# Nonbinary Representations in the NK and NKCS Models

**Larry Bull**
*Department of Computer Science and Creative Technologies*
*University of the West of England*
*Frenchay*
*Bristol, BS16 1QY, UK*

The NK model has been used widely to explore aspects of natural evolution and complex systems. Traditionally, the model has used a binary representation scheme. This paper introduces a modified form of the NK model through which to systematically explore the effects of discrete, nonbinary representations on evolution over rugged fitness landscapes. Results suggest the basic properties of the original model remain but changes are seen in walk lengths to optima and the sensitivity to mutation rates, in particular. The variation to the case of coupled fitness landscapes, the NKCS model, is also extended in the same way. Again, similarities and differences to the binary case are found.

*Keywords*: coevolution; evolution; fitness landscapes; mutation

## 1. Introduction

Kauffman and Levin [1] introduced the NK model to allow the systematic study of various aspects of organisms evolving on rugged fitness landscapes, and it has since been applied to many aspects of natural evolution (e.g., see [2]). Given its abstract nature, the model has also been used widely within complexity science, particularly around organization and management studies (e.g., see [3]). The NK model was later extended, in what is termed the NKCS model [4], to explore coevolution, that is, the evolutionary dynamics of ecosystems containing multiple species. Again, versions of the model have been applied to non-biological systems, particularly spatially extended versions, for example, in "patches" to receiver-based communication optimization [5].

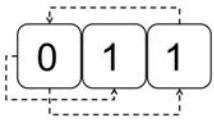
In the vast majority of known cases, the underlying representation in the model is binary. This is clearly a significant simplification for most systems of interest. Perhaps most markedly, natural biological systems use an underlying quaternary representation, of course. That is, both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are made up of four nucleobases. Moreover, synthetic biologists have recently created four new bases (e.g., [6]), opening the possibility of

https://doi.org/10.25088/ComplexSystems.31.1.87

up to octal representations in DNA in the future. This paper introduces a new parameter to the NK and NKCS models that enables the systematic exploration of the effects of altering the size of the alphabet $A$ of the underlying representation. Results suggest that a number of the basic properties of the original binary models remain, while aspects such as the time taken to reach optima, the sensitivity to the number of mutations experienced and the alphabet used by coevolving partners can significantly vary behavior.

## 2. The NK Model

In the standard NK model, the features of the fitness landscapes are specified by two parameters: $N$, the number of genes in a genome; and $K$, the number of genes that have an effect upon the fitness contribution of each gene. Hence increasing $K$ with respect to $N$ increases the epistatic linkage, increasing the ruggedness of the fitness landscape. The increase in epistasis increases the number of optima, increases the steepness of their sides and decreases their correlation (see [7]). As noted earlier, genes are traditionally from a binary alphabet $A = 2$. The model assumes all intragenome interactions are so complex that it is only appropriate to assign random values to their effects on fitness. For each of the possible $K$ interactions, a table of $A^{(K+1)}$ fitnesses is created for each gene with all entries in the range 0.0 to 1.0, such that there is one fitness for each combination of traits (Figure 1). The fitness contribution of each gene is found from its table. These fitnesses are then summed and normalized by $N$ to give the selective fitness of the total genome.




| $K_0$ | $K_1$ | Fitness |
|---|---|---|
| 0 | 0 | 0.76 |
| **0** | **1** | **0.23** |
| 1 | 0 | 0.56 |
| 1 | 1 | 0.98 |

**Figure 1.** An example standard NK model. Here $N = 3$, $K = 1$, $A = 2$, showing how the fitness contribution of each gene depends on $K$ random genes (left). Therefore there are $A^{(K+1)}$ possible allele combinations per gene, each of which is assigned a random fitness. Each gene of the genome has such a table created for it (right, left gene shown). Total fitness is the normalized sum of these values.

Kauffman [7] used a mutation-based hill-climbing algorithm, where the single point in the fitness space is said to represent a converged species, to examine the properties and evolutionary dynamics

of the NK model. That is, the population is of size one and a species evolves by making a random change to one randomly chosen gene per generation. A mutation here means that a new value in $A$ is chosen at random to replace the current value. The "population" is said to move to the genetic configuration of the mutated individual if its fitness is greater than the fitness of the current individual; the rate of supply of mutants is seen as slow compared to the actions of selection. Ties are broken at random. Figure 2 shows example results. All results reported in this paper are the average of 10 runs (random start points) on each of 10 NK functions, that is, 100 runs for 20 000 generations. Here $0 \leq K \leq 7$, for $N = 10$ and $N = 100$.

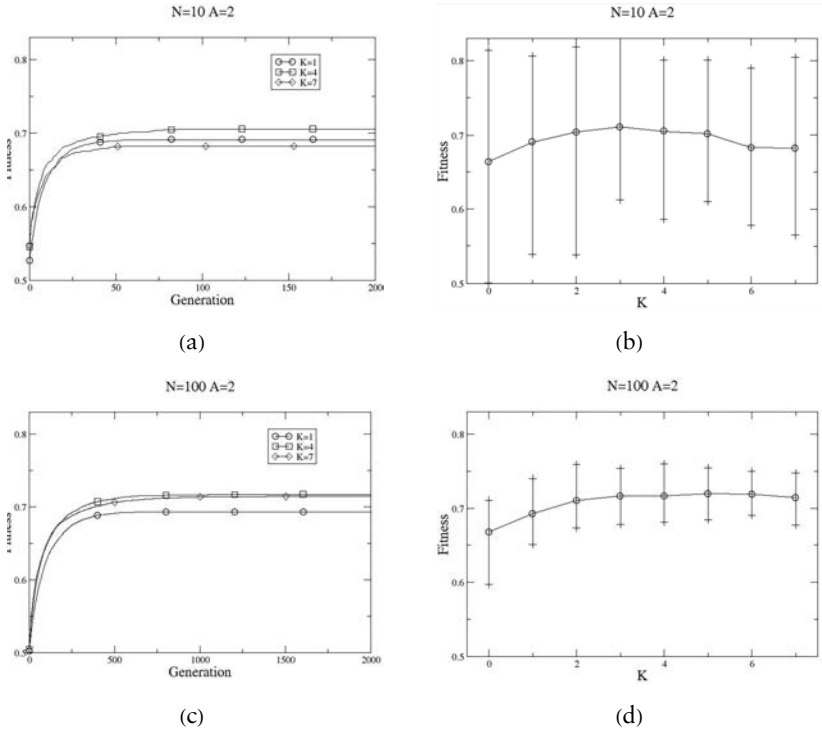



**Figure 2.** Typical behavior and fitnesses reached after 20 000 generations on NK landscapes of varying ruggedness $K$ and length $N$ with binary genes $A = 2$. Error bars show min and max values.

Figure 2 shows examples of the general properties of adaptation on such rugged fitness landscapes identified by Kauffman (e.g., [7]), including a "complexity catastrophe" as $K$ tends to $N$. When $K = 0$, all genes make an independent contribution to the overall fitness and, since fitness values are drawn at random between 0.0 and 1.0, order

statistics show the average value of the fit allele should be 0.66. Hence a single global optimum exists in the landscape of fitness 0.66. At low levels of $K$ ($0 < K < 8$), the landscape buckles up and becomes more rugged, with an increasing number of peaks at higher fitness levels, regardless of $N$. Thereafter the increasing complexity of constraints between genes means the height of peaks typically found begins to fall as $K$ increases relative to $N$: for large $N$, the central limit theorem suggests reachable optima will have a mean fitness of 0.5 as $K$ tends to $N$. Figure 2 shows how the optima found when $K > 6$ are significantly lower for $N = 10$ compared to those for $N = 100$ (T-test, $p < 0.05$).

As described, in the traditional NK model each of the $N$ elements is seen as a gene with one of two possible values, that is, $A = 2$. The size of the alphabet can also be varied; for example, with $A = 4$, each of the $N$ elements can be seen as representing the (transcribed) nucleobase values of DNA. Figure 3 shows the effects of doubling $A$ in this way over the same parameter ranges used in Figure 2. As can be seen,

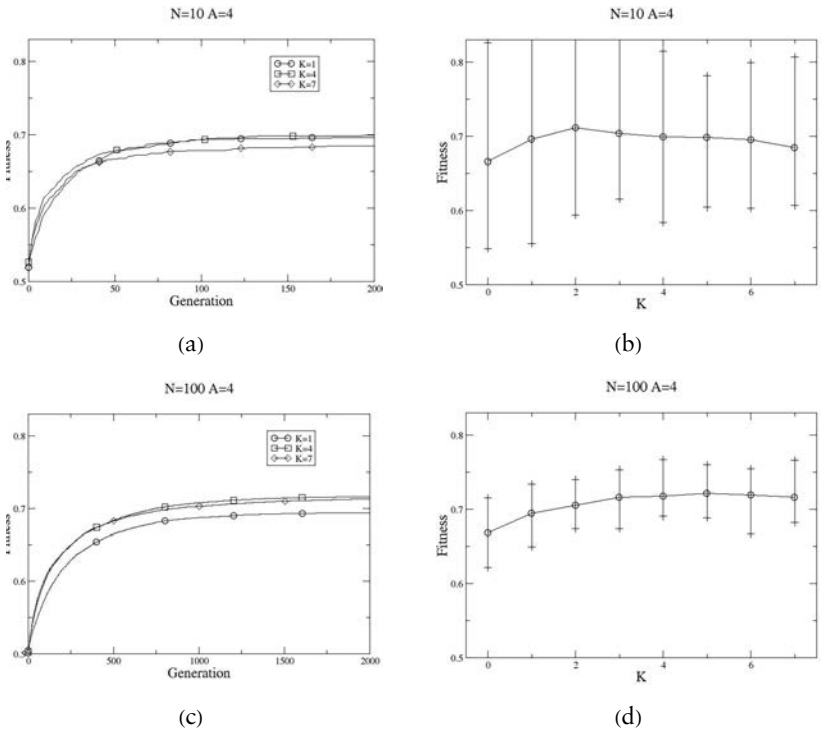



**Figure 3.** Typical behavior and fitnesses reached after 20 000 generations on NK landscapes of varying ruggedness $K$ and length $N$ with a quaternary alphabet, that is, $A = 4$.

the same well-established general properties and behavior mentioned previously are again observed. Figure 4 shows examples of how this remains the case for $2 \leq A \leq 16$, with fitnesses not significantly changed (T-test, $p \geq 0.05$).

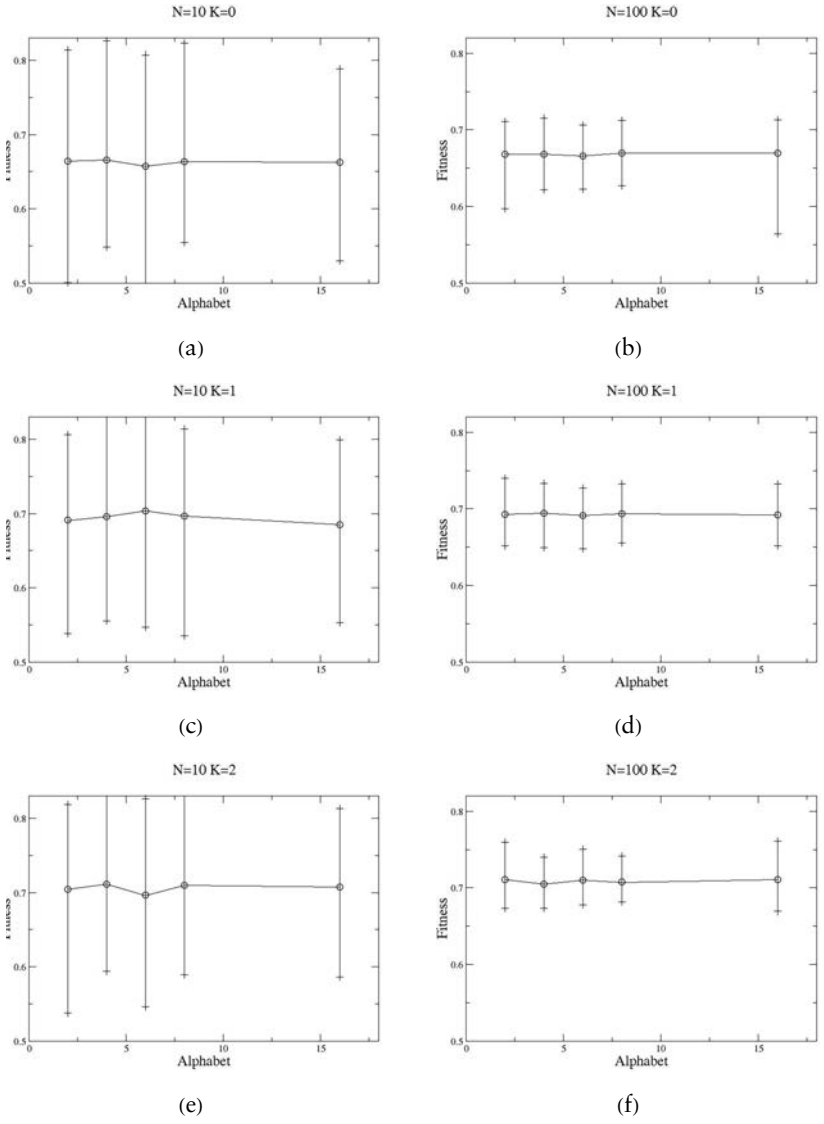



**Figure 4.** Typical behavior and fitnesses reached after 20 000 generations on NK landscapes of varying ruggedness $K$ and length $N$ with alphabets of different sizes $A$.

While the general properties of the NK model appear to be maintained with larger alphabets, it is clear by comparing Figures 2 and 3 (left) that the time taken to reach optima increases with $A$. Figure 5 shows how the walk length increases with increasing $A$ for $2 \leq A \leq 16$. However, as can be seen, despite the representation capacity increasing exponentially, the average walk length increases roughly linearly with $A$.

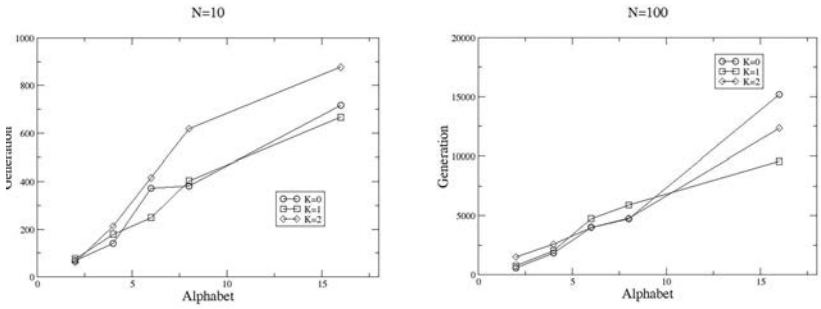



**Figure 5**. Typical walk lengths on landscapes of varying ruggedness $K$ and length $N$ for various alphabet sizes $A$.

As has been noted elsewhere, an increase in the alphabet would potentially have increased the mutation rate early in evolution (e.g., [8]). One way to consider this effect while maintaining the single mutation point scheme used earlier is to compare the behavior of individuals whose length is altered in a way proportional to their alphabet such that $A^N$ is constant. Figure 6 shows results where lengths are adjusted using $A = 4$ and $N = 10$ or 100 as the baseline. Hence the equivalent case with $A = 2$ has $N = 20$ or 200, and with $A = 16$ has $N = 5$ or 50. No significant difference in fitnesses is seen (T-test, $p \geq 0.05$). Alternatively, as shown in Figure 7, the number of mutations per offspring $M$ can be increased proportionally to $A$, here $M = A/2$. Perhaps coincidentally, $A = 4$ is always either the single fittest alphabet or in the set of fittest alphabets, with $A = 16$ almost always the opposite. More specifically, all alphabets perform equally well when $K = 0$, regardless of $N$, except $A = 16$ is significantly worse when $N = 10$ (T-test, $p < 0.05$). When $K = 1$, with $N = 10$, the equally fit alphabets are $A = 4, 6$, reducing to $A = 4$ with $N = 100$. When $K = 2$, with $N = 10$, the fittest alphabets are $A = 4, 6, 8$, increasing to $A = 4, 6, 8, 16$ with $N = 100$.

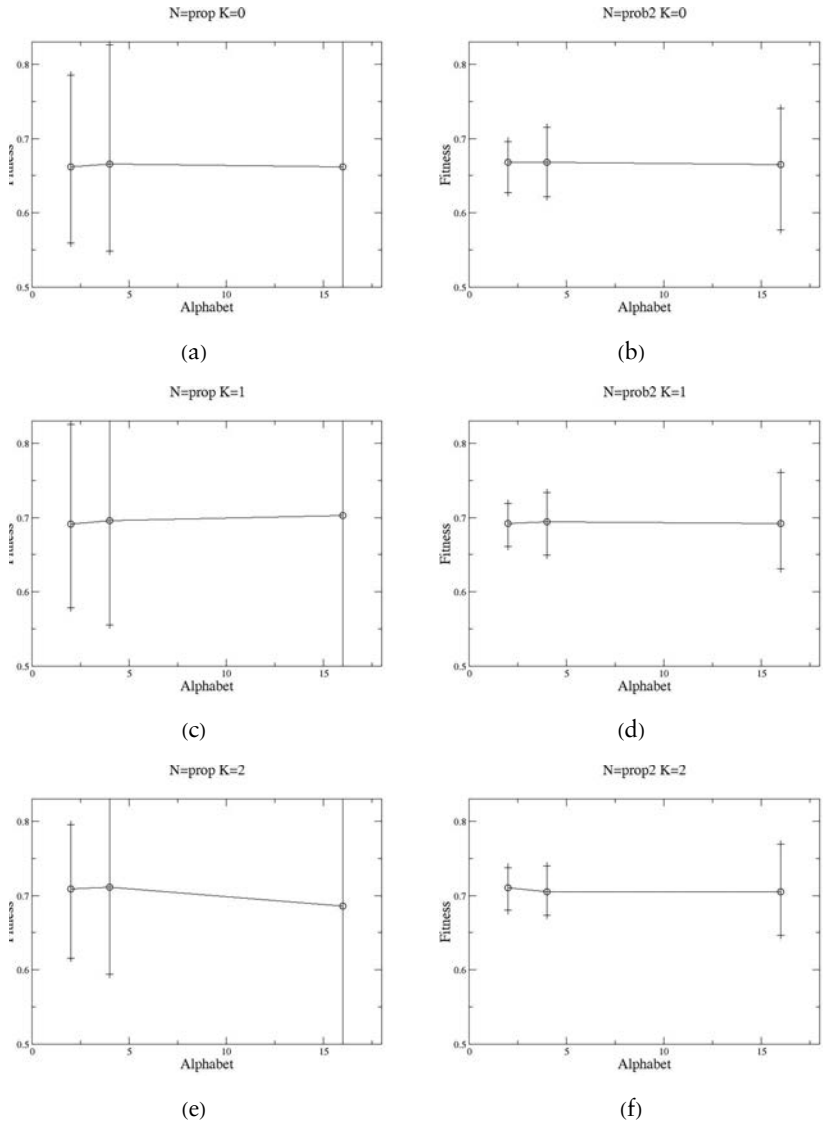



**Figure 6.** Typical behavior and the fitness reached after 20 000 generations on landscapes of varying ruggedness $K$ and alphabets $A$. Length $N$ is decreased proportional to alphabet size, starting with $A = 2$ and $N = 20$ (left column) and $N = 200$ (right column).

https://doi.org/10.25088/ComplexSystems.31.1.87

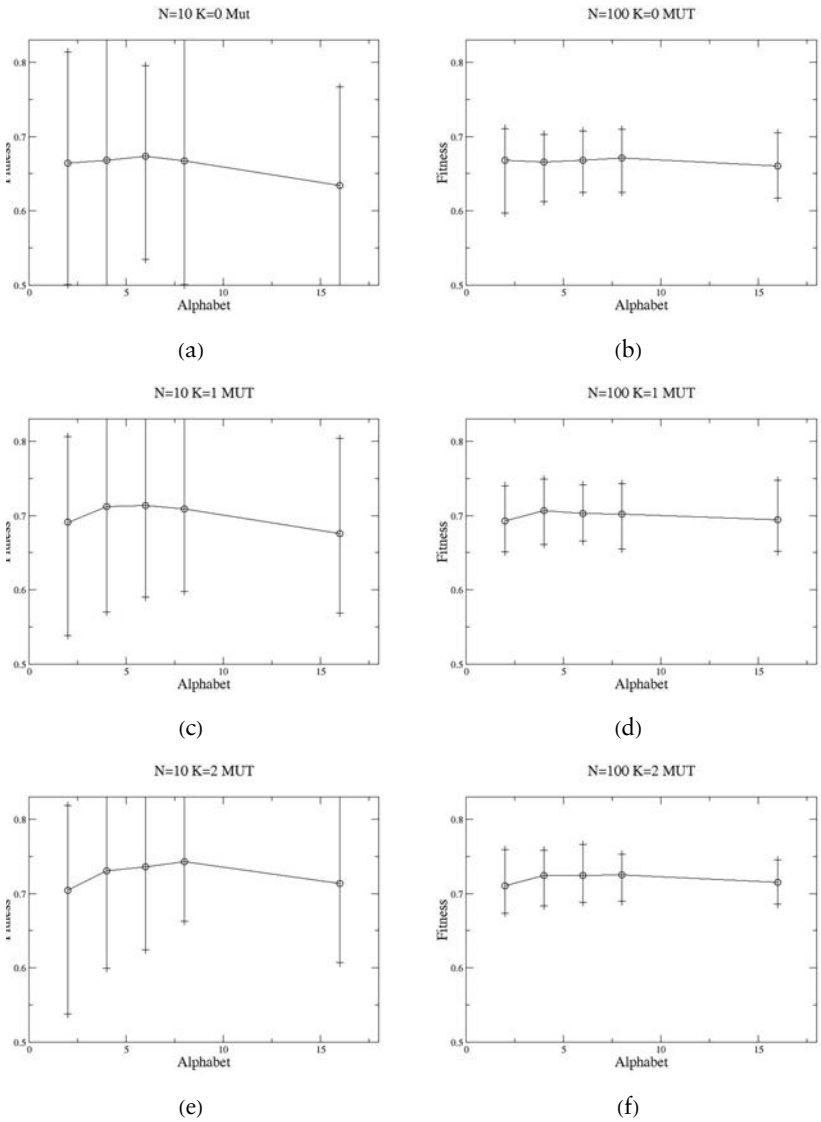



**Figure 7.** Typical behavior and the fitness reached after 20 000 generations on landscapes of varying ruggedness $K$ and length $N$ with alphabets of different sizes $A$ and $A/2$ mutations per reproduction event.

Figure 8 further shows the effects of varying the number of mutation points in an offspring; here $0 < M \leq 5$ for various $A$. With $N = 100$, there is very little effect from varying $M$. However, with $N = 10$, fitnesses are typically highest when $M \geq 3$, with the notable

exception of $A = 2$, where fitnesses become significantly lower under such conditions when $K < 2$ (T-test, $p < 0.05$). That is, for low $K$ and high $M$, higher values of $A$ are beneficial.

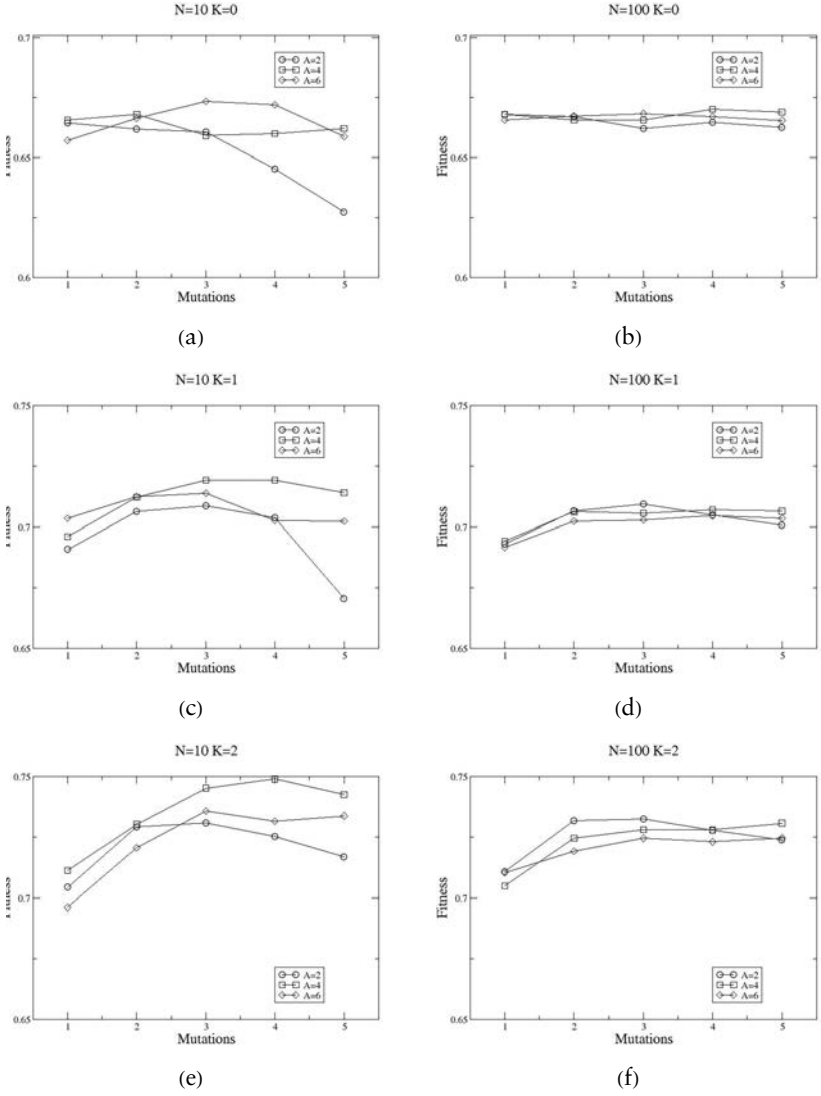
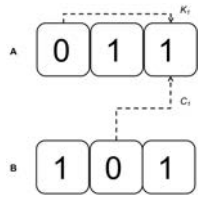



**Figure 8.** Typical behavior and the fitness reached after 20 000 generations on landscapes of varying ruggedness $K$ and length $N$ with alphabets of different sizes $A$ and $A/2$ mutations per reproduction event.

## 3. The NKCS Model

Kauffman and Johnsen [4] subsequently introduced the NKCS model to enable the study of various aspects of coevolution. At an abstract level, coevolution can be considered as the coupling together of the fitness landscapes of the interacting species. Hence the adaptive moves made by one species in its fitness landscape cause deformations in the fitness landscapes of its coupled partners. In this extension to the NK model, each gene is also said to depend upon $C$ randomly chosen genes in each of the other $S$ species with which it interacts. The adaptive moves by one species may deform the fitness landscape(s) of its partner(s). Altering $C$, with respect to $N$, changes how dramatically adaptive moves by each species deform the landscape(s) of its partner(s). Again, for each of the possible $K + (SxC)$ interactions, a table of $A^{(K+(SxC)+1)}$ fitnesses is created for each gene, with all entries in the range 0.0 to 1.0, such that there is one fitness for each combination of traits. Such tables are created for each species (Figure 9, the reader is referred to [7] for full details).




| $K_0$ $K_1$ $C_1$ | Fitness |
|---|---|
| 0 0 0 | 0.76 |
| 0 1 0 | 0.23 |
| **1 0 0** | **0.56** |
| 1 1 0 | 0.98 |
| 0 0 1 | 0.43 |
| 0 1 1 | 0.32 |
| 1 0 1 | 0.65 |
| 1 1 1 | 0.03 |

**Figure 9.** An example standard NKCS model. Each gene is connected to $K$ randomly chosen local genes and to $C$ randomly chosen genes in each of the $S$ other species. A random fitness is assigned to each possible set of combinations of genes. These are normalized by $N$ to give the fitness of the genome. Connections and table shown for one gene in one species for clarity.

Figure 10 shows example results for one of two coevolving species where the parameters of each are the same and hence behavior is symmetrical. All results reported in this paper are the average of 10 runs (random start points) on each of 10 NKCS functions, that is, 100 runs for 20 000 generations. Here $0 \le K \le 4$, $1 \le C \le 3$, for $N = 10$ and $N = 100$.

When $C = 1$, Figure 10 shows examples of the general properties of adaptation on such fitness landscapes identified in the NK model, that is, where $C = 0$. Figure 10 further shows how increasing the degree of connectedness $C$ between the two landscapes causes fitness levels to fall significantly (T-test, $p < 0.05$) when $C \ge K$ for $N = 10$. That is, as $K$ tends to $N$, a high number of peaks of similar height typically exist in each of the fitness landscapes and so the effects of

switching between them under the influence of $C$ is reduced since each landscape is very similar. Note this change in behavior around $C = K$ was suggested as significant in [7], where $N = 24$ only was used throughout. However, Figure 10 also shows how with $N = 100$ fitness always falls significantly with increasing $C$ (T-test, $p < 0.05$), regardless of $K$. That is, an increase in the degrees of freedom in movement over a larger fitness landscape is generally more disruptive to the search process when the landscapes are coupled.

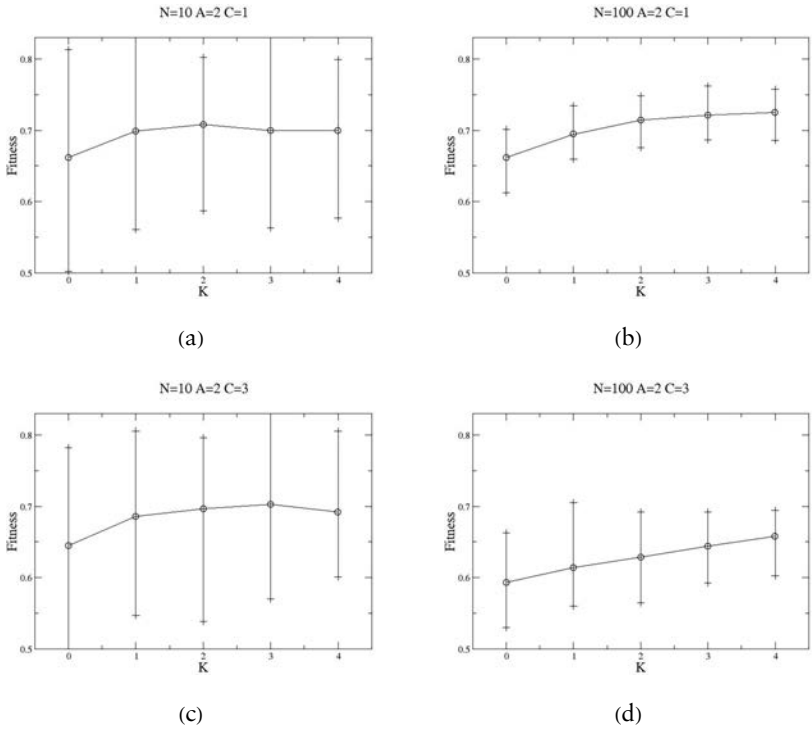



**Figure 10.** The fitness reached after 20 000 generations $M = 1$ on landscapes of varying ruggedness $K$, coupling $C$ and length $N$, with binary genes $A = 2$.

As in the traditional NK model, the size of the alphabet in the NKCS model can also be varied. Figure 11 shows the effects of doubling $A$ over the same parameter ranges used in Figure 10. As can be seen, the same general behavior with $A = 2$ is again observed with $A = 4$.

Using $N = 24$ and $A = 2$, Kauffman and Johnsen [4] explored varying the number of mutations per offspring in the NKCS model, with $0 < M \leq 24$. For various $K$ and $C = 1$, they report fitnesses are typically highest for $2 \leq M \leq 4$, and with increasing $C$ the optimal

mutation rate $M$ typically decreases. Figure 12 shows how increasing the alphabet to $A = 4$, with $N = 10$, various $K$ and $C = 1$, has no significant effect on coevolution. When $C = 3$, the optimal mutation rate typically decreases with $K = 3$ but fitnesses drop significantly compared to $A = 2$ when $M > 2$ and $K = 1$ (T-test, $p < 0.05$). With $N = 100$, fitnesses with $A = 4$ are almost always significantly higher than with $A = 2$, and fitnesses are typically higher with lower $M$. Hence such coevolutionary systems appear more sensitive to the mutation rate as the size of the alphabet increases compared to the standard NK model.

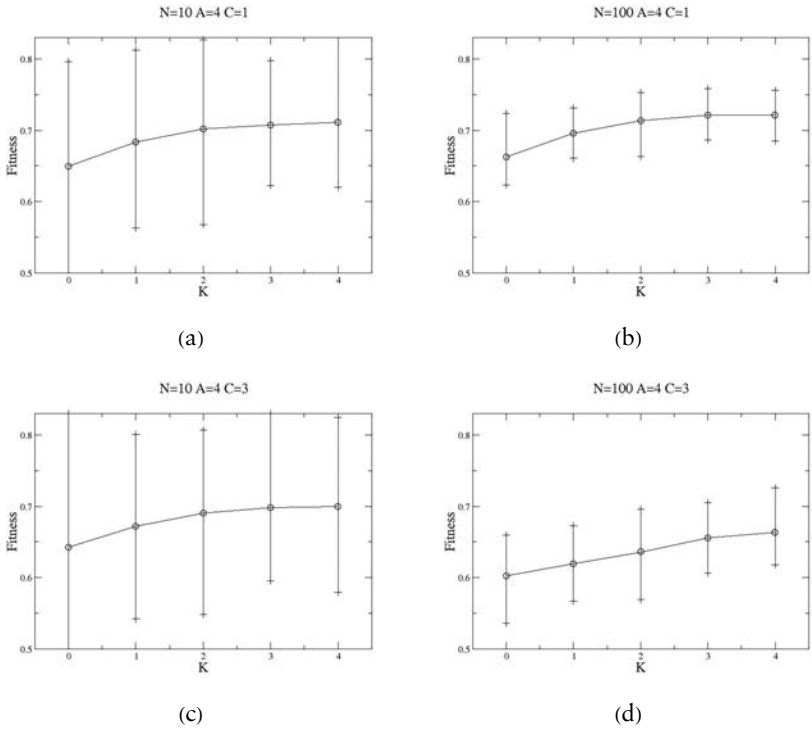



**Figure 11**. The fitness reached after 20 000 generations $M = 1$ on landscapes of varying ruggedness $K$, coupling $C$ and length $N$, with quaternary genes $A = 4$.

The standard NKCS model assumes symmetry among all species, with all experiencing the same $N$, $K$ and $C$. Kauffman and Johnsen [4] explored varying the $K$ value of two coevolving species, finding that a high-$K$ partner increased the relative fitness of a low-$K$ partner, for example. Similarly, the standard model assumes all species evolve

at the same rate. Bull et al. [9] introduced a new parameter $R$ to specify the relative rate of reproduction, showing how increasing $R$, with high $C$, can result in a rapid drop in fitness of the slower partner(s), for example.

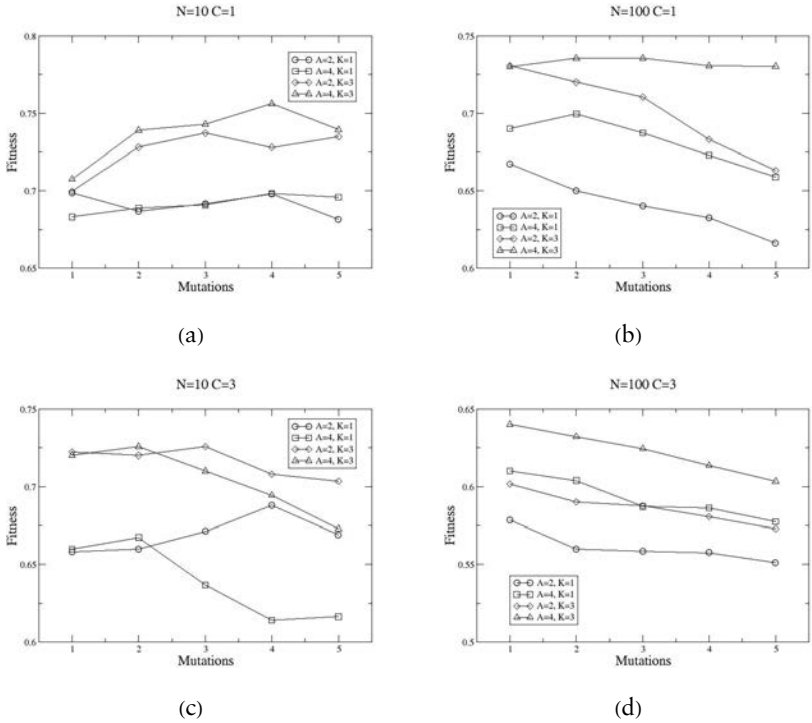



**Figure 12.** Typical behavior and the fitness reached after 20 000 generations on landscapes of varying ruggedness $K$, coupling $C$ and length $N$, with a different number of mutations in a genome per generation $M$.

Figure 13 shows example results from coevolving species using different alphabets, one with $A = 2$ and one with $A = 4$. With $N = 10$, regardless of $C$, there is no significant difference in behavior for either species in comparison to coevolving with a partner species with the same $A$ (not shown). The same is true for $N = 100$ when $C = 1$ (not shown). However, when $C = 3$, the $A = 4$ species experiences a significant drop in fitness, whereas the $A = 2$ species experiences a significant increase in fitness (T-test, $p < 0.05$). Moreover, an $A = 2$ species coevolved with an $A = 6$ partner does even better than with an $A = 4$ partner (not shown). It is here suggested that being partnered with a species with a longer typical walk length to an optimum, due to a higher $A$, can be beneficial to the species with the shorter walk length.

https://doi.org/10.25088/ComplexSystems.31.1.87

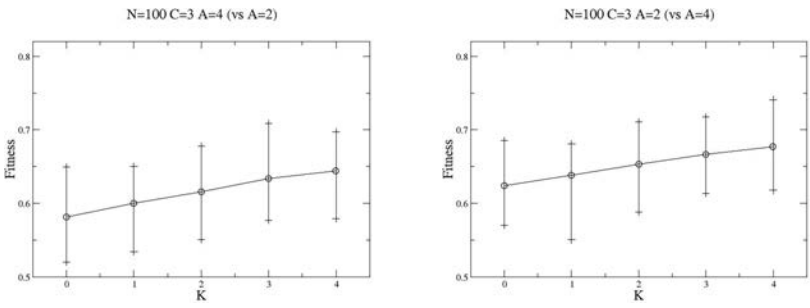



**Figure 13.** Typical behavior and the fitness reached after 20 000 generations on larger landscapes $N = 100$ of varying ruggedness $K$, with higher coupling $C = 3$ between an $A = 4$ (left) and $A = 2$ (right) species.

## 4. Conclusion

The well-known NK model assumes a discrete representation scheme for its genes/component parts, where these have almost exclusively been binary. This paper has explored the effects of increasing the size of the alphabet within the model, finding that the general properties of the landscapes are seemingly preserved. That is, landscapes become increasingly rugged with increasing $K$ and experience increasingly similarly sized optima as $K$ approaches $N$. Similarly, the extended version of the NK model to the case of coupled adapting systems, the NKCS model, again appears to maintain its general properties as the alphabet is increased from the binary case. Differences were seen in terms of the typical walk length to optima and the sensitivity of alphabets to different mutation rates.

That the alphabet of biological systems consists of two base pairs, that is, $A = 4$, has been suggested as a frozen property from an originally RNA world where it was potentially optimal in terms of catalytic activity (e.g., [8]). That is, despite the later shift to the improved catalytic power of proteins, the alphabet did not subsequently increase. In the NK model here, for shorter genome lengths and low epistasis, a quaternary alphabet was shown to be robust to an increasing amount of mutation (Figure 8). Such small, simple systems with low replication accuracy can be envisaged as typical early in evolution. Specifically, when $N = 10$, $K = 0$ or 1, and $M = 4$, $A = 4$ was shown to be beneficial over both $A = 2$ and $A = 6$ (T-test, $p < 0.05$). $A = 4$ was also shown to be beneficial over both $A = 2$ and $A = 6$ when $M = 5$ and $K = 1$ (T-test, $p < 0.05$), and over $A = 2$ when

$M = 5$ and $K = 0$ (T-test, $p < 0.05$), with no difference over $A = 6$ (T-test, $p \geq 0.05$). For larger $N$, such benefits disappear but reappear when an explicit degree of coupling to other adapting entities is included, that is, in the NKCS model, where $A = 4$ is again seen to be generally significantly beneficial over $A = 2$ (Figure 11). There is no significant difference with $A = 4$ compared to $A = 6$ (T-test, $p \geq 0.05$, not shown). Thus, the results here suggest a relatively widespread selective advantage to a low—but not the lowest—alphabet.

## References

[1] S. A. Kauffman and S. Levin, "Towards a General Theory of Adaptive Walks on Rugged Landscapes," *Journal of Theoretical Biology*, **128**(1), 1987 pp. 11–45. doi:10.1016/S0022-5193(87)80029-2.

[2] L. Bull, *The Evolution of Complexity: Simple Simulations of Major Innovations*, London: Springer, 2020.

[3] B. McKelvey, M. Li, H. Xu and R. Vidgen, "Re-thinking Kauffman's NK Fitness Landscape: From Artifact and Groupthink to Weak-Tie Effects," *Human Systems Management*, **32**, 2013 pp. 17–42. doi:10.3233/HSM-130782.

[4] S. A. Kauffman and S. Johnsen, "Co-evolution to the Edge of Chaos: Coupled Fitness Landscapes, Poised States and Co-evolutionary Avalanches," in *Artificial Life II: Held February 1990 in Santa Fe, New Mexico* (C. G. Langton, C. Taylor, J. D. Farmer and S. Rasmussen, eds.), Redwood City, CA: Addison-Wesley, 1992 pp. 325–370.

[5] S. A. Kauffman, *At Home in the Universe: The Search for Laws of Complexity*, New York: Oxford University Press, 1995.

[6] H. Hoshika, N. Leal, M. J. Kim, M. S. Kim, N. B. Karalkar, H. J. Kim, A. M. Bates, et al., "Hachimoji DNA and RNA: A Genetic System with Eight Building Blocks," *Science*, **363**(6429), 2019 pp. 884–887. doi:10.1126/science.aat0971.

[7] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, New York: Oxford University Press, 1993.

[8] E. Szathmáry, "Four Letters in the Genetic Alphabet: A Frozen Evolutionary Optimum?," *Proceedings of the Royal Society of London B*, **245**(1313), 1991 pp. 91–99. doi:10.1098/rspb.1991.0093.

[9] L. Bull, O. Holland and S. Blackmore, "On Meme-Gene Coevolution," *Artificial Life*, **6**(3), 2000 pp. 227–235. doi:10.1162/106454600568852.