Convergent evolution to an aptamer observed in small populations on DNA microarrays

2010 Phys. Biol. 7 036007

(http://iopscience.iop.org/1478-3975/7/3/036007)

Phys. Biol. **7** (2010) 036007 (11pp) doi:10.1088/1478-3975/7/3/036007

Convergent evolution to an aptamer observed in small populations on DNA microarrays

W Rowe[1,2,6,7], M Platt[3,7], D C Wedge[1,2], P J R Day[1,4], D B Kell[1,2] and J D Knowles[1,5]

[1] Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3] School of Chemistry and Chemical Biology, University College Dublin, Belfield campus, Dublin 4, Ireland
[4] School of Translational Medicine, The University of Manchester, Oxford Road, Manchester M13 9PT, UK
[5] School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL, UK

E-mail: william.rowe@manchester.ac.uk

Received 20 February 2010
Accepted for publication 2 August 2010
Published 1 September 2010
Online at stacks.iop.org/PhysBio/7/036007

Abstract
The development of aptamers on custom synthesized DNA microarrays, which has been demonstrated in recent publications, can facilitate detailed analyses of sequence and fitness relationships. Here we use the technique to observe the paths taken through sequence-fitness space by three different evolutionary regimes: asexual reproduction, recombination and model-based evolution. The different evolutionary runs are made on the same array chip in triplicate, each one starting from a small population initialized independently at random. When evolving to a common target protein, glucose-6-phosphate dehydrogenase (G6PD), these nine distinct evolutionary runs are observed to develop aptamers with high affinity and to converge on the same motif not present in any of the starting populations. Regime specific differences in the evolutions, such as speed of convergence, could also be observed.

S Online supplementary data available from stacks.iop.org/PhysBio/7/036007/mmedia

1. Introduction

Directed evolution has revolutionized the field of protein engineering exploiting the process of Darwinian selection on a laboratory scale [48]. The technique works towards optimization on the vast landscape of possible sequence permutations by iteratively screening, modifying and selecting from a library of variants based on phenotypic improvement. The successes have been manifold, endowing proteins with improved stability [15], varied specificity and robustness to novel environments [1, 22, 40, 50]. In addition directed evolution has been applied to nucleic acids, for example during the *ab initio* development of nucleic acid ribozymes [23]. For those more accustomed to rational design it is often difficult to perceive that an essentially stochastic process can prove so effective, independent of any knowledge of structure and function. A rarely exploited by-product in the development of novel biological macromolecules through directed evolution is the ability to monitor evolutionary processes as they happen [39]. This has provided insight at the molecular level that is simply unavailable at the level of whole organisms.

The central features of directed evolution (screening, selection and modification) bear close resemblance to evolutionary computation methods used in computer science

[6] Author to whom any correspondence should be addressed.
[7] These authors contributed equally to this work.

and engineering [4, 16, 19]. Yet, despite this similarity there is little transfer of methods and processes between the two disciplines, possibly because of the simple isolation of the two fields. A further reason may be that many of the systems studied in the field of evolutionary computation are *in silico,* and such experiments are cheap in comparison with their *'in vitro'* counterparts. Consequently computational optimization is performed over numerous generations to provide an optimal (or near optimal) solution, whereas the aim of directed evolution is to develop a variant suitable for a particular purpose as quickly as possible. In real terms this means that the directed evolution practitioner often resorts to a method known as 'declaring wild type' [23]. This is an extreme form of selection whereby the best individual from those currently generated is selected to be the parent for future generation of variants until another variant surpasses this individual's fitness. Whilst high selection pressures can reduce the number of generations required in directed evolution experiments, it can also lead to convergence on a highly suboptimal solution [49]. Lower selection pressures are commonly used in genetic algorithms and have been shown to produce superior performance [10, 34].

One area where directed evolution *has* borrowed from evolutionary computation is the incorporation of recombination (crossover). The rationale behind the process known as 'DNA-shuffling' is the belief that low levels of mutation in combination with crossover are sufficient for a genetic algorithm to evolve complex solutions [44]. It is known within the evolutionary computing literature that this is a simplistic representation of optimization as the incorporation of crossover is not guaranteed to be the most efficient method in all instances. Like high selection pressures, crossover may, in some cases, in fact diminish algorithm performance. This is described by the No Free Lunch Theory of optimization that states 'the average performance of any pair of algorithms across all possible problems is identical'. A 'one size fits all' algorithm cannot, therefore, be selected that will be superior on all optimization problems [51]. As a result, any prior knowledge of an algorithm's performance on a given problem is invaluable; however, for biomolecules this knowledge is currently sparse at best.

Like proteins, DNA (as well as RNA) can form complex three-dimensional structures giving them desired functions as, for example, DNAzymes [53] and riboswitches [11]. Oligonucleotide sequences that can selectively bind target molecules are known as aptamers. They are ideal candidates for building sensors, rather like their antibody counterparts. As with proteins, the combinatorial search space for nucleic acids is vast (10^{18} molecules for a 30 mer). Searching these large populations for binding candidates can be fraught with difficulties. Conventionally aptamers are developed using a procedure known as SELEX (systematic evolution of ligands by exponential enrichment), in which sequences with high affinity to a target ligand are iteratively screened, selected and amplified over a number of iterations [13, 47]. SELEX is not directly reliant upon modification to generate genetic diversity to optimize to the best aptamer (the technique relies on the diversity in the starting population), and so the technique differs considerably from the directed evolution of proteins.

Microarrays have recently emerged as a platform for the development of aptamers [3, 27, 38]. In previous work it has been shown that arrays can provide information regarding sequence-functionality relationships in the well-studied thrombin aptamer by systematically mutating the original aptamer sequences generated via SELEX and monitoring the binding intensities [37]. In an experiment which reversed the above procedure, a random population of sequences on a high density microarray was used as the starting population for aptamer development to the thrombin target.

Knight *et al* demonstrated the power of this array-based technique, termed closed loop aptameric directed evolution (CLADE), with the evolution of aptamers with high affinity and specificity to the fluorescent protein allophycocyanin [27]. In CLADE the oligonucleotides on the array are used as a population in a genetic algorithm. Sequences were assayed for their interaction with the target ligand via standard hybridization techniques, and then selection and modification were carried out *in silico*. Sequences for the next generation were produced by the genetic algorithm and synthesized on a custom array synthesiser. This process was carried out over nine generations, after which a sequence with a high affinity to allophycocyanin had evolved. The use of microarrays entails a much smaller initial pool size from which to select aptamers than during SELEX. However, the reduction in initial diversity is compensated by an increase in knowledge, as the sequences of all potential aptamers are known. This knowledge may be used in the development of superior algorithms for modifying and selecting sequences in each generation. In this work, we report a comparison of three different algorithms which result in aptamers with binding strengths comparable to those obtained from SELEX using initial pool sizes of just 1500.

When developing aptamers on microarray surfaces we have the opportunity to assess the performance of several algorithms simultaneously on one array. The spots on the arrays can be subdivided into individual populations, allowing the application of a different algorithm to each sub-population. Further, we show in this paper that the knowledge of sequences obtained through microarray technology enables the observation of any convergent or divergent evolution, by comparing the aptamers generated in each sub-population. As such we can 'replay the evolutionary tape' [17] and study commonalities between different algorithms and within replicate algorithms. While this approach may not inform the development of aptamers through conventional solution-based SELEX, the development of aptamers using a genetic algorithm 'on-chip' requires efficient search of the fitness landscape and so is more closely related to the optimization of proteins in directed evolution.

Aptamer arrays offer the potential for quantification of biological systems at both the metabolomic and proteomic levels [42]. Quantifying metabolically important biological variance is key to the generation of metabolic systems biology models [26]. Glucose-6-phosphate dehydrogenase (G6PD) is an enzyme in the pentose phosphate pathway that serves to generate NADPH and pentose sugars, and is a key biomarker in haemolytic anaemia. The choice of G6PD as a target was made for practical reasons based upon the wider goals of our research

group. As part of the Manchester Centre for Integrative Systems Biology (MCISB) one of our interests is to quantify flux at the transcriptomic, metabolomic and proteomic levels (particularly in the pentose phosphate pathway). We are therefore interested in the feasibility of aptamer arrays for protein quantification.

The evolution of aptamers to G6PD from *Saccharomyces cerevisiae* was studied through subdividing a 90K DNA array into nine separate evolving populations, which represent three different genetic algorithms in triplicate. Each of the algorithms is designed to represent a standard (μ, λ) algorithm [4] with similar parameters to those commonly employed in protein directed evolution; however, one algorithm works through mutation alone, the second incorporates crossover and the third uses a statistical model to reduce the effect of deleterious mutations. We show that despite the differences between these algorithms, performance is comparable, with the algorithms converging on similar sequences. In this case, this is a consequence of the structure of the fitness landscape, which we characterize effectively via the sequence-activity information that is generated 'on chip'.

2. Materials and methods

Microarray synthesis was performed using a Combimatrix B3 custom array synthesiser, details of which are given elsewhere [38]. Microarray chips were made up of 93 311 individual spots, 25 μm in diameter spaced 20 μm apart; each spot is composed of a single sequence that can be custom synthesized. As in the study described by Knight *et al* [27, 38] sequences were synthesized as 30 mers, and therefore had lengths typical of many aptamers. Each 30 mer sequence was synthesized in duplicate within each array and duplicate arrays, each with their own random positioning of sequences, were synthesized.

Bovine serum albumin, potassium chloride, sodium chloride, Tween 20, sodium dihydrogen phosphate and di-sodium hydrogen phosphate (>99.5%) were purchased from Sigma-Aldrich; biotinylated G6PD and unmodified G6PD (Sigma G3386, G4134) were used as purchased without further purification; all hybridizations were performed in phosphate buffer saline (PBS) (1 × PBS = 0.15 M NaCl, 20 mM phosphate buffer (pH 7.2)) at 37 °C and a protein concentration of 0.15 μM. Arrays were first incubated in prehybridization (prehyb) solution (5% BSA 0.5 Tween 1 × PBS) for a period of 30 min. Chips were then washed once with 1 × PBS, before the hybridization solution containing G6PD was incubated with the arrays for 1 h. After hybridization the arrays were washed twice with 1 × PBS before being incubated with Strep-Cy5 (0.02 mg ml^{-1}) for 2 min before two further washes in 1 × PBS and scanning at 5 μm resolution using a Genepix 4000B scanner (Axon instruments).

Median intensity values were recorded for each of the spots and spatial normalization and scaling between chips was performed as described within Knight *et al* [27, 38]. During each generation 2400 spots were replicated from the previous generation to permit inter-generational normalization; these spots were selected by fitness uniform selection [21]. Inter-generational normalization was performed by fitting a linear transformation (orthogonal regression using observed variances) to these replicates from the preceding generation [27, 38]. High correlations of 0.9 were observed between replicate chips. The fluorescence levels observed from the analysis were taken as direct measures of protein binding in our algorithm evaluation. The synthesis of sequences on the array is done using standard amidite chemistry, with a coupling efficiency of approximately 95%. As such the number of complete full length sequences available for binding can be difficult to calculate. To confirm the binding between G6PD and the aptamer sequences, and to remove any doubt that these incomplete/partial sequences are responsible for the observed binding, the binding characteristics were tested on an alternative technology by measuring binding constants using surface plasmon resonance (SPR), as described in the supplementary information available at stacks.iop.org/PhysBio/7/036007/mmedia.

2.1. Algorithm design

The genetic algorithms assessed within this study were based upon a standard (μ, λ) algorithm [4]. In this type of algorithm, μ represents the number of 'parent' sequences in each generation, and λ represents the total number of 'offspring' sequences produced. In our case, in each generation, the λ new offspring sequences are synthesized on the chip and screened for binding affinity. To obtain the μ parent sequences for the subsequent generation from these, the best μ in terms of binding are identified and selected.

The value of λ was 4733 (to allow all experiments of a generation to be conducted on a single microarray), except in generation one (see below). The value of μ was 50 in all cases. In all algorithms, modifications were induced by applying point mutations to each of the bases within the sequence with a probability of 1/30; each mutated base was altered to one of the other three bases, selected at random with uniform probability. In addition, indel mutations (where one base is inserted and then another base deleted randomly from the sequence) were applied at the same rate. If the sequence remained unmodified after this process, the procedure was repeated until the sequence was distinct from the parent sequence.

We use three different algorithms. What differs between them is the method of generating the offspring sequences from the parents, a process which is done entirely *in silico* in all three algorithms. For the simplest algorithm, the mutation-only evolutionary algorithm, the λ offspring are produced from the μ parents by repeatedly (i.e. λ times) selecting a parent at random (with replacement), copying it and mutating the copy. In the recombination algorithm, each offspring is generated either by recombination and mutation or by mutation alone. Recombination is used with probability of 0.6 and mutation alone with a probability of 1–0.6. Standard uniform crossover [45] of two parents was used as the method of recombination; the two parents were selected randomly from the parent population. With this type of crossover there is an equal probability that any particular base will be passed to the child sequence from either of the two parents. The

implementation of crossover was the only feature that differed from the 'mutation only' algorithm as point mutations and indels were applied at the same rate.

The third algorithm used a statistical model to predict sequence binding affinities. The aim of incorporating a statistical model into the directed evolution procedure is to focus selection not just on genes but also on the mutations made to those genes. Fox *et al* applied a partial least-squares model to guide the decision as to which individuals to include in combinatorial libraries in the development of bacterial halohydrin dehalogenase, with improved volumetric production in a cyanation process [14]. This was described as an extension of the QSAR approach commonly used in the development of small molecule drugs.

The applicability of the partial least-squares approach to modelling molecular structure activity relationships has recently been reassessed for QSAR and directed evolution, as the model is unable to detect complex interactions between variables [36]. Non-linear models such as Random Forests [8] are capable of capturing these interactions and may be more appropriate for a wider range of systems. The Random Forest model has already been utilized in the study of aptamer binding relationships, displaying strong predictive power on unseen data [27].

In our model-based algorithm, a Random Forest statistical model (details below) was used to select sequences from a larger pool of child sequences (of size $10 \times \lambda$). These children were generated by either mutation or mutation and crossover *in silico* but not synthesized. From this pool, the best λ were chosen based upon their predicted binding affinity using the model trained with data produced from the previous generations of the evolution. A feature set was generated to describe the sequences, based on position-specific base composition, overall base composition, and the prevalence and position of specific monomers, dimers and trimers.

The algorithms were evaluated over five generations in three independent blocks on the same chip (see figure 1). Within each block there is one replicate of each algorithm type (a total of nine evolutions running in parallel on the same chip). In order for paired statistical tests to be used to compare the performance of the three algorithm types, the algorithms within a block are started from the same initial (generation 1) population. Thus, the 45 000 independent random sequences that are synthesized and assayed in generation 1 are divided into three (and not nine) independent populations: one for each block. Thus, in generation 1, λ is effectively 15 000 for each algorithm. Subsequently, in each generation, each of the nine algorithms will produce its own offspring population of size $\lambda = 4733$, which gets its own space on the chip, and thus these algorithms run independently from generation 2 onwards. In summary, each of the three blocks is entirely independent; within each block the starting population (generation 1) was the same for each of the three different algorithms being tested, allowing the variance due to starting population to be mitigated and thus slightly more powerful statistical tests to be performed when comparing the different algorithm types. (Note: the reason that $\lambda = 4733$ and not 5000 for generations 2–5 is that 267 spots are reserved for the purpose of inter-generational normalization.)

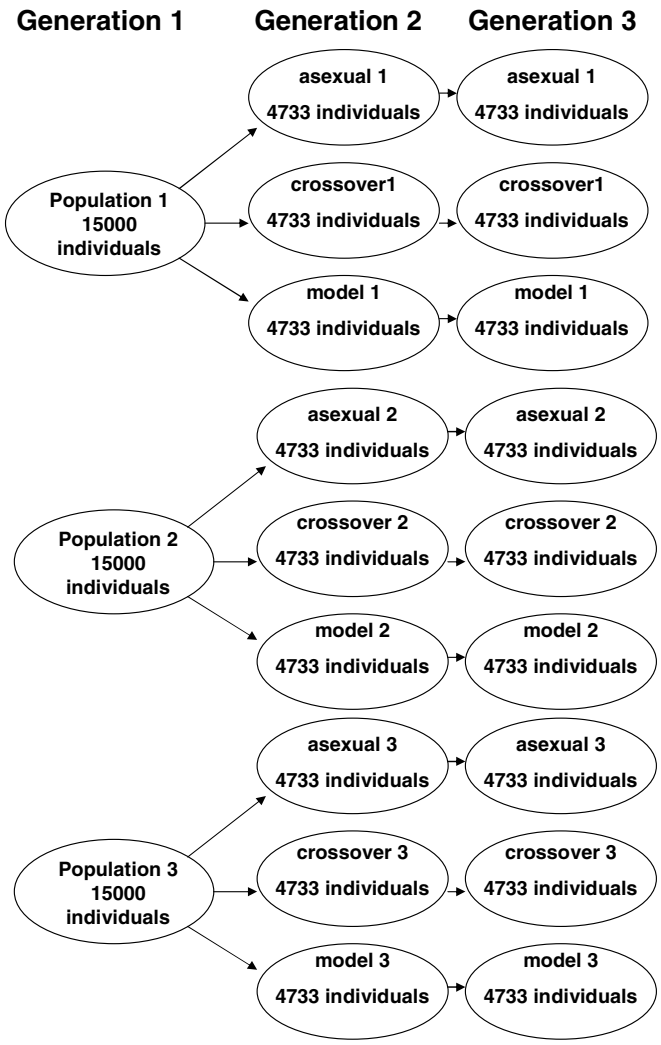

Figure 1. Population construction. Construction of populations within the directed evolution experiment. Sequences on the microarray were first divided into three to form the seeds for nine directed evolution experiments.

3. Results

Nine algorithms were assessed over five generations (asexual, crossover and model based). After five generations the binding constant of the highest affinity sequence from all evolutions was evaluated 'off chip' using SPR (as described in the supplementary information available at stacks.iop.org/PhysBio/7/036007/mmedia). This sequence was produced by the algorithm that included recombination. The sequence 'TTTAGAAGGATTAGTACCTTTTTAAAAAAT' was found to have a K_D of 245 nmol L^{-1}, which is comparable to other aptamers raised to protein targets through SELEX [2, 52]. Typically, we are interested in the best fitness achieved in each generation, rather than the population mean. Figure 2 plots this value for each algorithm (in the last generation of implementation, three of the model-based genetic algorithm, the data became corrupted and were lost, leaving only eight sets of sequences in the final generation). A Wilcoxon paired rank test confirms that there is no significant difference between crossover and asexual

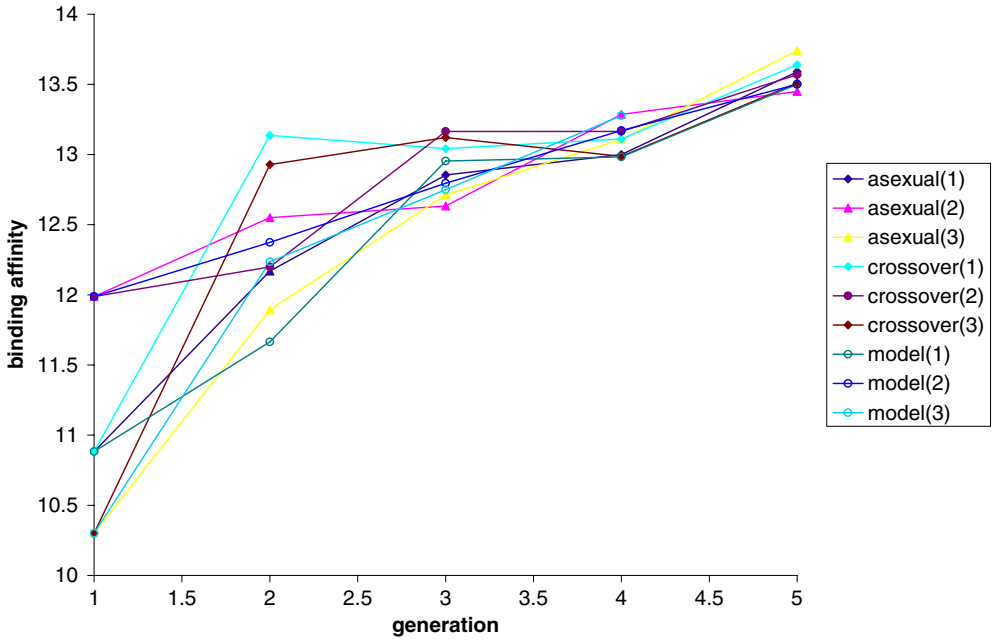

Figure 2. Algorithm performance. Plot displaying the performance of each of the nine algorithms assessed in terms of binding affinity of the best sequence produced in each generation. In the last generation of implementation 3 of the model-based genetic algorithm, the data became corrupted and were lost, leaving only eight sets of sequences in the final generation.

reproduction in terms of the best sequence binder produced in the final generation. While we are unable to determine a superior methodology for the evolution of aptamers to G6PD in these terms, we can study the performance of the separate algorithms in relation to the aptamer fitness landscape. In particular we focus on the performance of crossover versus asexual reproduction and on the relative performance of the model-based algorithm. The trends in the sequences produced by the algorithms and the performance of the algorithms over the four generations give us some insight into the way algorithms may work on other problems and the feasibility of future stepwise evolution experiments.

3.1. Convergent evolution

Convergent evolution describes the phenomenon where similar phenotypic traits evolve in unrelated organisms. Birds, bats and insects developed wings not because of common evolutionary history but because the wing structure represents a general solution to a problem upon which all three have independently converged. This is of course in contrast to the idea that, should the evolutionary tape be replayed, everything would be different. Recurrence is a similar concept to convergent evolution describing the phenomenon whereby different evolving populations follow identical trajectories from the same initial starting point. Recurrence has been demonstrated previously *in vitro* in multiple continuous evolution experiments developing a 150 nucleotide ligase ribozyme [30]. It was observed that in 13 different lineages, whilst sequences varied, the same nine mutations achieved fixation and dominated.

The probability of convergence occurring will depend upon a number of factors, including the rate and type of genetic operator employed, the shape of the biological fitness landscape and the composition of the initial populations [18]. When generating aptamers through stepwise evolutionary methods, such factors are important in determining whether sequences with high affinity to the target ligand can be evolved reproducibly with small initial populations. If the same sequence (or vital feature of a sequence) is repeatedly generated it can be indicative of a peak on the fitness landscape with a broad basin of attraction; if not, the landscape may be rugged with multiple local optima.

When studying the highest affinity sequences produced by the eight completed experiments the number of coincident bases at the 5′ end is striking, indicating a high level of convergence. A multiple sequence alignment generated by CLUSTALW [29] demonstrates that between the best sequences produced by these eight experiments, eight of the first nine bases are fully conserved, with the other base either A or G (see figure 3). It should be remembered that these eight runs are not fully independent. However they are derived from three distinct populations. It is unlikely that this motif has arisen by chance alone (e.g. randomly generated in the first generation) in all eight instances. This motif was not produced in any of the initial populations; the fact that each of the eight implementations produces the same motif indicates that the motif represents a peak on the fitness landscape with a broad basin of attraction.

Analysis of the best sequences produced overall by the asexual algorithm reveals that the nine-base motifs have not arisen randomly in the initial populations but are in fact derived from precursors to this sequence through a series of mutations (3, 2 and 2 mutations respectively; see figure 4).

Beyond the nine-base motif the level of sequence similarity between the highest affinity sequences from

algorithm		binding score
asexual(2)	TTTAGAAGG-A-TTATTTTACGTTTCCC---CCCT----	13.45
crossover(1)	TTTAGAAAGTA-TAATTTGAC-CTACCCA--CCC-----	13.64
model(2)	TTTAGAAGG-ACTAACA-AACG-TGTCGG--TACG----	13.50
asexual(3)	TTTAGAAGG-ATTAGT---ACCTTTTTAAAAAAT-----	13.74
asexual(1)	TTTAGAAAGAA-CAATA-GGTCCC-CGAA--ACCG----	13.59
crossover(3)	TTTAGAAAGAA-CCCCA-GGTCCT-T-ACCCACC-----	13.51
crossover(2)	TTTAGAAAGT--CCCTAAGAACCAGG-AA--ACC-C---	13.57
model(1)	TTTAGAAAG---CCC---GACCCAGT-AA--AGCGTAAG	13.50
	******* *	

Figure 3. Alignment of best sequences. Multiple sequence alignment of the best sequences produced by eight of the algorithms. Binding scores for each sequence are listed.

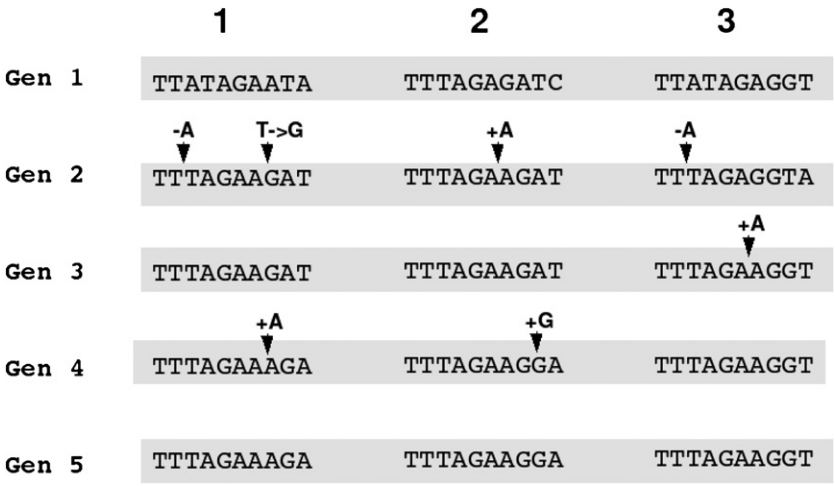

Figure 4. Evolutionary path of asexual algorithms. The evolutionary path of the three implementations of the asexual (μ, λ) algorithm indicating insertions (+), deletions (−) and point mutations (→), leading to the first ten bases of the best sequences in the final generation. These sequences are derived from three different starting populations.

each of the eight populations is not as clear. Sequence similarity was calculated between the eight aptamers using an implementation of the Needleman–Wunsch algorithm [35]. This revealed that on average each pair of aptamers shared 15.39 bases in common, although if the first nine bases are excluded only 6.93 bases are coincident. The latter figure is only slightly greater than the 5.25 (0.25×21) bases we would expect to be coincident by chance alone; this tallies well with the inevitable loss of diversity resulting from selection in a small population (i.e. drift), rather than indicating a particular direction or fitness bias of the search.

3.2. Recombination

Mutation alone (asexual reproduction) has been criticized as a method for creating genetic diversity both within the directed evolution and evolutionary computation literature [25, 41]. It has been shown, especially in small populations, that purely asexual reproduction can lead to the accumulation of deleterious mutations over a period of time to the detriment of the evolutionary process [33]. It is believed that recombination can reduce this effect through the removal of deleterious mutations while combining advantageous mutations.

The benefits of recombination are not assured and in the evolutionary computation literature it is well known that the advantage of employing recombination is dependent on the control parameters of the algorithm (e.g. population size, selection pressure and mutation rate) in addition to the properties of the fitness landscape assessed. There are numerous methods of performing recombination within the directed evolution literature, although direct comparisons of its performance with that of asexual reproduction have been limited [41].

After five generations there appears to be no advantage in using recombination within this study, with performance at each generation being comparable in terms of the affinities of the best aptamers produced (see figure 2). However, after three generations there is a clear advantage with all three implementations outperforming the six implementations of the other two algorithms. It is after this that the performances of the algorithms converge to similar levels.

3.3. Model-based genetic algorithm performance

The accuracy of the predictions made by the Random Forest at each stage of the evolution was assessed retrospectively with unseen data. These data were selected from the sequences that were not derived from the same starting population. One thousand points were selected by fitness uniform selection [21] and used to test the performance of the Random Forest using accumulated data from each of the generations produced

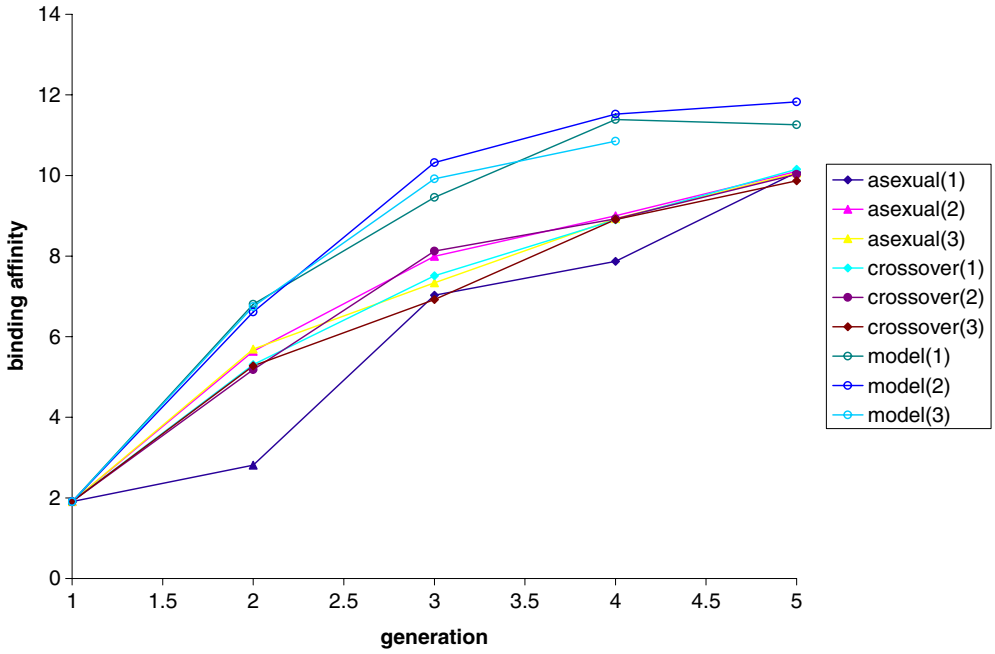

Figure 5. Mean binding scores. Plot displaying the performance of each of the nine algorithms assessed in terms of mean binding affinity of all sequences produced in each generation.

Table 1. Displaying correlation (Pearson correlation coefficient) between predicted and observed values for unseen values for a Random Forest model constructed with data produced from each generation of the evolution.

Generation	Accuracy
1	0.507
2	0.609
3	0.852
4	0.884

by implementation 1 of the model-based genetic algorithm. Model accuracy (as measured by the Pearson correlation coefficient of observed versus predicted scores) increases at each generation, reaching 0.88 by the final generation (see table 1). This indicates an improvement in the ability of the Random Forest to characterize the features that determine aptamer binding. This is the result of the greater volume of data within the training set in addition to a greater proportion of the data coming from sequences with higher affinities.

The accuracy of the model is reflected in the mean sequence affinity in each generation being higher than that for all other algorithms (see figure 5). Given this, why does the model-based algorithm not outperform the other two algorithms in terms of the *best* sequence produced within each generation? The entropy of the sequence describes the variability observed in an aligned column of bases (the uncertainty) [43]. The greater the variability within the column, the greater the entropy, so a column of random variables would have a value of 2, whereas a completely conserved column has a value of 0. Figure 6 shows a plot of the total uncertainty over all columns (bases). It can be seen that in the first generation the entropy in each algorithm is 60 indicating (as expected) completely random sequences. As each of the algorithms progresses the entropy decreases, and this is most evident within the model-based evolution. At each stage of the evolution the entropy for the model-based algorithm is lower than that of the other algorithms. This is a result of the model-based algorithm fixing bases it 'perceives' to be important to protein binding. Fixing bases reduces the search space that the algorithm can cover, which can have a beneficial or detrimental effect on algorithm performance. This balance of exploration versus exploitation is a well-known issue within the evolutionary computation literature [12]. Employing model-based selection of variants is similar to employing greater selection pressure in evolutionary search in that it strengthens the role of exploitation while limiting that of exploration. As such the population may be more prone to convergence to a local optimum.

The loss of diversity we observed in our model-based genetic algorithm does not rule out the possibility that other model-based methods may be able to raise the progress of evolution without causing such loss. For example, the use of a better cost function for training the model [7], less greedy selection based on the model predictions [6], or models that estimate their own uncertainty [21] may all go some way to preventing the kind of diversity loss we observed with our model-based algorithm.

3.4. Analysis of the landscape

While the Random Forest model can accurately predict the binding affinities of previously unseen sequences, the processes upon which it makes these decisions are opaque to the investigator. Like Random Forests, regression trees [9] are capable of modelling non-linear interactions between features. However, they have the advantage over the Random Forest model that the criteria which they use to make predictions are

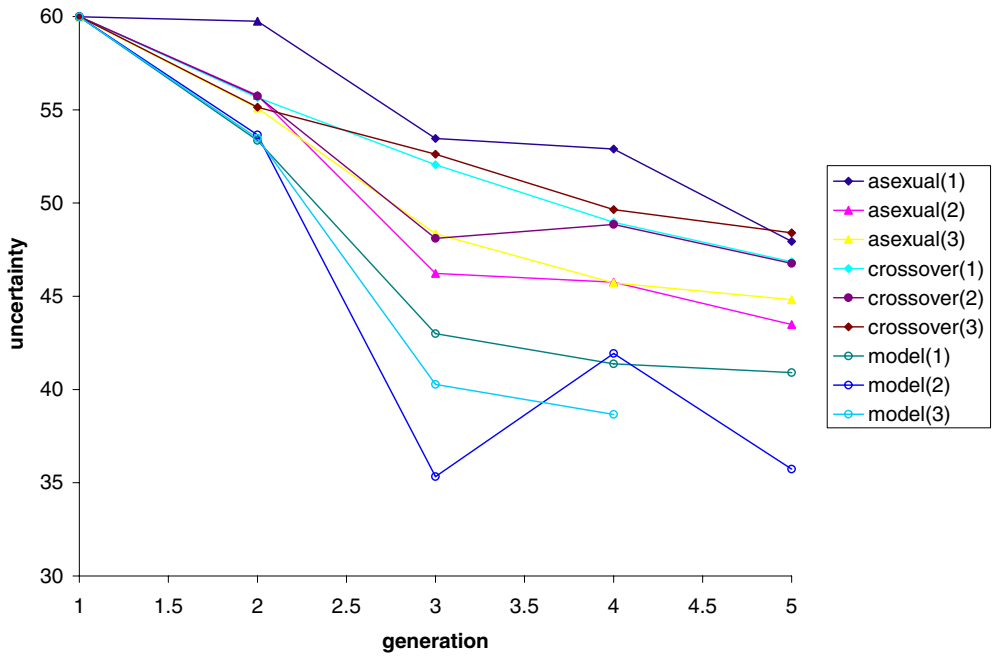

Figure 6. Entropy of sequences. Plot displaying the entropy (uncertainty) of the sequences for each of the nine algorithms assessed within each generation.

intelligible to the user. In the case of our aptamer dataset this should provide a clear indication of the sequence/structural features that determine protein binding. If we can determine these features we can make interpretations on the structure of the landscape and how this will have affected algorithm performance.

The sequences produced over five generations by implementation 1 of the model-based genetic algorithm were utilized as the training set for our regression tree model and the sequences generated by the other algorithms (previously used to test the Random Forest model) were employed as a test set. In addition to the features used to train the Random Forest model additional features were added, characterizing both predicted structures and commonly occurring motifs. Minimum free energy structures for each of the sequences were predicted using hybrid-ss-min (UNAFold) version 3.4 [31], using the DNA input mode. The temperature was set at 38 °C and the NaCl concentration given as 0.58 M (these values correspond to the conditions under which protein hybridization was performed). These structures were converted into discrete features using parsing software developed in-house.

Over-represented motifs were identified in 100 sequences from implementation 1 of the model-based genetic algorithm using MotifSampler v3.2 against a random background model [46]. To avoid partially the problems associated with a strong phylogenetic structure within the data, the 100 sequences were not the top 100 sequences in terms of binding affinity. Rather, a level of diversity between the sequences was enforced by selecting the top 100 sequences with fewer than 15 coincident bases (as measured using an implementation of the Needleman–Wunsch algorithm [35]) from the five generations. The training and testing sets were scanned using MotifScanner v3.2 for the five top motifs identified by MotifSampler and the start points and scores of the highest scoring occurrences were recorded.

The regression tree was built using a recursive partitioning algorithm implemented using the rpart package within R and pruning of the tree was performed according to Breiman's 'one standard error' rule [7]. The correlation between observed and predicted binding scores in the testing set was found to be 0.80 (somewhat lower than the Random Forest model). From the representation of the regression tree in figure 7 it can be seen that prediction of binding is based exclusively on features associated with the TTTAGAmn motif. The position and score of this motif appear to be the primary determinants of protein binding. These results may explain why there is parity between genetic algorithm performances, as optimizing eight bases in a fixed position within a DNA sequence is trivial for a genetic algorithm given the number of evaluations. Therefore it will be hard in this case to discriminate between the different forms of evolutionary search.

4. Discussion

Previous evolutions of aptamers to the proteins thrombin and allophycocyanin on microarrays have also indicated strong dependences on the bases at the 5′ end of the DNA sequence [27, 38]. These sequences are projected furthest away from the microarray surface and so are more likely to interact with the protein target. Aptamers derived through SELEX are often characterized by their secondary and tertiary structures. However, no structural features (based on the predicted secondary structure) were found to be associated with protein binding. Known aptamers with complex secondary and tertiary structures have been shown to be capable of binding their target ligands when synthesized on combimatrix 90k

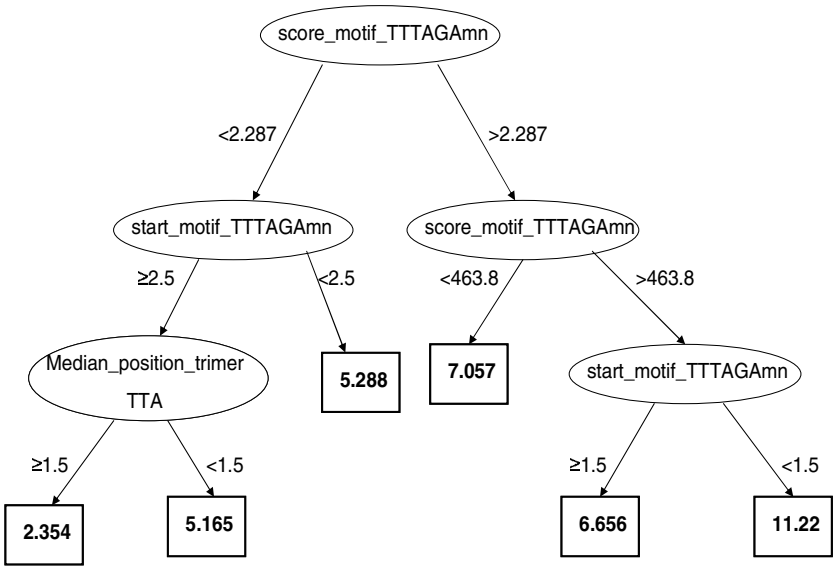

Figure 7. Decision tree derived from sequence-binding relationships. Displaying a decision tree built using data compiled from a single implementation of one of the algorithms assessed in this paper (model-based algorithm, replicate 1).

arrays [37] and so there is no physical limitation to evolving aptamers with secondary structures. The question arises: can complex structures be evolved by stepwise evolutionary methods using relatively small populations?

The development of aptamers through SELEX works not through directed search but by exploiting the immense sequence coverage generated in the initial libraries via purifying selection. Genetic algorithms, as used in this study, work differently to this. Strong binding aptamers are obtained because we use parameters for the evolution selection pressure and variation schemes (mutation or recombination) that are theoretically justified [10, 34] under mild assumptions, and have been empirically confirmed to work numerous times. The selection pressure we use, for instance, is much smaller than that used in SELEX, i.e. fitter sequences are only slightly more likely to have 'offspring' sequences than their less fit counterparts. This means that if no strong binders are found initially, the population will remain diverse and will continue to explore the fitness landscape via mutation. Only when stronger binders are found, do they get selected more frequently, but even then the evolution is gradual, and allows further exploration to occur. This is why nine different evolutions starting from three entirely independent small populations can arrive at the same motif. In the model-based algorithm we also use sequence information to directly guide selection, but this seems to make only a small improvement to the rate of evolutionary progress.

It has previously been stated in a study investigating sequence variants of the IgE aptamer [24] that, since all of the 1-, 2- and 3-base mutants of the aptamer have poor IgE binding, it would be difficult to generate the aptamer through stepwise evolutionary methods. Similarly, binding of the thrombin aptamer has been shown to be dependent on an immutable quartet of guanine repeats that fold to form a unimolecular quadruplex structure. The consensus sequence of the thrombin aptamer, 'GG(A/T)TGGN3–5GGT(A/T)GG', has a 0.55 probability of being generated randomly within a 30 mer in 40 000 sequences on a microarray [38]. If a variant of this sequence can be found with a higher binding affinity than the background then this sequence can be selected and optimized. If the consensus sequence is immutable and such a sequence is not present then the evolution will fail. This scenario is described in the evolutionary computation literature as 'isolation' or 'needle in a haystack' [20], and represents a difficult search problem for genetic algorithms. Aptamers have been developed to the protein thrombin on DNA microarrays with populations of around 40 000 sequences [38]. However, the sequences developed did not possess the usual consensus sequence of a thrombin aptamer developed through SELEX. Instead binding was shown to be dependent on a short motif of GGTTGG, again at the 5′ end of the oligonucleotide sequence.

The short motif 'GGTTGG' was generated in the first generation (of random individuals). Binding to the aptamer was then optimized over the course of the evolution by developing a chain of thymine bases that may serve as a linker to prevent interaction between the protein and the array surface [38]. While the motif identified here is short, it has not occurred by chance. Rather, it is derived from (differing) precursor sequences that have evolved to a common sequence (see figure 4). This result indicates a biological fitness landscape with a broad basin of attraction, rather than a 'needle in a haystack' type landscape. This result is encouraging for the potential to evolve other aptamers on microarrays.

It should be noted that despite the aptamers evolved on microarrays being characterized by short motifs at the 5′ end of the sequence, they are capable of binding target ligands with binding constants comparable to more orthodox aptamer structures. This observation fits a trend that we previously noted during the evolution of aptamers to both thrombin and APC [27]. It is clear that a portion of the sequence close to the 3′ end is used as a spacer to project the binding site away from the surface. Due to this unique arrangement we

chose a surface-based technique to determine their binding characteristics, i.e. SPR. Whilst solution-based assays, such as calorimetry or the filter binding method, reveal a great deal of information about the binding strength and kinetics, CLADE was developed as a surface-based assay technique and as such we have opted to measure K_D in the environment in which it will ultimately be used. The tethering of one end of the sequence to a surface will most likely change its binding kinetics from those observed in its free state in solution. However, this enhanced or altered kinetics can be used as a powerful assay format [28].

No higher (secondary or tertiary) structural features were determined to be causal to protein binding in the G6PD aptamer. This is not to say that the aptamer does not possess structure. As the TTTAGAA(A/G)G motif does not vary significantly we cannot determine any covariance, which could indicate the formation of a small hairpin and so only structural studies (such as by NMR or protein crystallography) would reveal the conformation this motif adopts when binding the protein.

In evolutionary computation, the choice of control parameters can greatly affect performance for a given problem. As a result of the inherent expense of laboratory-based experiments in directed evolution and particularly array-based evolution of aptamers, this becomes of paramount importance. While in this instance we cannot determine a significant difference in the performance of the three algorithms assessed in terms of best aptamer produced, we can identify differences in how the algorithms are working. The algorithm incorporating recombination appears to converge on a high affinity sequence quicker than the other two algorithms. In addition we can define properties of the search space and how the algorithms perform relative to this landscape. This knowledge arises from evolution on the microarray platform, which permits a level of understanding of genotype-phenotype relationships that is currently unsurpassed. Microarrays have been used recently in the study of the interaction of transcription factors with all ten base pair permutations of a DNA sequence [5]. To study the interaction of proteins with more complex (longer) sequences, algorithms similar to those presented here could be applied to sample the sequence landscape and generate a model of the interaction profile. While this magnitude of information is not currently available for the directed evolution of proteins, advances in next generation sequencing technologies should aid our knowledge of these more complex fitness landscapes and aid in the design of evolutionary optimization.

The evolution of such short motifs is reminiscent of the word game described by John Maynard Smith as an analogy for protein spaces [32]. He compared protein evolution to words that could be converted to new viable words through a change of letter. We observe a biological fitness landscape where a network of interchangeable sequences represents a broad basin of attraction, leading to an optimal motif produced by all eight algorithms. This motif has not arisen randomly in each of the populations and then become fixed; it has evolved in each instance from an inferior sequence through a series of mutations, each of which has conferred a selective advantage on the mutant. Such adaptation is essential for stepwise evolution to be a viable method for developing aptamers.

Acknowledgments

We acknowledge sponsorship by the Biotechnology and Biological Sciences Research Council PBB/D000203/1. This is a contribution from the Manchester Centre for Integrative Systems Biology (www.mcisb.org/).

References

[1] Alexeeva M, Carr R and Turner N J 2003 Directed evolution of enzymes: new biocatalysts for asymmetric synthesis *Org. Biomol. Chem.* **1** 4133–7
[2] Allen P, Worland S and Gold L 1995 Isolation of high-affinity RNA ligands to HIV-1 integrase from a random pool *Virology* **209** 327–36
[3] Asai R, Nishimura S I, Aita T and Takahashi K 2005 *In vitro* selection of DNA aptamers on chips using a method for generating point mutations *Analytical Letters* (London: Taylor and Francis) pp 645–56
[4] Bäck T and Fogel D B 1997 *Handbook of Evolutionary Computation* (London: Taylor and Francis)
[5] Badis G *et al* 2009 Diversity and complexity in DNA recognition by transcription factors *Science* **324** 1720–3
[6] Berry D A and Fristedt B 1985 *Bandit Problems: Sequential Allocation of Experiments (Monographs on Statistics and Applied Probability)* (London: Chapman & Hall)
[7] Bishop C M 2006 *Pattern Recognition and Machine Learning* (New York: Springer)
[8] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[9] Breiman L, Friedman J, Stone C and Olshen R A 1984 *Classification and Regression Trees* (London: Chapman and Hall/CRC)
[10] Deb K and Agrawal S 1999 Understanding interactions among genetic algorithm parameters *Found. Genet. Algorithms* **5** 265–86
[11] Dixon N, Duncan J N, Geerlings T, Dunstan M S, McCarthy J E G, Leys D and Micklefield J 2010 Reengineering orthogonally selective riboswitches *Proc. Natl Acad. Sci. USA* **107** 2830–5
[12] Eiben A E and Schippers C A 1998 On evolutionary exploration and exploitation *Fundam. Inform.* **35** 35–50
[13] Ellington A D and Szostak J W 1990 *In vitro* selection of RNA molecules that bind specific ligands *Nature* **346** 818–22
[14] Fox R J *et al* 2007 Improving catalytic function by ProSAR-driven enzyme evolution *Nat. Biotechnol.* **25** 338–44
[15] Giver L, Gershenson A, Freskgard P O and Arnold F H 1998 Directed evolution of a thermostable esterase *Proc. Natl Acad. Sci. USA* **95** 12809–13
[16] Goldberg D E 1989 *Genetic Algorithms in Search and Optimization* (Reading, MA: Addison-Wesley)
[17] Gould S J 1989 *Wonderful Life: The Burgess Shale and the Nature of History* (New York: W W Norton)
[18] Hanczyc M M and Dorit R L 2000 Replicability and recurrence in the experimental evolution of a group I ribozyme *Mol. Biol. Evol.* **17** 1050–60
[19] Holland J H 1975 An introduction with application to biology, control and artificial intelligence *Adaptation in Natural and Artificial System* (Cambridge, MA: MIT Press)
[20] Horn J, Goldberg D E and Deb K 1994 Long path problems *Proc. 3rd Conf. on Parallel Problems Solving from Nature* pp 149–58

[21] Hutter M 2002 Fitness uniform selection to preserve genetic diversity *E-Commerce Technology, IEEE Int. Conf.* pp 783–8
[22] Jackel C, Kast P and Hilvert D 2008 Protein design by directed evolution *Annu. Rev. Biophys.* **37** 153–73
[23] Joyce G F 2004 Directed evolution of nucleic acid enzymes *Annu. Rev. Biochem.* **73** 791–836
[24] Katilius E, Flores C and Woodbury N W 2007 Exploring the sequence space of a DNA aptamer using microarrays *Nucleic Acids Res.* **35** 7626–35
[25] Kauffman S and Levin S 1987 Towards a general-theory of adaptive walks on rugged landscapes *J. Theor. Biol.* **128** 11–45
[26] Kell D B 2006 Systems biology, metabolic modelling and metabolomics in drug discovery and development *Drug Discov. Today* **11** 1085–92
[27] Knight C G, Platt M, Rowe W, Wedge D C, Khan F, Day P J, McShea A, Knowles J and Kell D B 2009 Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape *Nucleic Acids Res.* **37** e6
[28] Kraemer S *et al* 2010 From aptamer-based biomarker discovery to diagnostic and clinical applications: an aptamer-based, streamlined multiplex proteomic assay available from *Nature Precedings* http://hdl.handle.net/10101/npre.2010.4642.1
[29] Larkin M A *et al* 2007 Clustal W and Clustal X version 2.0 *Bioinformatics* **23** 2947–8
[30] Lehman N 2004 Assessing the likelihood of recurrence during RNA evolution *in vitro Artif. Life* **10** 1–22
[31] Markham N R and Zuker M 2008 UNAFold: software for nucleic acid folding and hybridization *Methods Mol. Biol.* **453** 3–31
[32] Maynard Smith J 1970 Natural selection and the concept of a protein space *Nature* **225** 563–4
[33] Maynard Smith J 1988 The evolution of recombination *The Evolution of Sex: An Examination of Current Ideas* (Sunderland, MA: Sinauer Associates, Inc.)
[34] Mühlenbein H 1992 How genetic algorithms really work: I. Mutation and hillclimbing *Parallel Problem Solving from Nature* vol 2 pp 15–25
[35] Needleman S B and Wunsch C D 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins *J. Mol. Biol.* **48** 443–53
[36] Patel Y, Gillet V J, Howe T, Pastor J, Oyarzabal J and Willett P 2008 Assessment of additive/nonadditive effects in structure-activity relationships: implications for iterative drug design *J. Med. Chem.* **51** 7552–62
[37] Platt M, Rowe W, Knowles J, Day P J and Kell D B 2009 Analysis of aptamer sequence activity relationships *Integr. Biol.* **1** 116–22
[38] Platt M, Rowe W, Wedge D C, Kell D B, Knowles J and Day P J 2009 Aptamer evolution for array-based diagnostics *Anal. Biochem.* **390** 203–5
[39] Poelwijk F J, Kiviet D J, Weinreich D M and Tans S J 2007 Empirical fitness landscapes reveal accessible evolutionary paths *Nature* **445** 383–6
[40] Romero P A and Arnold F H 2009 Exploring protein fitness landscapes by directed evolution *Nat. Rev. Mol. Cell Biol.* **10** 866–76
[41] Rowe L A, Geddie M L, Alexander O B and Matsumura I 2003 A comparison of directed evolution approaches using the beta-glucuronidase model system *J. Mol. Biol.* **332** 851–60
[42] Rowe W, Platt M and Day P J 2009 Advances and perspectives in aptamer arrays *Integr. Biol.* **1** 53–8
[43] Shannon C E 1951 Prediction and entropy of printed English *Bell Syst. Tech. J.* **30** 50–64
[44] Stemmer W P 1994 DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution *Proc. Natl Acad. Sci. USA* **91** 10747–51
[45] Sywerda G 1989 Uniform crossover in genetic algorithms *Proc. 3rd Int. Conf. on Genetic Algorithms, George Mason University, USA* (San Mateo, CA: Morgan Kaufmann Publishers)
[46] Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P and Moreau Y 2001 A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling *Bioinformatics* **17** 1113–22
[47] Tuerk C and Gold L 1990 Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase *Science* **249** 505–10
[48] Voigt C A, Kauffman S and Wang Z G 2000 Rational evolutionary design: the theory of *in vitro* protein evolution *Adv. Protein Chem.* **55** 79–160
[49] Wedge D C, Rowe W, Kell D B and Knowles J 2009 *In silico* modelling of directed evolution: implications for experimental design and stepwise evolution *J. Theor. Biol.* **257** 131–41
[50] Williams G J, Nelson A S and Berry A 2004 Directed evolution of enzymes for biocatalysis and the life sciences *Cell. Mol. Life Sci.* **61** 3034–46
[51] Wolpert D H and Macready W G 1997 No free lunch theorems for optimization *IEEE Trans. Evol. Comput.* **1** 67–82
[52] Zhan L-S, Zhuo H-L and Wang H-Z 2005 Screening and characterization of aptamers of hepatitis C virus NS3 helicase *Prog. Biochem. Biophys.* **32** 245–50
[53] Zhou M G, Liang X G, Mochizuki T and Asanuma H 2010 A light-driven DNA nanomachine for the efficient photoswitching of RNA digestion *Angew. Chem., Int. Ed.* **49** 2167–70