

Research Article

Selective Phenome Growth Adapted *NK* Model: A Novel Landscape to Represent Aptamer Ligand Binding

Andrew Brian Kinghorn and Julian Alexander Tanner

School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong

Correspondence should be addressed to Andrew Brian Kinghorn; kinghorn@hku.hk

Received 16 January 2017; Revised 10 May 2017; Accepted 23 May 2017; Published 24 July 2017

Academic Editor: Pietro De Lellis

Copyright © 2017 Andrew Brian Kinghorn and Julian Alexander Tanner. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aptamers are single-stranded oligonucleotides selected by evolutionary approaches from massive libraries with significant potential for specific molecular recognition in diagnostics and therapeutics. A complete empirical characterisation of an aptamer selection experiment is not feasible due to the vast complexity of aptamer selection. Simulation of aptamer selection has been used to characterise and optimise the selection process; however, the absence of a good model for aptamer-target binding limits this field of study. Here, we generate theoretical fitness landscapes which appear to more accurately represent aptamer-target binding. The method used to generate these landscapes, selective phenome growth, is a new approach in which phenotypic contributors are added to a genotype/phenotype interaction map sequentially in such a way so as to increase the fitness of a selected fit sequence. In this way, a landscape is built around the selected fittest sequences. Comparison to empirical aptamer microarray data shows that our theoretical fitness landscapes more accurately represent aptamer ligand binding than other theoretical models. These improved fitness landscapes have potential for the computational analysis and optimisation of other complex systems.

1. Introduction

1.1. Background. Aptamers are single-stranded nucleic acid sequences capable of specific high-affinity binding [1–4]. This makes them attractive candidates as recognition molecules in diagnostics and therapeutics. Aptamers are isolated by systematic evolution of ligands by conventional exponential enrichment (SELEX), which involves several iterative steps of incubation with target, washing away of weak binders, and PCR amplification of strong binders.

Aptamer selection is complex. Many variables such as library size, quantity of target, temperature, selection buffer, pH, degree of PCR amplification, and use of mutation or recombination diversification need to be considered. Due in part to these factors, less than 30% of classical SELEX experiments are successful in isolating aptamers with dissociation constants less than 30 nM [5]. Understanding the dynamics of the selection process is extremely important; but what does this entail? The DNA required to fully represent the number of permutations in a 75-base aptamer library would be roughly equal to the mass of the moon [6]. In

order to represent this immense sequence space, an initial SELEX library may contain up to 10^{15} molecules. A rigorous empirical analysis of anything close to this number of library members is simply not feasible.

Nevertheless, empirical analyses of smaller fractions of a DNA aptamer libraries have been undertaken. The two main empirical selection analysis techniques used are high density DNA microarrays and high-throughput sequencing (HTS). Briefly, high density microarrays can contain up to approximately 1 million features, each representing an aptamer in a library. After array incubation with fluorescent target and a washing step, the binding affinity score of all aptamers on the array can be measured by fluorescence scanning. Platt et al. [7], Knight et al. [8], and Rowe et al. [9] used microarrays to both evolve aptamers and gain insight into an aptamer binding landscape. Additionally, DNA microarray data has been applied to aptamer specificity landscapes [10], fitness landscape morphology [11], and aptamer affinity maturation [12]. In comparison, the possible sequence space coverage using HTS is much greater, with Illumina's HiSeq HTS capable of yielding sequence data for more than

70 million sequences from a single lane [13]. Using this approach the copy number of a sequence is used as a proxy for its target binding strength so that the fitness of individuals in the library pools can be estimated. Cho et al. used this HTS approach to monitor microfluidic aptamer selection rounds and gauge enrichment [14]. PCR bias may distort this copy number/binding correlation but, by using a motif based statistical framework such as MPBind [15], the binding potential of aptamers can be predicted, eliminating error from PCR bias. Although both DNA microarrays and HTS led to major breakthroughs in understanding library sequence space fitness and selection, these techniques are only capable of analysing a small fraction of a given library's sequence space. Another approach to analysing aptamer selection is via computational simulation. The challenge in simulating aptamer selection is the design of a suitable model for aptamer binding fitness.

Computational approaches to model aptamer fitness by virtue of folding include secondary structure prediction by minimum free energy [16] and inverse folding [17]. These approaches can be computationally expensive and while being excellent models for folding they may not capture the higher complexity of molecular binding. Hoinka et al. coded a program to simulate the aptamer selection process called "AptaSim" [18]. The binding model used chose aptamer affinities at random without relevance of sequence. While AptaSim was an important step forward in simulating selection enrichment and mutation copy number, AptaSim cannot appropriately represent heritability or represent binding correlation between related sequences required for the study of genetic systems. Oh et al. used a string matching function as a binding fitness model to simulate aptamer selection [19]. This model does include heritability and binding correlation between related sequences, but as only close range epistasis is possible by using one "optimal solution" aptamer the landscape is cone shaped and would often be unrepresentative of a true aptamer binding landscape.

Kauffman's NK model is a robust mathematical model, related to study of autocatalytic sets, which serves as an objective function relating genotypic sequence to phenotypic fitness [20]. Derivations of the original NK model have been used to describe complex interacting systems in areas as diverse as immunology [21], evolutionary biology [22], and economics [23]. The NK model describes a fitness landscape whose size is determined by the number of components in system (N) and the ruggedness of the landscape can be tuned using the degree of interaction of these components (K). This system is perhaps best described when used to represent problems in evolutionary biology as originally intended by Kauffman. A population of genomes where each contains N genes are given fitness values based on the sum of fitness contributions from each of their genes. The fitness contribution of each gene is determined by its interaction with K other genes within its own genome. The interacting genes can be positioned sequentially, randomly, or by some other gene interaction pattern predetermined by an interaction map. The allelic sequence of these interacting genes is designated a fitness contribution, usually from a generated random distribution. In this way, the allelic substitution of

one interacting gene means there will be a completely different fitness contribution score for the entire collection of interacting genes. In the NK model by increasing K , the number of interactions between genes, the complexity of the system, and the ruggedness of the landscape are increased. In addition to the value of K , the position of these interactions on the interaction map is of great importance to the fitness landscape.

Typically the NK model is used as a scoring system for a population of genomes which can evolve via diversifications such as mutation or recombination. In this way the NK model is an objective function for a complex system. As mentioned earlier the NK model can be adapted to many other areas of study. In this paper we use the NK model to represent binding of an aptamer to a target analyte. In this representation N would be equal to length of the aptamer in the library and K would be equal to the interactions of bases within each aptamer. Many modifications to the original NK model have been made, some to optimise the model for a given research area. Herein we will describe some modifications to the NK model which are aimed at optimising the model to represent binding of an aptamer to a target analyte. The NK model was believed to resemble molecular fitness landscapes similar to the binding of an aptamer to an analyte [24]. In the NK model mutational additivity usually holds for non-interacting positions in sequences. This mutational additivity is biologically accurate as has been demonstrated for several proteins [25–33].

Wedge et al. used an NK model for the simulation of protein directed evolution (DE) [36], a similar field to aptamer selection. Binary strings of length 40 and 100 were used with random epistatic interactions varying from $K = 0$ to 10. Genetic algorithms utilising mutation, crossover, different library sizes, and selection pressures were simulated and compared to deduce general rules for protein directed evolution, which are of great use to DE experiments. As noted in this study, the "No Free Lunch Theorem" (NFL) [37] establishes that all search algorithms perform exactly the same when averaged over all possible problems. This infers that, for an optimisation algorithm, any elevated performance in one class of problem is exactly paid for in performance in another class. If there is discrepancy between a real life system and a model used to describe it, any elevated performance in optimisation using simulation of the model is exactly paid for in performance for the real life system. This illustrates the need for an accurate model when using simulation results to improve empirical ligand selection experiments.

Despite this biological accuracy in regard to mutational additivity, the classical NK model may have limitations in representing some biological systems. The NK model's greatest utility is that ruggedness can be tuned using the epistasis variable K . However, this epistasis is quite uniform throughout the sequence. For some biological applications, a higher amount of epistasis is desirable. As K increases the landscape tends to become more multi-peaked and rugged, to the point where it is too chaotic to allow adaptation. Kauffman refers to this phenomena as the "complexity catastrophe" [38]. Kauffman goes further to say "the complexity catastrophe is averted in the NK model for those landscapes which are sufficiently

smooth to retain high optima as N increases” [38]. Thinking along these lines provided a solution to the complexity catastrophe, creating complex landscapes which retained a smoother surface.

1.2. Constructional Selection of NK Landscapes. Altenberg developed an evolutionary approach to selecting epistatic interaction, thereby creating landscapes which were smoother than classic NK landscapes with the same degree of epistatic interaction [34]. Altenberg achieved this using selective genome growth, a type of constructional selection, to create modular interaction matrices. These selected matrices have reduced epistasis which give rise to smoother fitness landscapes [34, 39]. Selective genome growth is a process by which the genome of the fittest individual is expanded one gene at a time (Figure S1a in Supplementary Material available online at <https://doi.org/10.1155/2017/6760852>). The new gene is only kept if the fitness of a selected optimum genome is increased. In this way the probable global optima of the landscape is constructed and all other points on the landscape are relative to this optimisation. A similar method for creating landscapes was devised by Hebron et al., which uses preferential attachment growth process to add genes to a genome [40]. A problem with these two approaches, when applied to specific applications, is that due to the increasing returns of the selection system these methods attribute extremely high pleiotropy to a handful of genes (vertical lines in Figure 2(c)). This phenomenon of increasing returns of gene control is biologically appropriate and accurate for a system describing a group of genomes, but when describing the binding of an aptamer to an analyte this high aggregated pleiotropy is not biologically appropriate. Each base in an aptamer has a relatively low number of interactions due to its spatial capacity, meaning that high aggregated pleiotropy is not biologically representative for an aptamer.

Herein we have created a new model that we have termed “selective phenome growth.” Selective phenome growth is a constructional selection technique in which phenotypic contributing factors are added to a genotype-phenotype interaction map incrementally (Figure S1b) in such a way that each new phenotypic contributing factor increases the fitness of global or local optima. Additionally, comparison is made between selective phenome growth landscapes and aptamer binding landscapes.

2. Model and Methods

2.1. Selective Phenome Growth to Create a Genotype/Phenome Interaction Map. Selective phenome growth is a new method of constructing an interaction matrix one phenotypic contributor at a time. The method of representing the interaction map is the same as Altenberg’s [34], with slight modification to represent aptamers, and is as follows:

- (1) The aptamer consists of n binary valued bases that have influence over f phenotypic functions, each of which contributes a component to the total fitness.
- (2) Each base controls a subset of the f fitness components, and, in turn, each fitness component is

controlled by a subset of the n bases. This genotype-phenotype map can be represented by a matrix,

$$M = \|m_{ij}\|, \quad i = 1, \dots, n, \quad j = 1, \dots, f, \quad (1)$$

of indices $m_{ij} \in \{0, 1\}$, where $m_{ij} = 1$ indicates that base i affects fitness component j .

- (3) The columns of M , or epistasis vectors, $g_j = \|m_{ij}\|$, $i = 1, \dots, n$, give the bases controlling each fitness component j .
- (4) The rows of M , or pleiotropy vectors, $p_i = \|m_{ij}\|$, $j = 1, \dots, f$, give the fitness components controlled by each base i .
- (5) If any of the bases controlling a given fitness component is altered, the new value of the fitness component will be uncorrelated with the old one. Each fitness component φ_i is a uniform pseudo-random function of the genotype, $x \in \{0, 1\}^n$.
- (6) If a fitness component is affected by no genes, it is assumed to be zero.
- (7) The total fitness is the normalized sum of the fitness components:

$$w(x) = \frac{1}{f} \sum_{i=1}^f \varphi_i(x). \quad (2)$$

Similarly to Altenberg’s selective genome growth [34], a test sequence of length N is randomly generated. In the phenotype selection loop (Figure 1), K positions from a total of N are selected at random and added as a phenotypic contributor to the interaction matrix. If the new addition decreases the overall fitness of the test sequence it is removed; if the new addition increases or does not change the overall fitness of the test sequence it is kept in the interaction matrix. The selection loop is usually repeated until N phenotypic contributors are selected.

Selective phenome growth can be seen as randomly selecting one sequence as the fittest member of a prospective landscape and evolving the interaction map and therefore the fitness landscape around this fittest sequence. In this way phenotypic contributors are sequentially added in such a way that each increases the fitness of the selected fittest sequence.

2.2. Characterising Landscapes. Adaptive walks were used to find the two measures, accessibility of local optima and hamming distance from fittest optima similarly to Kauffman 1993 [38]. 100,000 adaptive walks from randomly selected points on each landscape were performed and the frequency each local optimum reached was recorded. The more adaptive walks terminating at an optimum, the larger the basin of attraction. In this way the accessibility of a local optimum is a measure of its basin of attraction. For the same 100,000 adaptive walks the hamming distance from the fittest optimum found for all other optima found was calculated. Hamming distance from fittest optima is a measure of peak clustering.

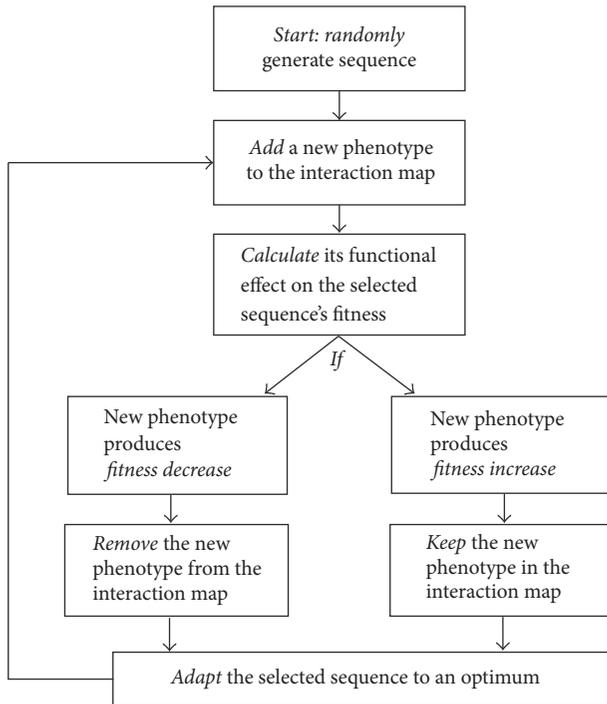


FIGURE 1: Schematic of the selective phenome growth algorithm: a flowchart representing the algorithm used in the selective phenome growth of interaction matrices. The scheme is similar to that used by Altenberg [34].

Mean path divergence (MPD) is a measure developed by Lobkovsky et al., 2013, which quantifies the degree of similarity among monotonic evolutionary trajectories with the same endpoints [11]. MPD was used to gauge the roughness of a given landscape.

2.3. Breaking the One-Gene One-Phenotypic Contributor Paradigm. Previously in NK model literature the number of phenotypic contributors has been equal to the number of genes. This rigid framework is not representative of biological systems as the number of phenotypes may be lower or may exceed the number of genes in a genome. This one phenotype per gene paradigm made sense with earlier interpretations of the NK model with local interactions, as each numbered row in an interaction matrix corresponded to a phenotypic contributor determined by its gene interacting with other genes. This one phenotype per gene paradigm where each locus always contributes to its own fitness contribution can be seen as the red diagonal lines in Figures 2(a) and 2(b).

Altenberg's genome growth constructional selected interaction maps do not have this pattern of each locus always contributing to its own fitness contribution as all interacting genes for a given phenotypic contributor are selected at random. In Altenberg's genome growth interaction maps the number of phenotypic contributors is usually predetermined and equal to the number of genes [34]; however this is not necessarily the case for our phenome selected interaction maps.

In the case of our phenome selected interaction maps the phenotypic contributors are added sequentially and so their number can be varied. This in effect varies the complexity of the landscape in such a way that the landscape can be tuned by adding or subtracting phenotypic contributors. The effect of this landscape tuning via adding or subtracting phenotypic contributors was analysed in Results and Discussion (Figure 6).

3. Results and Discussion

3.1. Comparison of Epistatic Maps. From Figure 2 the different types of interaction map can be compared. For locally distributed interactions (Figure 2(a)) the rows, or epistasis vectors, are equal to the columns, or pleiotropy vectors, which are equal to the K value plus one. For randomly distributed interactions (Figure 2(b)) the epistasis vectors are equal to the K value plus one and the pleiotropy vectors are a Poisson distribution with a mean of K .

For genome selected interactions (Figure 2(c)) the epistasis vectors are equal to the selected K values plus one which yields a mildly skewed distribution (Fisher-Pearson coefficient of 0.13). The pleiotropy vectors for genome selected interactions are heavily skewed (Fisher-Pearson coefficient of 1.90) towards a handful of loci which can be seen as the vertical lines in Figure 2(c). For phenome selected interactions (Figure 2(d)) the epistasis vectors are equal to the selected K value plus one which yields a skewed distribution (Fisher-Pearson coefficient of 0.69). The pleiotropy vectors for phenome selected are a normal distribution (Fisher-Pearson coefficient of 0.07). From Figure 2 it can be seen that phenome selected growth produces an interaction map with less aggregated pleiotropy when compared to genome selected maps.

In order to make a relevant comparison between genome selected and phenome selected landscapes the complexity of phenome selected growth interaction map was limited to be equal to that of the genome selected landscape. This was achieved by limiting the number of phenotypic contributors such that the total number of base interactions was limited to that of a previously generated genome selected interaction map. As the number of interactions a phenotype has is randomly selected, the actual number of interacting bases selected for phenome selected growth is greater than the enforced limit relating to the genome selected interaction map. Consequently, as the genome selection map was involved in 577 interactions, the phenome selected map was involved in 588 interactions.

3.2. Comparison of Theoretical Landscapes. The accessibility of local optima was used as a comparison for different landscapes. Explained briefly, 100,000 adaptive walks from randomly selected points on each landscape were performed and the number of times each local optimum was reached was recorded. The larger the basin of attraction of a local optimum is, the more adaptive walks will terminate at that optimum. In this way the accessibility of local optimum is a measure of an optimum's basin of attraction.

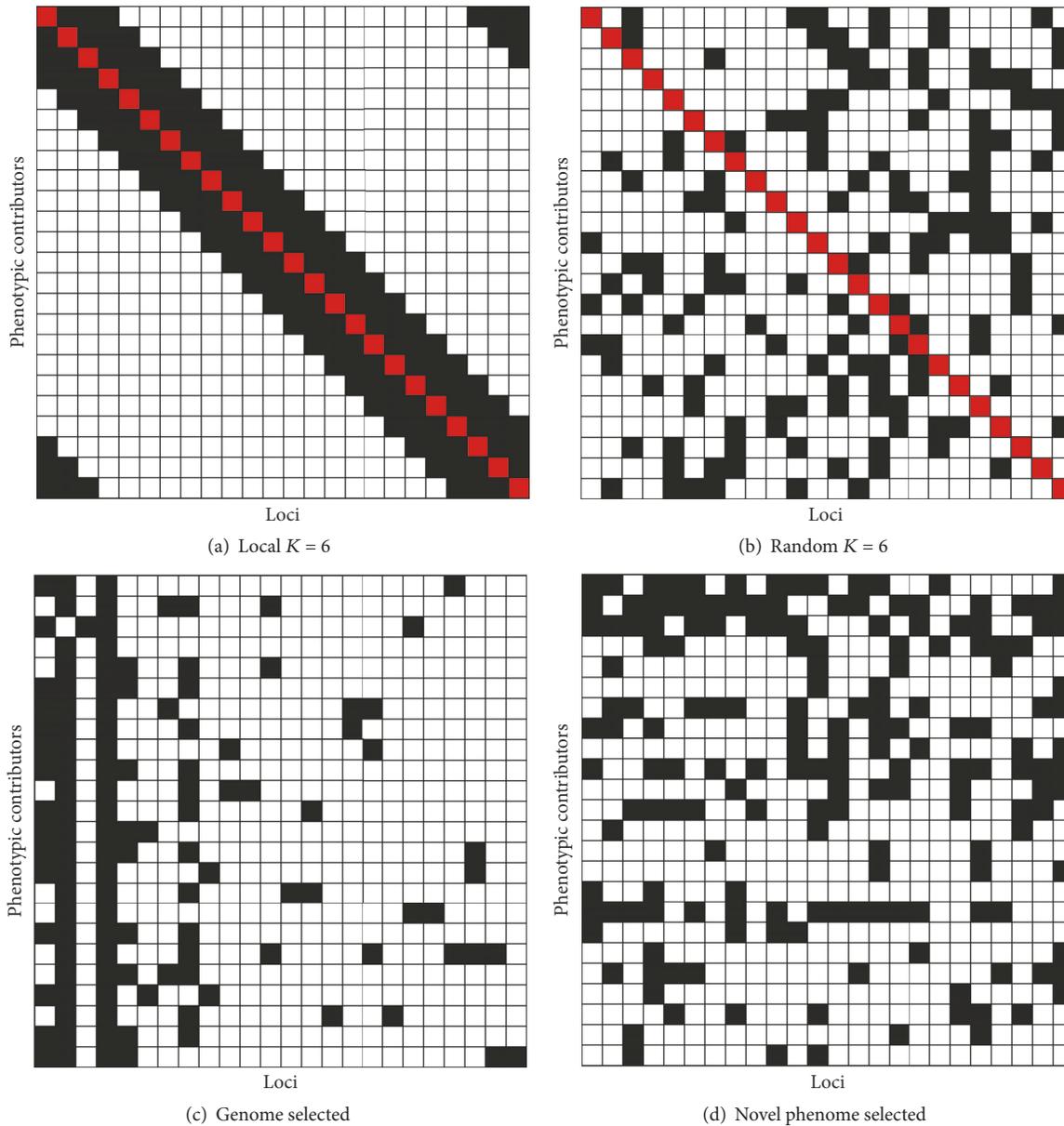


FIGURE 2: Comparison of interaction maps: in the interaction maps each row represents a phenotypic contributor with interacting loci coloured blue. The columns of each interaction map represent pleiotropy vectors while the rows represent epistasis vectors. In this context pleiotropy is a single base (dark squares) affecting multiple phenotypic contributors (rows), and epistasis is multiple bases (dark squares) affecting a single phenotypic contributor (row). (a) Kauffman's local interaction map [35] in which both the epistatic and pleiotropic vectors are equal to the K value of 5. (b) Kauffman's random interaction map [35] in which the epistatic vectors are equal to the K value of 5 and the pleiotropic vectors are a Poisson distribution with a mean of the K value of 5. (c) Altenberg's genome selected interaction map [34] which is highly pleiotropic at a few loci (Fisher-Pearson coefficient of 1.76) as indicated by the vertically connected dark squares. Epistatic vectors are relatively equal (Fisher-Pearson coefficient of 0.13). (d) Our novel phenome selected interaction map, presented in the current study, in which both pleiotropic and epistatic vectors are relatively equal (Fisher-Pearson coefficient of 0.07 and 0.69, resp.).

From Figure 3 the comparison of the accessibility of local optima demonstrates the difference in basin size between different types of binding landscapes. For low K value landscapes such as $K = 2$ (Figure 3(a)), the model exhibits a range of basin sizes, with an optimum's basin size loosely correlating to its fitness. At higher K value (Figure 3(b)) the correlation between an optimum's basin size and its fitness is

abolished. The basin size was reduced sufficiently that during the adaptive walk simulation only a single adaptive walker can reach each local optimum.

The landscape derived from selective genome growth (Figure 3(c)) exhibits a wider range of basin sizes than a random $K = 2$ landscape (Figure 3(b)) with a maximum walker tally of 9 and a mean walker tally of 1.006 with a

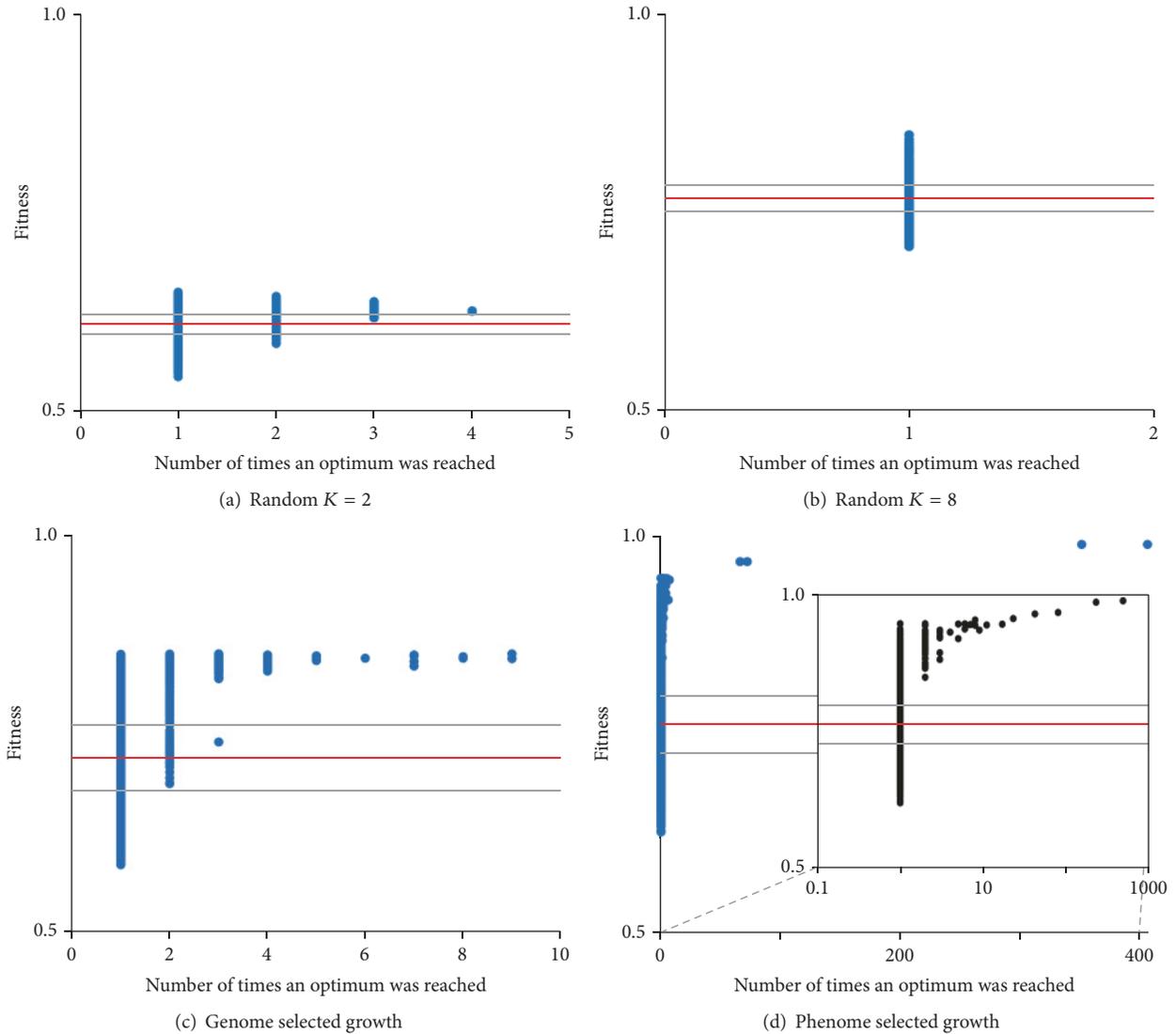


FIGURE 3: Comparison of basins: the accessibility of local optima discovered by 100,000 independent adaptive walks on four types of $N = 96$ landscape, shown in blue. (a) The $K = 2$ random landscape, 99249 unique optima. (b) The $K = 8$ random landscape, 100,000 unique optima. (c) The Genome selected landscape, 99381 unique optima. (d) Phenome selected landscape, 98977 unique optima. The log scale panel shows the same phenome selected landscape data in black. Red horizontal lines show mean and grey horizontal lines show one standard deviation from the mean.

standard deviation of 0.113. In this landscape an optimum's basin size correlates to its fitness.

The landscape derived from our selective phenome growth (Figure 3(d)) exhibits a much wider range of basin sizes than the genome selected landscape (Figure 3(c)) with a maximum walker tally of 407 and a mean walker tally of 1.010 with a standard deviation of 1.735. These statistical results indicate that the phenome selected landscape has a larger range of basin sizes when compared to genome selected landscapes. Furthermore, the distribution of these basin sizes has more variation in phenome selected landscapes when compared to genome selected landscapes.

In the phenome selected landscape there is a stronger correlation between an optimum's basin size and its fitness when compared to genome selected landscapes. Additionally

there are a lower number of unique optima for phenome selected (98977) compared to genome selected (99381) landscapes. This lower number of unique optima found for phenome selected landscape indicates a lower number of basins corresponding to a smoother, less chaotic landscape.

One major criticism of NK landscapes when used as fitness landscapes is that with larger K values they become excessively chaotic and an unrealistic representation of reality. This observed smoothness of the genome and phenome selected landscapes despite the landscape complexity suggests that the model is more similar to that of a real binding landscape.

From Figure 4 the correlation between an optimum's fitness and its similarity to the fittest optima found can be seen for different landscapes. Typically hamming distance from

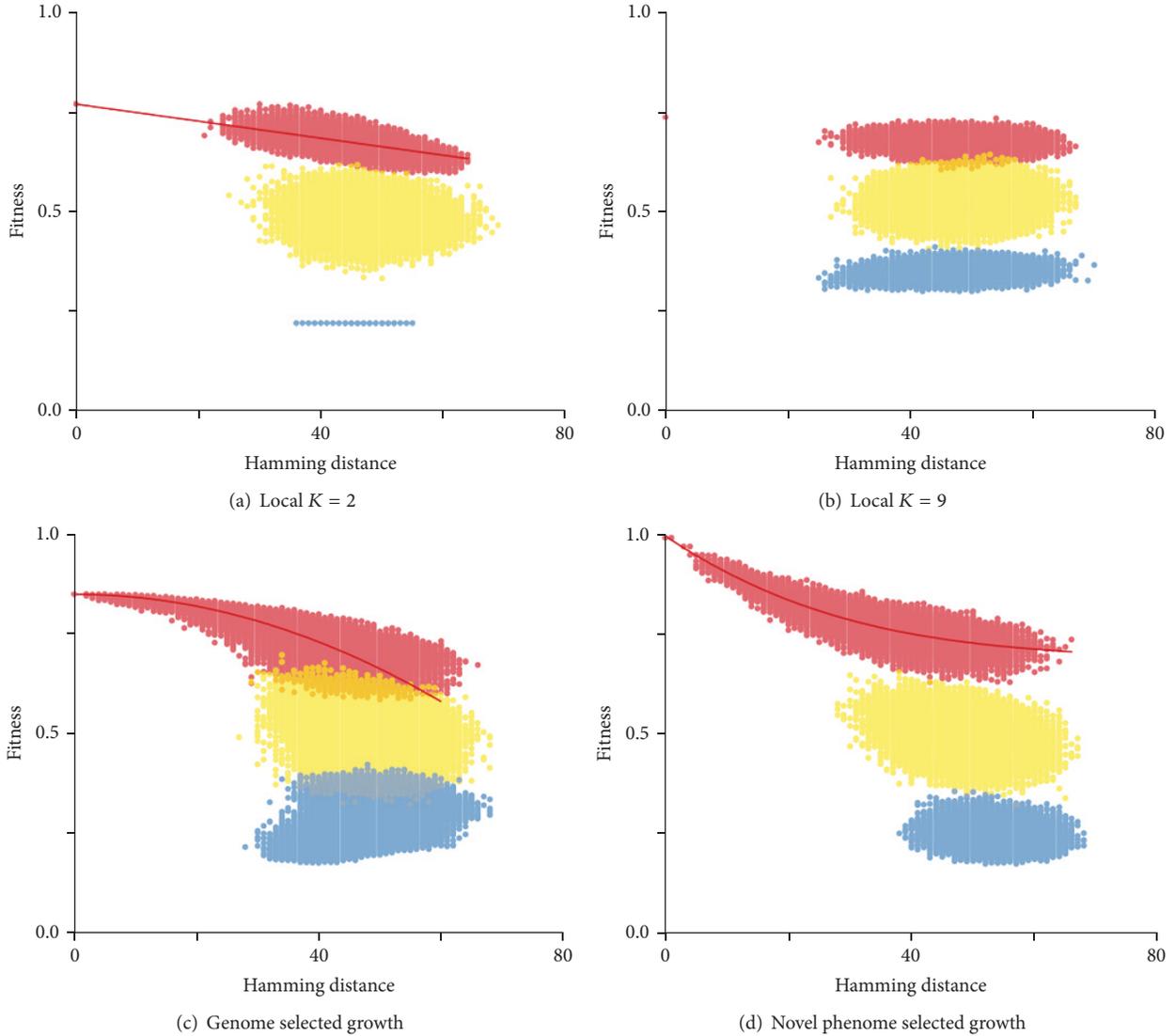


FIGURE 4: Hamming distance from fittest optima comparison: red points are the hamming distance from fittest optima found to other optima found over 100,000 adaptive walks. Yellow points are the hamming distance from fittest optima found to 100,000 randomly generated sequences. Blue points are the hamming distance from fittest optima found to the minima found over 100,000 downhill adaptive walks. (a) $K = 2$ local landscape; (b) $K = 8$ local landscape; (c) Genome selected landscape; (d) Phenome selected landscape.

fittest optima plots only includes other points from optima reached during adaptive walks, shown as red in Figure 4 [35, 40], but here we have included the two additional groups in order to better represent the landscapes. The first is a random sequence group, shown as yellow in Figure 4, which gives some indication of the average sequence within the landscape. The second is a minima group, shown as blue in Figure 4, which gives some indication of the troughs or valleys in the landscape.

For the $K = 2$ landscape (Figure 4(a)) the optima group, displayed as red points, show an inverse correlation between fitness of an optimum and its distance from the fittest optima. This correlation is indicative of the “Massif Central” global structure described by Kauffman [35]. The $K = 2$ random group, displayed as yellow points, show a range of fitness

values and no discernable relationship between fitness of a random sequence and its distance from the fittest optima. The $K = 2$ minima group, displayed as blue points, are all of a similar fitness which indicates a convergence on a small number of related low fitness sequences. This may be an artefact of the low complexity of the $K = 2$ landscape.

For the higher complexity $K = 8$ landscape (Figure 4(b)) the optima group, displayed as red points, show no correlation between fitness of an optimum and its distance from the fittest optima. The absence of this correlation shows a disruption of the “Massif Central” type landscape observed with higher K values.

In such a landscape, optima’s position in sequence space occurs more randomly in a chaotic fashion. The $K = 8$ random group displayed as yellow and the minima group

displayed as blue show sequences with the same similarity of the fittest optima as the optima group. This is indicative of a chaotic landscape.

For the genome selected landscape (Figure 4(c)) the optima group, displayed as red points, show an inverse relation between fitness of an optimum and its distance from the fittest optima. The line of best fit for this relation shows a decreasing exponential-like trend with a decreasing rate of fitness increase with closer hamming distance to the fittest optima found. This seems to represent a landscape in which many similar fitness optima exist at a relatively distant hamming proximity from the fittest optima found, indicative of a plateaued “Massif Central” fitness landscape.

The genome selected random group displayed as yellow points show quite an even distribution with a slight skew towards the inverse relation between fitness of a random point and its distance from the fittest optima. The genome selected minima group displayed as blue points show a relation between fitness of an optimum and its distance from the fittest optima. This relation is the exact opposite to that the optima group displayed. Intriguingly, this indicates that lowest minima are found closer to the fittest optima than the average minima. This pattern was observed on multiple genome selected interaction maps, so it is not an artefact.

One explanation for this effect is that for genome selected interaction maps the high pleiotropy attributed to a few loci means that there exist sequences with great sequence homology to the fittest optima but with relatively low fitness as these high pleiotropy and high fitness contributing loci are not optimised. To test this the optimum sequence was mutated at each of the high pleiotropy loci according to the interaction map and it was found that just 9 mutations yielded a sequence with a fitness score of 0.35. This sequence was not a minimum as it was not represented by one of the blue minima points in Figure 4(c). This example shows that it is useful to investigate minima and not just optima when characterising a landscape.

For the phenome selected landscape in Figure 4(d) the optima group, displayed as red points, show an inverse relation between fitness of an optimum and its distance from the fittest optima. The line of best fit for this correlation shows an increasing exponential-like trend with a higher rate of fitness increase with closer hamming distance to the fittest optima found. This seems to represent a landscape in which fitness level drops steeply from the fittest optima and gradually levels out. This phenome selected steeple type “Massif Central” landscape shows the opposite effect of the plateaued “Massif Central” described for the genome selected landscape. As mutational additivity is observed in empirical landscapes [25–33], the most realistic theoretical aptamer binding landscapes are Mount Fuji-like in nature [42] with fewer distinct fittest sequences. The steeple nature of the phenome selected landscape fits this Mount Fuji-like approximation well. Furthermore mutational additivity, or gradient of the red lines of best fit in Figure 4, holds for phenome selected landscape sequences close to the fittest sequence (Figure 4(d)); however, for genome selected

landscapes mutational additivity for sequences close to the fittest sequence does not hold (Figure 4(c)). This difference indicates that phenome selected landscapes are a more accurate representation of aptamer binding fitness landscapes.

The phenome selected random group displayed as yellow points show a slight skew towards the inverse relation between fitness of an optimum and its distance from the fittest optima, similar to that of the optima group. The phenome selected minima group displayed as blue points show an even distribution with no discernable correlations.

3.3. Optimising Phenome Selected Landscapes. As described earlier, phenome selected landscapes have the unique capability to alter the number of phenotypic contributors thereby dismissing the one-locus one-phenotype contributor paradigm. By changing the number of phenotypic contributors in an interaction map another mechanism, aside from altering K value, can be used to tune the landscape. Similarly to using K value to tune a landscape, the number of phenotypic contributors can be used to tune the landscape from states of relative order to states of relative chaos. In Figure 5(a) four interaction maps with varying numbers of phenotypic contributors were used to create fitness landscapes. These landscapes were assessed in two ways, firstly by their accessibility of optima (Figure 5(b)) and secondly by their optima’s hamming distance from the fittest optima found (Figure 5(c)).

For a low number of phenotypic contributors (Figure 5(a)(i)) local optima are inaccessible (Figure 5(b)(i)). At such low numbers of phenotypic contributors there exist loci in the interaction map which are not represented in any phenotypic contributors. This means that they have no bearing on sequence fitness. Mutational relatives in which these loci are altered are, from a phenotypic scoring point of view, equal but, from a sequence point of view, different. Therefore convergence on a single sequence representing a scoring optima is more difficult. Additionally, for a low number of phenotypic contributors, the optima’s hamming distance from the fittest optima (Figure 5(c)(i)) shows an inverse relation between fitness of an optimum and its distance from the fittest optima. However, despite this relation there is a relatively large fitness and hamming distance gap between the fittest optima and the next fittest optima. This gap reflects a large evolutionary jump from the optima to the fittest optima found.

For a low to intermediate number of phenotypic contributors (Figure 5(a)(ii)) the local optima (Figure 5(b)(ii)) are accessible with a reasonably large basin size for the fittest individuals and a range of basin sizes which correlate to fitness. This observed reasonably large basin size and basin size correlation to fitness are indicative of a well-tuned landscape. Additionally, the hamming distance from fittest optima (Figure 5(c)(ii)) shows a strong inverse relation between fitness of an optimum and its distance from the fittest optima. The distance between the fittest optima and the other optima is minimal indicating good mutational additivity and a well-tuned landscape.

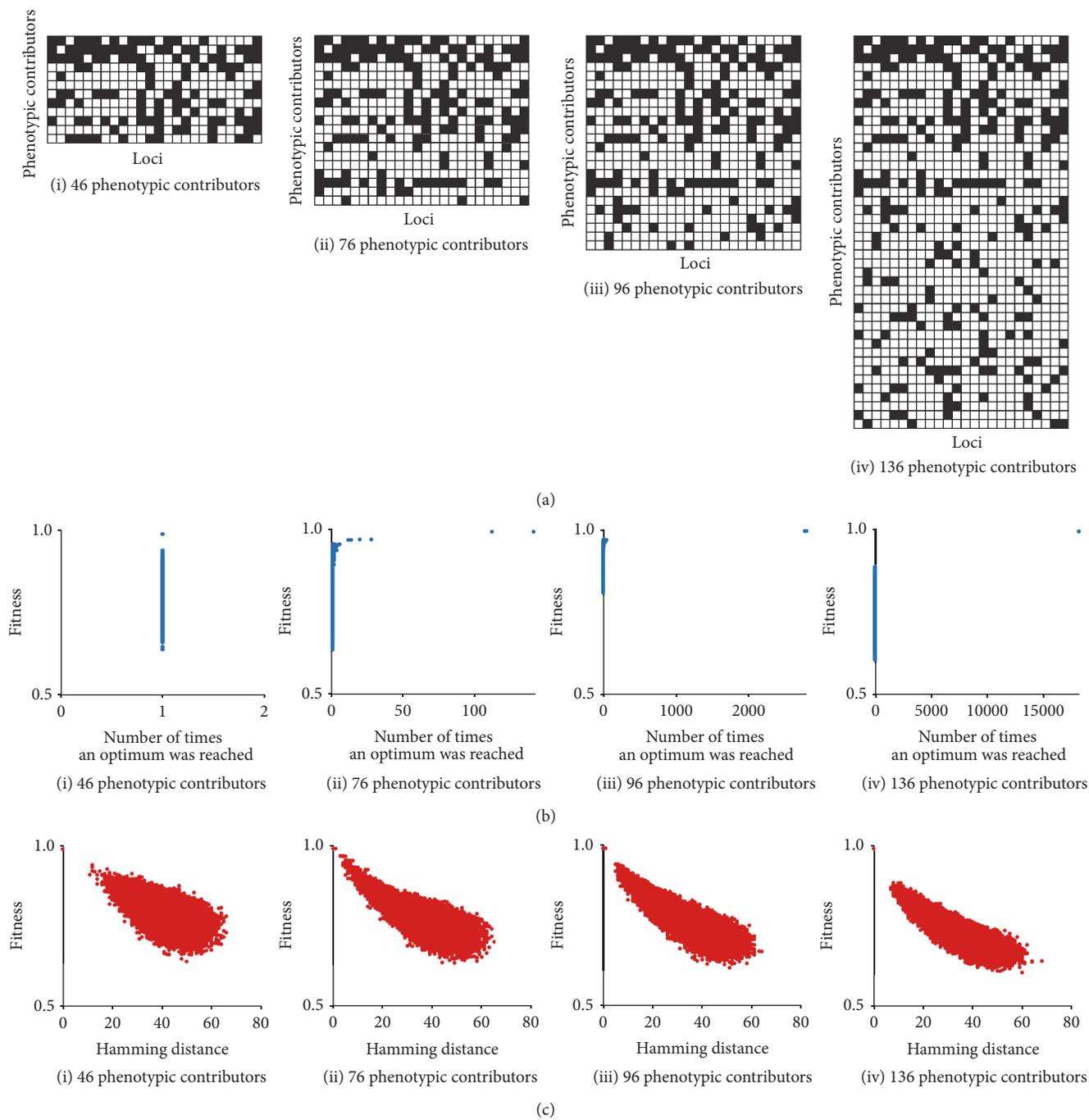


FIGURE 5: Comparison of numbers of phenotypic contributors for interaction map: (a) comparison of examples of interaction map relative dimensions. (b) Comparison of basin sizes by accessibility of local optima. For each of the four interaction maps 100,000 adaptive walks were used to discover the accessibility of the fitness landscape’s local optima. (c) Comparison of hamming distances from fittest optima. For each of the four interaction maps 100,000 adaptive walks were used to discover optima and the hamming distance of the fittest optima to all other optima versus optima fitness plotted. The four interaction maps compared are (i) 46 phenotypic contributors, (ii) 76 phenotypic contributors, (iii) 96 phenotypic contributors, and (iv) 136 phenotypic contributors.

For the intermediate to high range of phenotypic contributors (Figure 5(b)(iii)) the local optima (Figure 5(b)(iii)) are accessible albeit with a skewed distribution of basin sizes. The hamming distance from fittest optima (Figure 5(c)(iii)) shows an inverse relation between fitness of an optimum and its

distance from the fittest optima. There is a relatively large fitness and hamming distance gap between the fittest optima and the next fittest optima. This gap indicates overoptimisation towards the lead aptamer sequence, with the landscape becoming more rugged and optima being of a lower score.

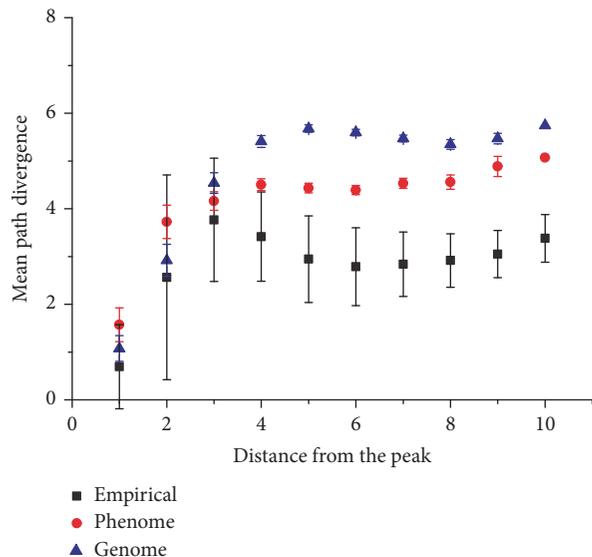


FIGURE 6: Comparison of mean path divergences among theoretical and empirical landscapes: mean path divergences, a measure of landscape smoothness, were compared between theoretical and empirical landscapes. The empirical landscape was binding data (black squares) for 10-base aptamers against the protein allophycocyanin [41]. The theoretical landscapes were genome-selected [34] (blue triangles) and the novel ones were phenome-selected (red circles), both with a string size of $N = 10$ in order to correspond with the empirical dataset. The phenome selected landscape is more similar to the empirical dataset.

For the high number of phenotypic contributors (Figure 5(a)(iv)) only one local optimum (Figure 5(b)(iv)) is accessible and its basin size is extremely large. This is indicative of landscape that has been overoptimised for the lead aptamer sequence to the point that it has converged on a singularity. The hamming distance from fittest optima (Figure 5(c)(iv)) shows an inverse relation between fitness of an optimum and its distance from the fittest optima. However there is a larger fitness and hamming distance gap between the fittest optima and the next fittest optima. This gap indicates severe overoptimisation towards the lead aptamer sequence, with the landscape becoming increasingly rugged and the optima being of a lower score.

This progression through the number of phenotypic contributors depicted in Figure 5 shows that there is an optimal number of phenotypic contributors when designing an interaction map for a fitness landscape.

3.4. Comparison to an Empirical Binding Landscape. To compare theoretical binding models to empirical aptamer-protein binding, DNA microarray data was used. Previously, Rowe et al. analysed a complete DNA-protein affinity landscape using DNA microarrays [41]. In this study all possible DNA oligomer variants of 10 bases were synthesised onto a microarray. The microarray was incubated with fluorescent allophycocyanin protein and fluorescent scanning performed to reveal the entire protein-binding landscape [41]. We used this complete empirical binding dataset for comparison to our theoretical landscapes.

To compare theoretical and empirical aptamer binding data we used mean path divergence of monotonic trajectories, a method of measuring smoothness of a landscape [11]. Briefly described, a monotonic function is entirely nonincreasing or nondecreasing; ergo a monotonic trajectory starts at point A and finishes at point B without changing the polarity of its slope. Path divergence is calculated by averaging the deviation within a set of monotonic trajectories between two set points, one point being a local maximum. Mean path divergence (MPD) is the average of path divergence values between all possible points on a landscape [11]. MPD is defined by Lobkovsky et al. as

$$D = \frac{\sum_i P_i d(p_i, p_0)}{\sum_i P_i}, \quad (3)$$

where P_i is the probability the occurrence of p_i and $d(p_1, p_2)$ is the distance between the trajectories p_1 and p_2 [11]. The current selected local maximum is p_0 . In this way MPD is a measure of the smoothness of landscape, one important feature for which there may be disparity between a theoretical model and empirical data. Each individual path divergence measurement is binned according to its hamming distance between the two given points before averaging so a mean path divergence profile is produced. The mean path divergence for selected and empirical landscapes is similar (Figure 6). Closer to the local maximum lower, MPD is observed. MPD increases with distance from the local maximum, levelling out at a local maximum distance of around 3.

The phenome selected landscape is more similar in path divergence to the empirical data than the previously described genome selected landscape (Figure 6). This similarity to empirical data in terms of mean path divergence shows that the phenome selected landscape is more similar in smoothness to the empirical landscape and therefore a more realistic representation of aptamer-protein binding.

4. Conclusions

Herein we have described a method for generating genotype-phenotype interaction maps with lower aggregated pleiotropy vectors which yield smooth fitness landscapes. These fitness landscapes have a steeple type “Massif Central” structure which appears to be more biologically accurate when representing the binding of an aptamer to an analyte.

Furthermore we have removed the one phenotype per gene paradigm that is the norm for NK model literature and used the varying number of phenotypic contributors to tune our new fitness landscapes. Comparisons between phenome selected landscapes and the similarly constructed genome selected landscapes have been made. The most striking difference is the complementary correlation line of best fit for hamming distance to fittest sequence (Figure 4). Perhaps future work would combine both approaches, constructionally selecting both genotypic and phenotypic contributors, to yield a landscape with a linear correlation line of best fit for hamming distance to fittest sequence. This difference seen in the correlation line of best fit for hamming distance to fittest sequence between genome and phenome relates to both

mutational additivity and the global landscape structure. In situations where a higher number of divergent fitness optima should occur, such as genome evolution, genome selected landscapes seem more representative. In situations where a lower number of convergent fitness optima should occur, such as the binding of an aptamer to a specific analyte, phenome selected landscapes seem more representative. Our phenome selected landscape seems to model aptamer-target binding better than any other tested model. A future application of the phenome selected fitness landscape model would be its use in simulation of aptamer selection. As the model seems to be more accurate, the simulation would be a more accurate representation of real life aptamer selection so any selection condition optimisations should be more applicable and transferable to real life aptamer selections. Although phenome selected landscapes seem to represent this biological system more accurately than their standard *NK* landscape counterparts, caution should be used when applying them to other biological systems as they may not be well described.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

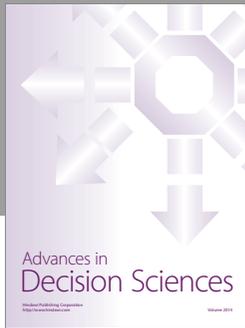
Acknowledgments

This research was funded by Hong Kong RGC GRF Grants 17127515 and 17119814 awarded to Julian Alexander Tanner, by HKU Outstanding Young Researcher Award awarded to Julian Alexander Tanner, and by Seed Fund Grant 2014090160039 awarded to Julian Alexander Tanner. The authors thank Dr. William Rowe for providing the allophycocyanin-DNA microarray binding data.

References

- [1] A. D. Ellington and J. W. Szostak, "In vitro selection of RNA molecules that bind specific ligands," *Nature*, vol. 346, no. 6287, pp. 818–822, 1990.
- [2] C. Tuerk and L. Gold, "Systemic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase," *Science*, vol. 249, no. 4968, pp. 505–510, 1990.
- [3] A. D. Ellington and J. W. Szostak, "Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures," *Nature*, vol. 355, no. 6363, pp. 850–852, 1992.
- [4] R. M. Dirkzwager, A. B. Kinghorn, J. S. Richards, and J. A. Tanner, *Chemical Communications*, vol. 51, pp. 4697–4700, 2015.
- [5] A. Ozer, J. M. Pagano, and J. T. Lis, *Molecular Therapy Nucleic Acids*, vol. 3, p. e183, 2014.
- [6] W. H. Michael Jr. and W. T. Blackshear, "Recent results on the mass, gravitational field and moments of inertia of the moon," *The Moon*, vol. 3, no. 4, pp. 388–402, 1972.
- [7] M. Platt, W. Rowe, J. Knowles, P. J. Day, and D. B. Kell, *Integrative Biology*, vol. 1, pp. 116–122, 2009.
- [8] C. G. Knight, M. Platt, W. Rowe et al., "Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape," *Nucleic Acids Research*, vol. 37, no. 1, article e6, 2009.
- [9] W. Rowe, M. Platt, D. C. Wedge, P. J. Day, D. B. Kell, and J. Knowles, *PLoS*, vol. 7, 036007, 2010.
- [10] C. D. Carlson, C. L. Warren, K. E. Hauschild et al., "Specificity landscapes of DNA binding molecules elucidate biological function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 10, pp. 4544–4549, 2010.
- [11] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, "Quantifying the similarity of monotonic trajectories in rough and smooth fitness landscapes," *Molecular BioSystems*, vol. 9, no. 7, pp. 1627–1631, 2013.
- [12] A. B. Kinghorn, R. M. Dirkzwager, S. Liang et al., "Aptamer Affinity Maturation by Resampling and Microarray Selection," *Analytical Chemistry*, vol. 88, no. 14, pp. 6981–6985, 2016.
- [13] A. E. Minoche, J. C. Dohm, and H. Himmelbauer, "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems," *Genome Biology*, vol. 12, no. 11, article no. R112, 2011.
- [14] M. Cho, Y. Xiao, J. Nie et al., "Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 35, pp. 15373–15378, 2010.
- [15] P. Jiang, S. Meyer, Z. Hou et al., "MPBind: A Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers," *Bioinformatics*, vol. 30, no. 18, pp. 2665–2667, 2014.
- [16] P. Schuster and P. F. Stadler, *Discrete Models of Biopolymers*, Springer, 2004.
- [17] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshfte für Chemie Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1989.
- [18] J. Hoinka, A. Berezhnoy, P. Dao, Z. E. Sauna, E. Gilboa, and T. M. Przytycka, "Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery," *Nucleic Acids Research*, vol. 43, no. 12, pp. 5699–5707, 2015.
- [19] I. S. Oh, Y.-G. Lee, and R. McKay, "Simulating chemical evolution," in *Proceedings of the 2011 IEEE Congress of Evolutionary Computation, CEC 2011*, pp. 2717–2724, June 2011.
- [20] S. Kauffman and S. Levin, "Towards a general theory of adaptive walks on rugged landscapes," *Journal of Theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.
- [21] M. W. Deem and H. Y. Lee, *Physical Review Letters*, vol. 91, Article ID 068101, 2003.
- [22] M. Hall, K. Christensen, S. A. Di Collobiano, and H. J. Jensen, "Time-dependent extinction rate and species abundance in a tangled-nature model of biological evolution," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 66, no. 1, Article ID 011904, p. 011904/10, 2002.
- [23] S. Kauffman and W. Macready, "Complexity," vol. 1, pp. 26–43, 1995.
- [24] S. Klussmann, *The Aptamer Handbook: Functional Oligonucleotides and Their Applications*, John Wiley & Sons, 2006.
- [25] Y. Takeda, A. Sarai, and V. M. Rivera, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, pp. 439–443, 1989.
- [26] A. Sarai and Y. Takeda, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, pp. 6513–6517, 1989.
- [27] L. Serrano, A. G. Day, and A. R. Fersht, "Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability," *Journal of Molecular Biology*, vol. 233, no. 2, pp. 305–312, 1993.

- [28] W. S. Sandberg and T. C. Terwilliger, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, pp. 8367–8371, 1993.
- [29] X.-J. Zhang, W. A. Baase, B. K. Shoichet, K. P. Wilson, and B. W. Matthews, “Enhancement of protein stability by the combination of point mutations in t4 lysozyme is additive,” *Protein Engineering, Design and Selection*, vol. 8, no. 10, pp. 1017–1022, 1995.
- [30] M. M. Skinner and T. C. Terwilliger, “Potential use of additivity of mutational effects in simplifying protein engineering,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 20, pp. 10753–10757, 1996.
- [31] P. V. Nikolova, J. Henckel, D. P. Lane, and A. R. Fersht, “Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14675–14680, 1998.
- [32] T. Aita, N. Hamamatsu, Y. Nomiya, H. Uchiyama, Y. Shibanaka, and Y. Husimi, “Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution,” *Biopolymers*, vol. 64, no. 2, pp. 95–105, 2002.
- [33] J. A. Wells, “Additivity of mutational effects in proteins,” *Biochemistry*, vol. 29, no. 37, pp. 8509–8517, 1990.
- [34] L. Altenberg, “Evolving better representations through selective genome growth,” in *Proceedings of the 1st IEEE Conference on Evolutionary Computation*, pp. 182–187, 1994, IEEE World Congress on Computational Intelligence.
- [35] S. A. Kauffman, *Biomathematics and Related Computational Problems*, Springer, 1988.
- [36] D. C. Wedge, W. Rowe, D. B. Kell, and J. Knowles, “In silico modelling of directed evolution: Implications for experimental design and stepwise evolution,” *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 131–141, 2009.
- [37] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.
- [38] S. A. Kauffman, *The Origins of Order: Self Organization and Selection in Evolution*, Oxford University Press, 1993.
- [39] L. Altenberg, *Evolution and Biocomputation*, Springer, 1995.
- [40] T. Hebborn, S. Bullock, and D. Cliff, “NK α : Non-uniform epistatic interactions in an extended NK model,” in *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pp. 234–241.
- [41] W. Rowe, M. Platt, D. C. Wedge, P. J. Day, D. B. Kell, and J. Knowles, “Analysis of a complete DNA - protein affinity landscape,” *Journal of the Royal Society Interface*, vol. 7, no. 44, pp. 397–408, 2010.
- [42] T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo, and Y. Husimi, “Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and thermolysin,” *Biopolymers*, vol. 54, no. 1, pp. 64–79, 2000.




Hindawi

Submit your manuscripts at
<https://www.hindawi.com>

