	CGACATAATGCTTACCGCGACTCGTAGACAGGCCCCAGTTGAGGCCCTCCGTCTCGAAAG (1.967)
99	2	1	100	AGCCGCCATCCATGATGCAGGGCATGTCCGTTCGACTGTTTGGAGGGCCTTTGTGGTGAATATAAGGACTTATATACCATGCCACCCAAACCGGGAGTC (1.997)
150	2	1	100	GGTTCAGCAGGAGTTGTGGAAGTGGCGACGTTATGAAGAACAGACGACGAGGCTTTAGATAACTCATGTGGTACCGATATCGACTTTACGGCCCTTACTAGCCTACCATCAGTTCCACATGCCGCAAGCAAGTGGGAAATTTAAACGCGG (1.992)
150	2	10	100	TCTGCCCGCTCGAAGGATCTCGGGGGAGCGCCGGCACCCGTCGGTTTTTGCATGTCAATAGAGTGCGGTTCCGACATAACGAATGACCTCGACATGTTCCTTTAGAGACGACCGTGGCACTCCATAACCGGAGAGTATTCTGCAATTGTG (1.992)
300	2	1	100	AGTTTCCACTATTCTTACGCACTACAGATCCGCCTTTCGATCTGCTCGCCTTGTTTTTCCCGTTACTCAATTTTGGCACGGGTGTGCTAATAGCCATTTAGTACTCATTTGCTTTAACCCACACAGTCTTGCACACTATCAGGGTTACCTTTCGAAAGCTGCCGGCAGTGAAGACTTGACGATTCCTGAATAACACTCTCCCAGGACTTTATGGCTGCACATCAAAAAGTTACAGCACGGGCTGTAGCTCGCCCACAGGGCCCTCAGCTTAGCCCGGTCGCGTCGAATGTCCCAGAGAGC (1.985)

References

Adami, C. 2004. Information theory in molecular biology. *Physics of Life Reviews*. 1: 3-22.

Andersson, D.I., Jerlström-Hultqvist, J., Näsvall, J. 2015. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harbour Perspectives in Biology*. 7: a017996.

Kauffman, S.A. & Levin, S. Towards a General Theory of Adaptive Walks on Rugged Landscapes. *Journal of Theoretical Biology*. 128: 11-45.

Kauffman, S.A. & Weinberger, E.D. 1989. The NK Model of Rugged Fitness Landscapes And Its Application to Maturation of the Immune Response. *Journal of Theoretical Biology*. 141: 211-45.

Schlötterer, C. 2015. Genes from scratch – the evolutionary fate of *de novo* genes. *Trends in Genetics*. 31(4): 215-19.

Whitlock, M.C. & Bürger, R. 2004. Fixation of New Mutations in Small Populations. In: *Evolutionary Conservation Biology*. R. Ferrière, U. Dieckmann, & D. Couvet, Eds. Cambridge: Cambridge University Press. 155–170. Cambridge University Press.