Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2013, **9**, 1627

Quantifying the similarity of monotonic trajectories in rough and smooth fitness landscapes

Alexander E. Lobkovsky, Yuri I. Wolf and Eugene V. Koonin*

Received 4th December 2012,
Accepted 21st February 2013

DOI: 10.1039/c3mb25553k

www.rsc.org/molecularbiosystems

When selection is strong and mutations are rare, evolution can be thought of as an uphill trajectory in a rugged fitness landscape. In this context the fitness landscape is a directed acyclic graph in which nodes are genotypes and edges lead from lower to higher fitness genotypes that differ by a single mutation. Because the space of genotypes is vastly multi-dimensional, classification of fitness landscapes is challenging. Many proposed summary characteristics of fitness landscapes attempt to quantify biologically relevant and intuitive notions such as roughness or peak accessibility in alternative ways. Here we explore, in different types of landscapes, the behavior of the recently introduced mean path divergence which quantifies the degree of similarity among evolutionary trajectories with the same endpoints. We find that monotonic trajectories in empirical and model fitness landscapes are significantly more constrained, with low median path divergence, than those in purely additive landscapes. By contrast, transcription factor sequence specificity (aptamer binding affinity) landscapes are markedly smoother and allow substantial variability in monotonic paths that can be greater than that in fully additive landscapes. We propose that the smoothness of the specificity landscapes is a consequence of the simple dependence of the transcription factor binding affinity on the aptamer sequence in contrast to the complex sequence-fitness mapping in folding landscapes.

1 Introduction

The construct of the fitness landscape is useful in cases when the fitness of an organism depends only on its genotype and not on the environment or population structure.[1–4] The landscape idea allows the separation of the population dynamic effects such as clonal interference from the effect of genetic background on the fate of mutations.[5–7] Even when the population dynamics are complex, and the population contains a distribution of clones with different genotypes, evolution can still be envisioned as an ensemble of splitting and merging trajectories on the fitness landscape.[8]

Because fitness landscapes are vastly multi-dimensional, their characterization is difficult. A number of summary statistics have been proposed to reflect the properties of landscapes relevant to evolution.[3,9–14] Here we apply the recently introduced mean path divergence metric to a number of landscapes of radically different origins.[8,15] Mean path divergence, defined in detail in the following section, quantifies the degree of similarity of trajectories that share the starting and ending points. Computing mean path divergence requires that each trajectory have an associated probability of occurrence within some population dynamics on the landscape. The magnitude of path divergence reflects the expected distance between a randomly picked trajectory and the most probable trajectory, and can therefore be used to assess the degree of evolutionary repeatability, or in other words, to quantify the difference in outcomes of two evolutionary experiments performed under exactly the same conditions.[8,15]

Here we employ mean path divergence to investigate the behavior of evolutionary trajectories in several types of landscapes where fitness is defined using substantially different criteria. The three classes of landscape exploit models of RNA[16] or protein folding[17] where fitness is defined as folding robustness or energy; mutational data where fitness corresponds directly to an organism's growth rate[18] or to enzyme specificity for a particular substrate;[19] and exhaustive data on aptamer binding for a variety of transcription factors.[20–23] We show that while the folding landscapes exhibit low path divergence, or in other words, impose strong constraints on the trajectories available for evolution, binding specificity landscapes are smooth, with substantially higher mean path divergence. We hypothesize that the differences in path divergence reflect major differences in the complexity of the sequence-fitness mapping.[24–29]

National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA. E-mail: koonin@ncbi.nlm.nih.gov

2 Mean path divergence

When mutations are rare, they appear sequentially. Therefore, at any time the population contains only the wild type and the unique mutant genotype. When selection pressure is strong and population size is large, beneficial mutations are fixed with certainty whereas the probability of fixation of neutral and deleterious mutations is vanishingly small. Evolution is therefore a strictly monotonic trajectory on the fitness landscape.[30] Although the ''weak mutation strong selection'' approximation might rarely if at all apply to real world situations, it allows the statistical properties of the landscape to be disentangled from the complexities introduced by clonal interference and drift.

When attempted mutations are random, the fixation probability of a particular beneficial mutation is simply the inverse of the total number of available beneficial mutations. In other words, the probability of taking a particular uphill step on the landscape is the inverse of the number of the available distinct uphill steps. Therefore, the total probability P of a trajectory is the inverse of the product of the number of available beneficial mutations at each point along the trajectory.

Let us now consider an ensemble $\{p_i\}$ of monotonic trajectories, illustrated in Fig. 1b, that start at the same point on the landscape and terminate at the same peak. Let p_0 be the trajectory of maximum probability of occurrence in this ensemble (shown in bold in Fig. 1b). If several trajectories have the same maximum probability, one of them is selected at random. The mean path divergence D of the ensemble is defined as

$$D = \frac{\sum_i P_i d(p_i, p_0)}{\sum_i P_i}, \quad (1)$$

where P_i is the probability of occurrence of trajectory p_i defined above, and $d(p_1,p_2)$ is the distance between trajectories p_1 and p_2 defined in Fig. 1a. The normalizing factor in the denominator is needed because it is not feasible in many cases to find all monotonic trajectories. In practice only trajectories with the probability of occurrence above a certain threshold are included in the ensemble.

The properties of a landscape are computed as follows. Given a point Q in the landscape, the highest accessible peak S is located. Suppose that the Hamming distance between Q and S is H. All strictly uphill trajectories from Q to S with the probability of occurrence $P > 10^{-10}$ are found and the mean path divergence D among these trajectories is computed using eqn (1). The path divergence values for all points on the landscape are binned according to the distance H from their associated peaks. The distribution of path divergences in each H bin is then summarized and plotted using the standard box and whiskers representation.

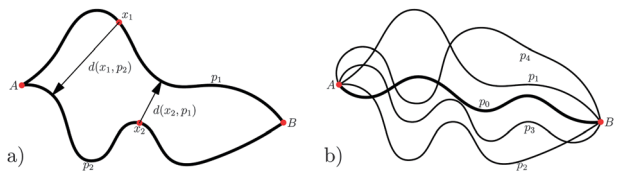

Fig. 1 (a) Computation of the inter-trajectory distance $d(p_1,p_2)$ involves finding the closest point on p_2 from every location x_1 on p_1 and *vice versa*, summing these minimal distances $d(x_1,p_2)$ and $d(x_2,p_1)$ and dividing the sum by the combined length of p_1 and p_2. (b) If p_0 is the most likely trajectory in the ensemble (shown in bold), the path divergence is the sum of the inter-trajectory distances $d(p_i,p_0)$ between p_0 and each of the trajectories p_i in the ensemble weighted by the respective probabilities of occurrence P_i.

3 Results and discussion

We analyzed empirical and model landscapes of different origins, summarized in Table 1, and described in more detail below.

3.1 Synthetic landscapes

We considered three types of synthetic landscapes in which the space of short sequences is explored completely and the fitness is derived either from random distributions or from models of folding polymers.

3.1.1 *NK* epistatic landscapes. In the *NK* model each site's contribution to the total fitness is influenced by K other sites.[31]

Table 1 Analyzed landscapes

Landscape	Method of construction	Size	H = 4 path div.
Additive	Random fitness effects drawn from a uniform distribution	65 536	M = 2 : 0.70 M = 4 : 0.91
Random epistatic	*NK* model with M = 2, N = 12	4096	K = 4 : 0.53 K = 10 : 0.38
Folding	Folding robustness of model 3D polymers	19 924	M = 4 : 0.41
RNA folding	Secondary structure similarity and thermal stability	52 656	0.31
A. niger	Mycelium growth rate	256	0.5
Sesquiterpene synthase	Preponderance of a single catalytic product	419	0.57
c-abl	DNA binding affinity	32 896	0.44
gata4	DNA binding affinity	32 896	0.52
jumonji	DNA binding affinity	32 896	0.61
p53	DNA binding affinity	32 896	0.72
nkx	DNA binding affinity	131 072	0.74
pa1	DNA binding affinity	131 072	0.86
pa2	DNA binding affinity	65 536	0.85
pa3	DNA binding affinity	32 896	0.95
tbp	DNA binding affinity	129 073	0.85

The purely additive landscape is the $K = 0$ special case of the *NK* family of landscapes. An uncorrelated landscape in which the fitness of every point is drawn from the same distribution independently is obtained when $K = N - 1$. We considered binary *NK* landscapes (*i.e.* the alphabet size of $M = 2$) with $N = 12$ and a range of K. Because the properties of landscapes could depend on a particular realization of the random fitness effects, we averaged our findings over 100 instances of the landscape for each value of K.

3.1.2 Additive landscapes. Additive landscapes are a special case of the *NK* landscapes in the absence of epistasis, *i.e.* when the every substitution in the most fit sequence has a fixed additive negative fitness effect. When only a single type of substitution is allowed in each site, *i.e.* when the alphabet size is $M = 2$, all trajectories that do not include back mutations are monotonic in fitness. In other words, if a mutant is different from the most fit sequence by substitutions at H sites, the mutations which restore the most fit sequence can occur in any order and increase fitness at every step. Because for $M = 2$ back mutations always decrease fitness, there are H uphill directions at each point on the landscape. Thus, there are always exactly $H!$ uphill trajectories and the mean path divergence is a function of only H. When the number of allowed substitutions is $M > 2$, the situation is different. The number of available uphill steps depends on the fitness effects, and therefore the number of uphill trajectories and the mean path divergences fluctuate among points at the same distance from the peak. We analyze the case of alphabet size $M = 4$ and sequence length $N = 8$ numerically in the same way as the empirical and other model landscapes.

3.1.3 RNA folding landscapes. Landscapes with fitness values derived from the prediction of RNA secondary structures have been used extensively to probe evolutionary questions.[16,32,33] We follow Cowperthwaite *et al.*[16] and use the Vienna package version 1.8.4 (ref. 34) to examine the secondary structures of all RNA sequences of $N = 12$ nucleotides. The most stable structure is chosen as the target. The fitness of any sequence is derived from the Hamming distance H between the parenthesis representations of its predicted structure and the target structure. Because the space of RNA secondary structures is highly degenerate, we resolve the landscape further by using the ΔG value and assigning higher fitness to sequences with lower ΔG if they share secondary structure. Roughly 20% of sequences with trivial predicted secondary structure (unfolded) were excluded from the landscape.

3.1.4 Folding robustness landscapes. We have previously introduced landscapes derived from the folding robustness of off-lattice 3D "protein-like" model heteropolymers.[15,17] The fitness of a sequence was defined as the probability of folding to a target structure. The folding polymers were modeled as flexible, length $N = 25$, chains of monomers of 4 types interacting *via* the Lennard-Jones and screened Coulomb potentials. Folding probabilities were computed by considering the configuration space overlap of the fluctuating conformation of the query sequence with that of the target structure. The landscapes included all sequences with non-zero folding probabilities that had a neighbor with a folding probability above a certain threshold. This analysis includes a single representative target structure landscape but the results of our previous work indicate that other structures produced statistically similar landscapes.

3.2 Empirical landscapes

We consider empirical landscapes that can be abstracted from experimental data in which the fitness value is measured in different ways.

3.2.1 Sesquiterpene synthase specificity. O'Maille *et al.* found that the primary product of the catalytic activity of a family of sesquiterpene synthase enzymes could be switched *via* a set of 8 amino-acid substitutions.[19] When a subset of the 8 substitutions were present, several reaction products were produced in varying proportions. We computed the enzyme specificity, *i.e.* a measure of the dominance of one product over the others as the fitness value.[15]

3.2.2 *Aspergillus niger* growth rate. Franke *et al.*[18] measured the mycelium growth rate of a library of *A. niger* mutants with combinations of up to 8 phenotypic marker mutation. The fitness value in this landscape is the measured growth rate.

3.2.3 Aptamer binding. High-throughput microarray experiments can interrogate the entire sequence space of 8 to 10 nucleotide aptamers that interact with a DNA-binding molecule.[23] The resulting sequence specificity landscapes (SSL) are large and dense. The fitness value is the binding affinity of a transcription factor to the DNA sequence motif. We obtained and analyzed complete SSL's[23] for a number of DNA binding proteins: c-abl, gata4, jumonji, nkx-2.5, p53, pa-1, pa-2, pa-3, and tbp (private communication, A. Z. Ansari).

4 Mean path divergence in different classes of landscapes

Fig. 2 shows the distributions of path divergences in most of the analyzed landscapes as a function of the distance from the peak. There is a clear difference between the true fitness landscapes and the binding affinity landscapes, the latter being markedly smoother with substantially larger path divergences. Note that there are points at a distance $H = 1$ from the peak with multiple uphill trajectories to the peak. These trajectories necessarily contain reverse mutations.

Fig. 3 shows the dependence of the path divergence on the distance H from the peak. We compute the median path divergence among all points that are at a distance H from the peak and have more than one trajectory to the peak. To account for the cases in which there is a unique trajectory to the peak we multiply the median path divergence among the points with more than one trajectory by the fraction of points with more than one trajectory.

Path divergence partitions the landscapes we considered here into roughly three classes. Path divergence in folding landscapes, including the model polymer and RNA landscapes, exhibits a roughly flat dependence on the distance from the peak. By this measure, the folding landscapes appear similar to

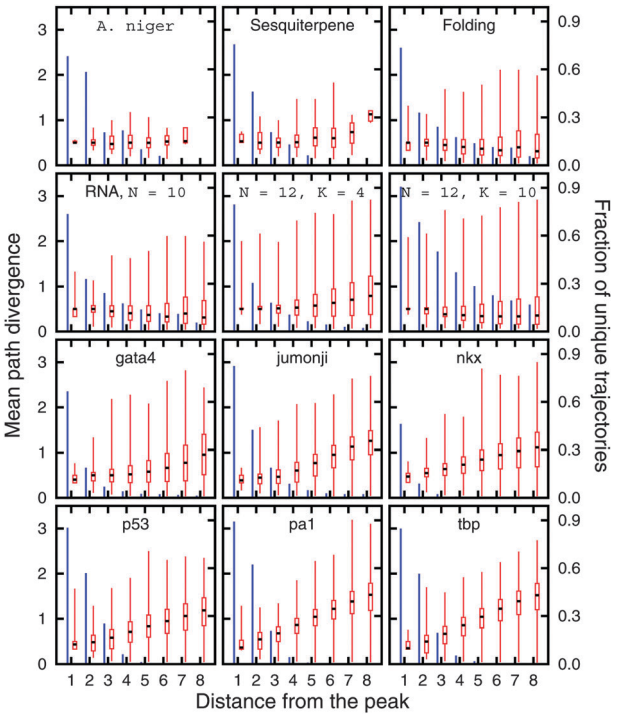

Fig. 2 Mean path divergences in the model and empirical landscapes indexed by their distance to the peak. Blue bars denote the fraction of points which have a single uphill trajectory to the peak. Red boxes and whiskers summarize the distribution of the path divergences of the remaining cases in which there are multiple uphill trajectories to the peak.

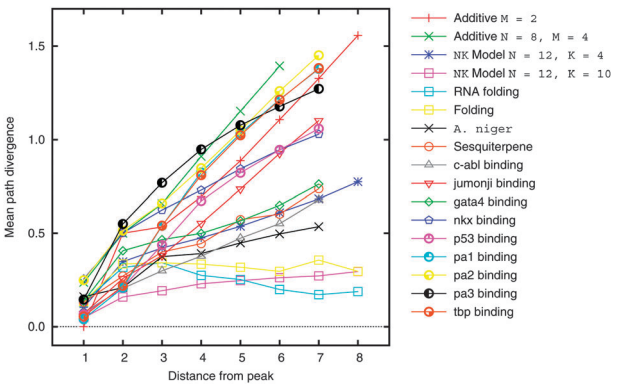

Fig. 3 The median path divergence (*cf.* the black line in the red boxes in Fig. 2) multiplied by the fraction of cases in which there are multiple uphill trajectories *vs.* the distance H from the peak.

almost uncorrelated NK landscapes with $N = 12$ and $K = 10$ (recall that $K = N - 1$ is the completely uncorrelated case).

The second class of landscapes includes both the true fitness landscapes (sesquiterpene synthase and *A. niger*) and the c-abl and gata4 binding affinity landscapes. These landscapes show a moderate increase in path divergence away from the peak and cluster with the moderately correlated NK landscapes with $N = 12$ and $K = 4$.

Finally, path divergence in the rest of the binding affinity landscapes is similar to that in the additive landscapes.

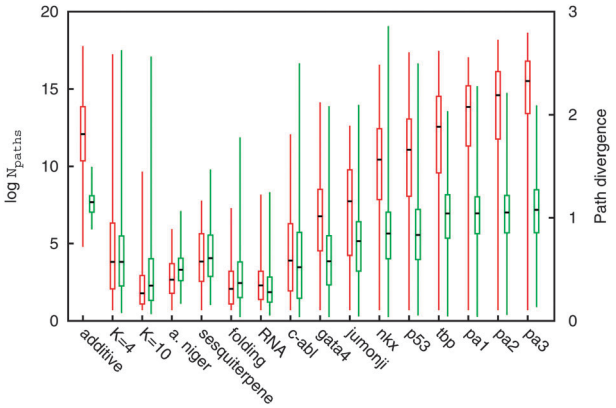

Fig. 4 Summaries of the distributions of the logarithm of the number of uphill paths (red, left axis) and the path divergence (green, right axis) for points $H = 5$ away from the peak.

An additional distinction can perhaps be made between the slightly more rough p53, jumonji and nkx binding affinity landscapes and the rest.

The distributions of the number of uphill trajectories shown in Fig. 4 for all points at a distance $H = 5$ from the peak correlates strongly with path divergence indicating that the increase in path divergence is caused primarily by the increase in the number of available uphill trajectories. Notable exceptions to this rule are the binding affinity landscapes for pa1, pa2 and pa3 in which there are more uphill trajectories than in the additive landscape whereas the path divergence is lower indicating that the uphill trajectories cluster around the most probable route.

5 Conclusions

Evolution can be frequently represented by the dynamics of a single or multiple stochastic uphill walkers on a vastly multi-dimensional fitness landscape. Statistical properties of landscapes have therefore become the focus of a large number of theoretical and recently experimental studies. With empirical fitness landscapes becoming available, it is important to develop metrics that are relevant to evolution and therefore can be used to classify the landscapes. It is also of interest to compare fitness landscapes to model landscapes designed to mimic them and to empirical landscapes of alternative origins such as the sequence specificity or binding affinity landscapes.

We found that the path divergence metric clearly differentiates the studied landscapes into the almost uncorrelated, moderately smooth and super smooth ones (with higher availability of trajectories than in a purely additive landscape). The group of nearly uncorrelated landscapes with low characteristic path divergence includes the model RNA-folding and protein-folding landscapes in which median path divergence is substantially lower than it is in additive landscapes. The low median path divergence reflects strong constraints on the accessible evolutionary trajectories in these rugged landscapes. In contrast, binding specificity landscapes are smooth, with high path divergence that reflects minimal constraints on evolutionary trajectories. The only

"true" fitness landscapes that we had an opportunity to study, that derived from the *A. niger* growth data, and the enzyme catalytic specificity, showed intermediate properties. We surmise that the simple, effectively additive dependence of the transcription factor-binding affinity on the distance of an aptamer sequence from the consensus binding sequence translates into multiple monotonic paths toward the "fitness peak". In contrast, the complex processes of protein and RNA folding, even in their model incarnations, allow few monotonic paths, hence low path divergence and in the extreme nearly deterministic evolution. Put another way, the binding affinity landscapes are simple with respect to the sequence-fitness (phenotype) mapping whereas the folding landscapes are complex. It is less clear why the empirical fitness landscapes yield intermediate values of path divergence because generally one would expect complex mapping of sequence to phenotype in these cases as well. We suspect that because the available data cover only a tiny fraction of these landscapes, it is insufficient to obtain robust characteristics of the landscape as a whole. Comprehensive analysis of information-rich landscapes using these and other formal measures could provide insights into the relationship between determinism and randomness in evolution.

Acknowledgements

The authors' research is supported by the intramural funds of the DHHS (National Institutes of Health, National Library of Medicine). The authors wish to thank Dr Devesh Bhimsaria and Prof. Aseem Ansari for providing the DNA aptamer binding data.

References

1 J. M. Smith, *Nature*, 1970, **225**, 563–564.
2 M. Kogenaru, M. G. de Vos and S. J. Tans, *Crit. Rev. Biochem. Mol. Biol.*, 2009, **44**, 169–174.
3 M. Carneiro and D. L. Hartl, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(suppl 1), 1747–1751.
4 P. Schuster, *Theory Biosci.*, 2011, **130**, 71–89.
5 M. M. Desai, D. S. Fisher and A. W. Murray, *Curr. Biol.*, 2007, **17**, 385–394.
6 K. C. Kao and G. Sherlock, *Nat. Genet.*, 2008, **40**, 1499–1504.
7 C. R. Miller, P. Joyce and H. A. Wichman, *Genetics*, 2011, **187**, 185–202.
8 A. E. Lobkovsky and E. V. Koonin, *Front. Genet.*, 2012, **3**, 246.
9 T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo and Y. Husimi, *Biopolymers*, 2000, **54**, 64–79.
10 T. Smith, P. Husbands, P. Layzell and M. O'Shea, *Evol. Comput.*, 2002, **10**, 1–34.
11 D. M. Weinreich, R. A. Watson and L. Chao, *Evolution*, 2005, **59**, 1165–1174.
12 T. Aita, *J. Theor. Biol.*, 2008, **254**, 252–263.
13 S. Kryazhimskiy, G. Tkacik and J. B. Plotkin, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 18638–18643.
14 W. Rowe, D. C. Wedge, M. Platt, D. B. Kell and J. Knowles, *Bioinformatics*, 2010, **26**, 2145–2152.
15 A. E. Lobkovsky, Y. I. Wolf and E. V. Koonin, *PLoS Comput. Biol.*, 2011, **7**, e1002302.
16 M. C. Cowperthwaite, E. P. Economo, W. R. Harcombe, E. L. Miller and L. A. Meyers, *PLoS Comput. Biol.*, 2008, **4**, e1000110.
17 A. E. Lobkovsky, Y. I. Wolf and E. V. Koonin, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 2983–2988.
18 J. Franke, A. Klozer, J. A. de Visser and J. Krug, *PLoS Comput. Biol.*, 2011, **7**, e1002134.
19 P. E. O'Maille, A. Malone, N. Dellas, B. Andes Hess Jr, L. Smentek, I. Sheehan, B. T. Greenhagen, J. Chappell, G. Manning and J. P. Noel, *Nat. Chem. Biol.*, 2008, **4**, 617–623.
20 C. L. Warren, N. C. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips Jr and A. Z. Ansari, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 867–872.
21 C. G. Knight, M. Platt, W. Rowe, D. C. Wedge, F. Khan, P. J. R. Day, A. McShea, J. Knowles and D. B. Kell, *Nucleic Acids Res.*, 2009, **37**, e6.
22 W. Rowe, M. Platt, D. C. Wedge, P. J. Day, D. B. Kell and J. Knowles, *J. R. Soc., Interface*, 2010, **7**, 397–408.
23 C. D. Carlson, C. L. Warren, K. E. Hauschild, M. S. Ozers, N. Qadir, D. Bhimsaria, Y. Lee, F. Cerrina and A. Z. Ansari, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 4544–4549.
24 C. B. Anfinsen, *Science*, 1973, **181**, 223–230.
25 K. A. Dill, *Biochemistry*, 1990, **29**, 7133–7155.
26 J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins*, 1995, **21**, 167–195.
27 B. Dubertret, S. Liu, Q. Ouyang and A. Libchaber, *Phys. Rev. Lett.*, 2001, **86**, 6022–6025.
28 C. G. Kalodimos, N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens and R. Kaptein, *Science*, 2004, **305**, 386–389.
29 D. U. Ferreiro, I. E. Sanchez and G. de Prat Gay, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 10797–10802.
30 D. M. Weinreich, N. F. Delaney, M. A. Depristo and D. L. Hartl, *Science*, 2006, **312**, 111–114.
31 S. Kauffman and S. Levin, *J. Theor. Biol.*, 1987, **128**, 11–45.
32 L. A. Meyers, J. F. Lee, M. Cowperthwaite and A. D. Ellington, *J. Mol. Evol.*, 2004, **58**, 681–691.
33 M. C. Cowperthwaite, J. J. Bull and L. A. Meyers, *Genetics*, 2005, **170**, 1449–1457.
34 P. Schuster, W. Fontana, P. F. Stadler and I. L. Hofacker, *Proc. Biol. Sci.*, 1994, **255**, 279–284.