Published online 23 November 2008

Nucleic Acids Research, 2009, Vol. 37, No. 1 ***e6***
doi:10.1093/nar/gkn899

Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape

Christopher G. Knight[1,2,3,*], Mark Platt[1,2], William Rowe[1,2], David C. Wedge[1,2], Farid Khan[1,2], Philip J. R. Day[1,4], Andy McShea[5], Joshua Knowles[1,6] and Douglas B. Kell[1,2]

[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]Faculty of Life Sciences, The University of Manchester, Simon Building, Brunswick Street, Manchester M13 9PL, UK, [4]School of Medicine, The University of Manchester, Oxford Road, Manchester M13 9PT, UK, [5]Combimatrix Corporation, 6500 Harbor Heights Parkway, Suite #303, Mukilteo, WA 98275, USA and [6]School of Computer Science, The University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL, UK

Received September 20, 2008; Revised October 20, 2008; Accepted October 23, 2008

ABSTRACT

Mapping the landscape of possible macromolecular polymer sequences to their fitness in performing biological functions is a challenge across the biosciences. A paradigm is the case of aptamers, nucleic acids that can be selected to bind particular target molecules. We have characterized the sequence-fitness landscape for aptamers binding allophycocyanin (APC) protein via a novel Closed Loop Aptameric Directed Evolution (CLADE) approach. In contrast to the conventional SELEX methodology, selection and mutation of aptamer sequences was carried out *in silico*, with explicit fitness assays for 44 131 aptamers of known sequence using DNA microarrays *in vitro*. We capture the landscape using a predictive machine learning model linking sequence features and function and validate this model using 5500 entirely separate test sequences, which give a very high observed versus predicted correlation of 0.87. This approach reveals a complex sequence-fitness mapping, and hypotheses for the physical basis of aptameric binding; it also enables rapid design of novel aptamers with desired binding properties. We demonstrate an extension to the approach by incorporating prior knowledge into CLADE, resulting in some of the tightest binding sequences.

INTRODUCTION

Mapping between genotype and phenotype, and more specifically, understanding the landscape of fitness over the complete set of theoretically possible genotypes ('genotypic sequence space') is a key and longstanding challenge across the biosciences (1). Such landscapes are typically very large and potentially complex (2). Questions about genotype–phenotype mapping are most acute, and potentially most tractable, when genotype and phenotype are very closely allied, i.e. fitness is determined directly by the primary nucleotide sequence, without mediation by transcription or translation. Thus models have recently been developed for nucleosome positioning (3) and transcription factor binding (4) *in vivo* and *in vitro*, respectively. However, the most insightful and comprehensive fitness landscape work has been carried out in theory (5) and *in silico*, notably with nucleic acid structure prediction (6,7). The success of such *in silico* structure systems comes from the ability to carry out evolutionary computing experiments which explore otherwise intractably large regions of sequence space (8). For instance, a 30 nt sequence, as used here, lies within a landscape of $4^{30} \approx 10^{18}$ possible sequences. Even with the smallest standard 5 μm

*To whom correspondence should be addressed. Tel: +44 161 2755378; Email: chris.knight@manchester.ac.uk
Correspondence may also be addressed to Douglas B. Kell. Tel: +44 161 306 4492; Fax: +44 161 306 4556; Email: dbk@manchester.ac.uk
Present address:
Andy McShea, Theo Chocolate, 3400 Phinney Ave. N., Seattle, WA 98103, USA

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

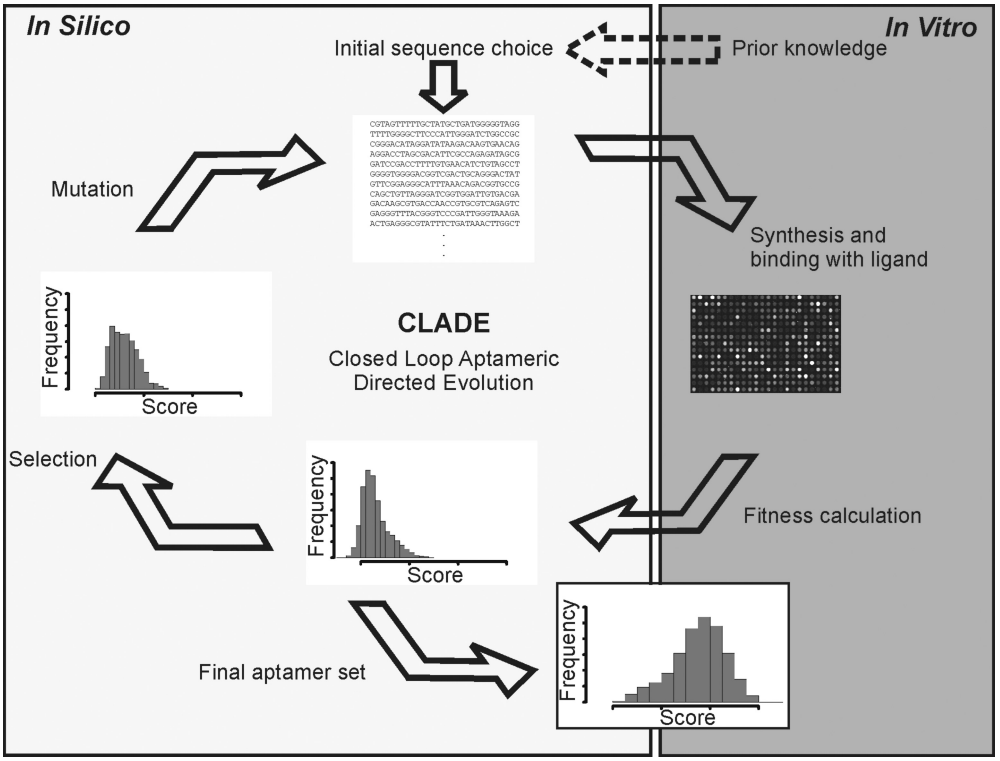

Figure 1. Schematic of the CLADE approach. CLADE as utilized in this study starts at the top of the schematic with an initial choice of DNA sequences. These sequences may be generated entirely *in silico*, or optionally, as with some of the sequences here, utilizing prior knowledge generated *in vitro*. These sequences are synthesized on a custom microarray and bound with the chosen ligand, here the APC protein. Analysis of binding intensities gives a distribution of fitnesses; the frequency distribution of Generation 1 binding to APC protein is shown by way of example. Some of these sequences are selected *in silico*, based on the *in vitro* score distribution, here using tournament selection (see Materials and methods section). These sequences are then mutated *in silico* to generate a new sequence set which can then be synthesised *in vitro*, and so on round the cycle as often as is required. The final aptamer set offers a greatly increased binding affinity to the ligand.

microarray features, a 29 km^2 array would be needed to assay them all *in vitro*. Here, we combine powerful aspects of an *in silico* evolutionary approach with an *in vitro* system to explore and characterize the sequence-fitness landscape for a specific bio-molecular interaction.

Aptamers are nucleic acids selected to bind particular target molecules (9,10). Though small, they can form complex structures, with or without their target molecule (11,12). Typically, aptamers are identified within sequence space via an evolutionary process (SELEX) (9,10) involving sequential rounds of *in vitro* selection from a large pool, possibly including some mutation (13). The variety of sequences and structures and the population dynamics can be complex (14,15); yet this 'black-box' process typically yields only a small number of known aptamer sequences at the end, and hence little information on the sequence-fitness landscape. In contrast, we assay known DNA sequences for aptameric binding using bespoke microarrays (16). On the basis of fitness values measured in the *in vitro* assay, selection and mutation of aptamer sequences is carried out explicitly, *in silico* (17) using known sequences at each generation. This yields a more directed, diverse and transparent evolution. To demonstrate our approach, that we call Closed Loop Aptameric Directed Evolution (hereafter CLADE, Figure 1), we choose as our binding target a large (110 kDa) fluorescent protein, allophycocyanin (APC). APC is an important fluorescent reagent widely used in drug discovery. Aptamers to such fluorescent proteins, as developed here, have great potential application in protein localization *in vivo* (18). More generally, proteins are important targets for specific binding to aptamers, offering the potential for protein-binding arrays (19), with many advantages over alternative technologies such as antibodies (20,21).

This study comprises two conceptually separate parts: (i) The evolution of DNA aptamers to the APC protein using the CLADE *in vitro*, *in silico* approach; (ii) The analysis and modelling of the sequence-fitness landscape for these aptamers. These are respectively the subjects of the first two sections of the results. The first could be carried out without the second and, in principle, given an appropriate dataset from another source, the second could be carried out without the first. These parts are however connected in that it is only the large number of sequence-fitness pairs that explore the landscape in an efficient way, as generated by CLADE, that enable effective modelling of such a landscape. Further, an important feature of CLADE is that it can incorporate prior knowledge into aptamer generation (Figure 1) and combining CLADE with modelling, as in this study, is itself one way to generate appropriate 'prior knowledge'. Incorporating prior knowledge into CLADE is the subject of the third section of the results. Doing so enhances the

CLADE process both in creating potentially useful aptamers and in generating understanding of a sequence-fitness landscape.

MATERIALS AND METHODS

Array synthesis and binding

Six thousand 30 nt DNA sequences were synthesized in duplicate on Combimatrix B3 synthesizer (for a detailed protocol see www.combimatrix.com). Briefly, each chip contains 12 544 spots onto which chosen sequences are synthesized electrochemically. A total of 544 spots are used for fabrication and quality control. The spatial position of each replicate of each aptamer sequence was randomized independently for each chip.

In vitro assays were performed in phosphate buffered saline (1× PBS = 0.15 M NaCl, 20 mM phosphate buffer, pH 5.4) at 37°C with 0.01 mg/ml APC protein. Prior to binding all chips were incubated with a pre-binding solution (5% BSA, 0.5% Tween 1× PBS) for 30 min at 37°C. Chips were incubated with the binding solution (APC 0.01 mg/ml, 1× PBS) for 1 h. After this the chips are washed five times in 1× PBS at room temperature and imaged immediately. Details of how the assay conditions were chosen are given in Supplementary Methods.

Absolute binding measurement

Absolute binding was assayed by Surface Plasmon Resonance (SPR) in a Biacore 3000 instrument. Two DNA sequences were taken from the aptamer evolution, a low binding-score sequence from the first generation (C1.0489, TTAAGGAATACATAGATTATATGGCA AGTT score 5.8 ± 0.1) and a high binding-score sequence from the final generation (G9.3415, ATCCCCCCCTC CCCTTATGTGCACCCGCAT score 12.6 ± 0.8). These sequences were synthesized with a 3′-biotin tag and immobilized on a streptavidin coated chip in adjacent flow cells, causing an increase in ~1700 resonance units (RU) in each case. One blank flow cell consisting of streptavidin only (without aptamer) was used as the control. Solutions of APC in the buffer used for binding in the evolution were passed over all of the flow cells at 20 µl min^{-1} at 37°C, and in each binding cycle the chip was regenerated with glycine buffer pH 1.5. Binding curves for APC at 214, 97, 76, 53 and 31 nM are shown in Figure 3 for the final generation sequence. Each curve was blank subtracted using the control flow cell. Curves were fitted separately to the data for 'on' and 'off' rates at each concentration using the BIA-evaluation 4.1 software.

Microarray analysis. Images were scanned at 5 µm resolution and any artefacts, damaged or obscured spots were excluded by visual inspection. Median intensity values were recorded for each spot. The principal systematic spatial effects we have identified on these chips are smooth spatial gradients. These were removed by normalizing median intensity values to a smooth surface fitted for each chip by means of a spatial general additive model (GAM) in which the degree smoothing is chosen as part of the fit (22) i.e. probe intensities were multiplied by the median value of the fitted surface divided by the surface value at the probe location. The mean-variance relationship in the data was then removed simultaneously with normalization to the same scale among chips within a generation by fitting a generalized log transform model (23). This yields data on a uniform, though arbitrary scale, where zero is below background binding levels. The resulting binding score estimates for each sequence were then averaged (first within, then between chips) to give a single score for each sequence. The error values quoted were calculated as the standard error (SE) among chips (i.e. the SE of the chip means for each sequence). Scores were then normalized to the preceding generation's scores as follows: A linear transformation was estimated by fitting a line (orthogonal regression, using observed variances) to the plot of scores in the new generation vs. scores in the preceding generation for all identical sequences appearing in both generations (control sequences and sequences selected but by chance not mutated, usually a total of around 1200 sequences). Test chip scores were placed back onto the same scale as that used for the chips in the evolution in a similar manner, but using the 500 control probes only. These 500 control sequences were chosen to represent as evenly as possible the range of binding seen in the course of the optimisation of assay conditions (see Supplementary Methods for details).

Evolution

The CLADE methodology comprises a repeated cycle of sequence synthesis, *in vitro* assay, *in silico* selection and *in silico* mutation (Figure 1). G1 Sequences comprised the 500 controls and 5500 30-base ($L = 30$) sequences generated with an equal probability of each base at each position. Following the *in vitro* assay, selection in G1 was from all 6000 sequences; in later generations only non-control sequences were used. 5500 sequences were selected *in silico* by a 'tournament selection'. In each case, the best scoring of four sequences picked uniformly at random with replacement from the current generation of sequences was taken. Selected sequences were subjected to point mutations and insertion–deletion events (indels). Point mutation comprised replacement of a base by a base generated uniformly at random with a probability of $1/L$ at each position. Indels were applied subsequently also with a probability of $1/L$ at each position. Indels were effected by inserting a new base (equal probability of all four bases) at a given position, followed by deletion of a base at a random position. These new sequences were then synthesized on a microarray and the analytical binding process repeated for a total of nine generations, i.e. until G9 had been synthesized, assayed and scored.

In order to demonstrate the validity of the CLADE method, a relatively simple *in silico* evolutionary algorithm was used (8). Other operations, including recombination could be in principle included. Recombination might be expected to find peaks in the landscape more effectively, though that would depend upon the ruggedness of that landscape. Theoretical studies also suggest that alternative parameter settings, including higher

mutation rates and higher selection pressures, may result in improved performance (24). Empirical validation of these findings will be the subject of further study.

Sequence analysis

Three hundred and two candidate explanatory variables (Supplementary Table 3) were extracted from each sequence. These comprised the frequency and median position of all single, double and triple base sequences, as well as 30 4-level factors for the particular base at each position and 29 16-level factors for each pair of bases in each position. Motifs characteristic of high or low scoring sequences were identified using MotifScanner v3.2 and associated software (25). Motifs were identified in two ways. Firstly, motifs were searched for using all the top and bottom scoring probes from each lineage. Motifs were identified in the top scoring probes, assuming a background model derived from the bottom scoring probes and vice versa. Lineages containing only a single probe or where the top and bottom probe scores differed by less than 0.1 were omitted. Secondly, only the top quartile of the top scoring probes in each lineage and the bottom quartile of the bottom scoring probes in each lineage were used to find motifs, the background model was derived from the G1 probes. In both cases, motifs length 4, 8 and 14 were searched for, using 100 separate runs of the software in each case. Similar motifs were merged and the best matches to all motifs identified in all sequences. Location and motif score for the best matches to each motif in each sequence were used as explanatory variables. The 5000 top-scoring examples of the motif were used to create the logo plot in Figure 7. The assumed prior probability of finding a motif in each sequence was raised until a hit to all motifs was found in all sequences. In most, but not all, cases there was a clear demarcation in the distribution of motif scores between motif scores that were $\sim$0 and all others. In cases of a motif score of $\sim$0, no location value was used.

DNA structure was predicted using hybrid-ss-min (UNAFold) version 3.4 (26), using the DNA input mode with settings of 37°C and 0.58 M NaCl concentration (equivalent to the binding conditions used in the *in vitro* assays). Discrete descriptions of the minimum free energy structure comprising the number of loops, bulges, hairpins and unmatched regions were made using a simple parsing algorithm developed in-house.

The proportion of pairs of nucleotides not able to undergo base stacking was also included (i.e. pyramidine dinucleotides, T-T, T-C and C-C) to indicate an expected bendability (27).

Modelling

Data preparation comprised removing one complete record that had a missing binding score followed by replacing missing values by the mode (for nominal variables) and the median for numerical variables. Regression of the binding score as a function of all 302 predictor was carried out using Breiman and Cutler's random forests™ for classification and regression (28). A forest of 200 trees was used with the single parameter *mtry* set to 100 (the integer value of the number of predictor variables divided by three), its default recommended value for regression.

When using the forest to make predictions on test data, missing values in the predictor variables were first replaced using the same method as for the training data.

Alternative models were also fitted using regression trees (29) and genetic programming (30). Results were very similar in each case (i.e. similar explanatory variables shown to be important), however random forests were chosen since they provided the highest correlation between observed and predicted scores when tested on subdivisions of the main data set (NB the choice of method was made entirely independently of the independent test set reported in Figure 7).

Software

Image analysis used Combimatrix Microarray Imager (https://webapps.combimatrix.com/customarray/customarrayHome.jsp). Data preparation and inspection used JMP 7.02 (31). Data normalization and analysis used R 2.6.0 (32), and the mgcv 1.3 (22) and VSN 3.3.1 (23) packages, the latter being part of the Bioconductor project (33). The randomForest (28) package v4.5 for R was used for modelling.

RESULTS

On-chip aptamer evolution reveals a multimodal fitness landscape

The CLADE procedure (Figure 1) was initiated with a population of 6000 30 nt DNA sequences synthesized in duplicate on a custom microarray. A total of 5500 of these were entirely random sequences (with an equal probability of each base at each position), while the remaining 500 were controls, repeated in each generation and chosen from a trial binding to have a range of binding to APC. A binding score to APC was assayed *in vitro* for each sequence. A new generation of 5500 sequences was then obtained by selection on the basis of the *in vitro* binding score followed by mutation *in silico* (see Materials and methods section). These new sequences were then synthesized on a microarray and the analytical binding process repeated.

Nine generations (G1–G9) of CLADE were completed, during which the population of aptamers evolved from low-scoring binding to consistent high-scoring binding (Figure 2 and Supplementary Table 2). To relate the relative binding-score scale used to an absolute affinity scale, sequences from the top and bottom of the binding score scale were assayed in an independent, Surface Plasmon Resonance (SPR) system. This indicated that, even using a different assay method, the high-scoring sequence bound APC with high affinity, whereas the low-scoring sequence did not measurably bind APC (though it was still possible to estimate the rate of wash-off, Supplementary Table 2). While not expecting a linear relationship between concentration and SPR signal, there are complexities in the binding, giving a reduction of the SPR signal at higher concentrations (Figure 3). However, an assumption of simple 1:1 binding indicates that the top of the

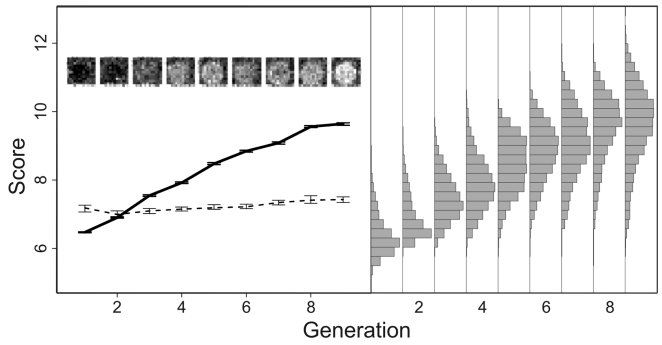

Figure 2. Evolution of binding scores. The left side shows the median score and 95% bootstrap confidence interval for 30-mer experimental probes in each generation (solid line) and the same for the 500 control sequences (dashed line). The right side shows the frequency distribution of all the binding scores in each generation. The spread of the score distributions generally increases (SD: G1 = 0.55, G2 = 0.6, G3 = 0.62, G4 = 0.73, G5 = 0.74, G6 = 0.78, G7 = 0.92, G8 = 0.92, G9 = 1.08). Scores are on an arbitrary scale based on a generalized log transform (50) of the raw intensity measurements. The raw scan of the median scoring sequence from each generation assayed side by side on a separate chip is shown on the top of the left panel.

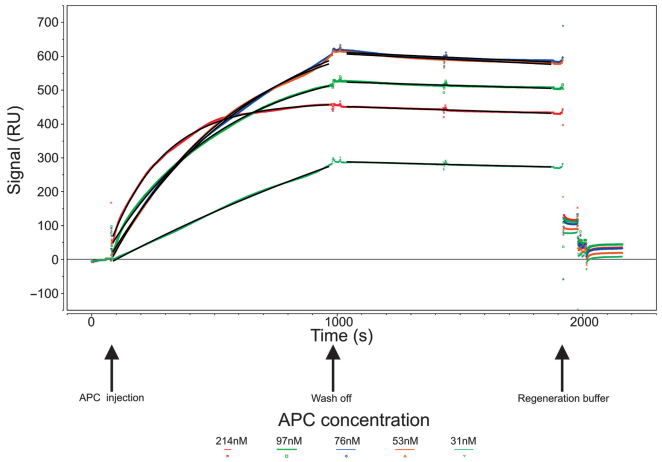

Figure 3. SPR data and fits for a high scoring DNA sequence (G9.3415, score 12.6 ± 0.8). SPR signal is expressed in resonance units (RU), arrows indicate event timings. All data have the values from a control, blank flow cell subtracted. The fitted curves give the results in Table 1. Following the injection of the regeneration buffer, the section at the end (following 2000s) is given to show the reconstitution of the surface to an SPR signal close to the original baseline.

Table 1. Absolute binding via SPR values from the fits shown in Figure 3

APC concentration (nM)	'on rate' k_a ($M^{-1}s^{-1}$)	'off rate' k_d(s^{-1})	Apparent dissociation constant K_D (M)
214	2.1E + 04	5.3E − 05	2.6E − 09
97	2.1E + 04	4.3E − 05	2.0E − 09
76	2.0E + 04	5.5E − 05	2.7E − 09
53	3.6E + 04	6.1E − 05	1.7E − 09
31	1.3E + 04	6.4E − 05	5.0E − 09

Binding constants for APC to a high-scoring sequence (G9.3415, binding-score 12.6 ± 0.8) fitted at five different APC concentrations. The dissociation constant (K_D) assuming 1:1 binding is simply a ratio of the 'on' and 'off' rates. A low-scoring sequence (C10489, binding-score 5.8 ± 0.1) did not give measurable binding.

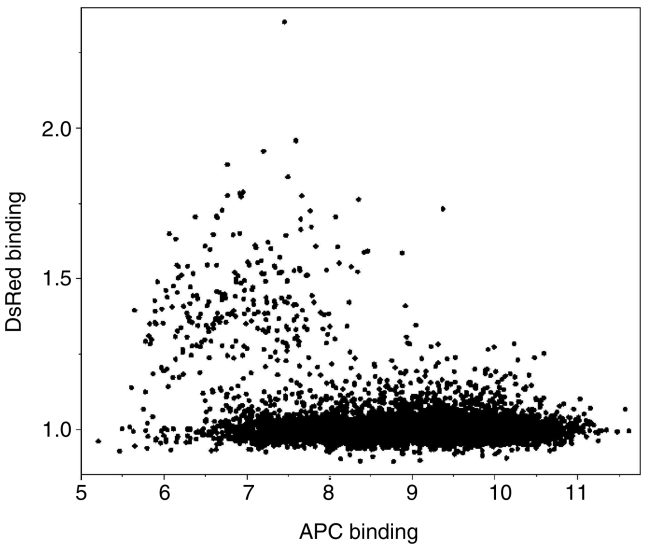

Figure 4. Binding by G7 sequences to APC and DsRed. Both scales are arbitrary, but the APC scale corresponds to that in Figure 2. The majority of sequences exhibiting a high degree of binding to APC possess binding scores around 1 with DsRed, which corresponds to undetectable binding. While some binding to DsRed was identified, the highest affinity APC binders all have low DsRed binding.

binding-score scale corresponds to an apparent disassociation constant (K_D) for the high scoring sequence in the low nanomolar range (Table 1 and Figure 3), well within the range of aptamers developed by other means (34). Direct comparison of aptamers derived by the CLADE approach to those derived to the same ligand by other means is the subject of a separate study (M. Platt. *et al*., Submitted for publication). The specificity of binding was tested by assaying binding on the microarray to DsRed, a structurally unrelated protein. While some binding was identified, the highest affinity APC binders all had low DsRed binding (Figure 4).

Figure 5A illustrates the way the population is made up of individual lineages and the way they vary in size through the course of the evolution [see also Supplementary Figure 8 for the complex dynamics of lineage sizes, similar to that seen in *in vitro* studies (15,35)]. How those lineages explore the sequence landscape is shown in Figure 5B and C, also Supplementary Figures 9 and 10. Sequences do not, in general, fall very close to the 'average' surface of the landscape (Figure 5C), however, much of the discrepancy comes from the representation of a highly multidimensional landscape in two or three dimensions (see Supplementary 'Methods' and Supplementary Figure 11). It is nonetheless clear that the sequence-fitness landscape is complex, with at least two major fitness peaks in APC binding (left and right in Figure 5B and C). This multimodality and a more general

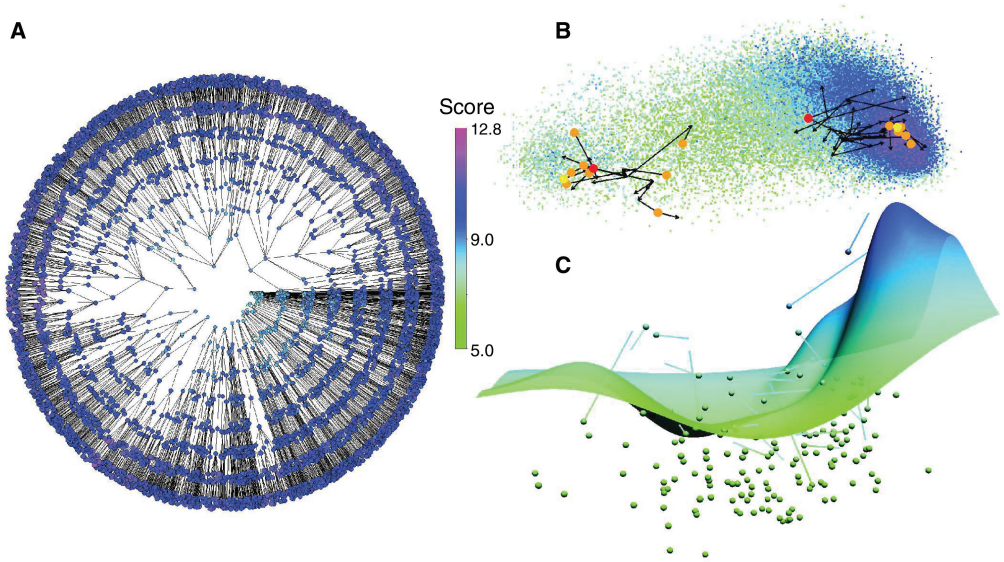

Figure 5. Evolution and the fitness landscape. Scores in the *in vitro* assay are indicated by the same colour scale in all three panels. (**A**) Successful lineages. All aptamers with descendants in the final generation are shown. Each circle represents a single aptamer sequence; lines connect parent and offspring sequences. Node size as well as colour indicates each sequence's binding score. (**B**) All 44 131 sequences are shown. 2D location represents relationship of sequence as determined by multidimensional scaling (see Supplementary 'Methods'), plotted in order from the lowest to highest scorers, so higher scoring points can obscure lower. Two example lineages are indicated. In each case, the red dot is the starting sequence, black arrows are mutations, orange dots are sequences in the penultimate generation of the lineage and the yellow dot is the highest scoring sequence in the lineage. (**C**) A smoothed fitness surface across all sequences on the same *x*- and *y*-axes as (B) but with score as the *z*-axis. Superimposed are 150 randomly picked lineages, with a ball indicating the starting sequence in G1 [equivalent to a red dot in (B)] and a line connecting it to the average position for all the probes in the penultimate generation of the lineage [equivalent to an average of the orange dots in (B)]. Balls without lines indicate sequences that did not have descendants in G3 or later.

'ruggedness' of the landscape is demonstrated explicitly by considering the relationship of sequences' distance from a peak in sequence space with how far below that peak their binding-score is (Figure 6). Further information about the degree and nature of this ruggedness is revealed by considering the fitness differences in individual evolutionary steps (Supplementary Figure 12). The average fitness change decreases with each subsequent generation, and within any individual lineage, the frequency distribution of beneficial mutations is typically roughly exponential, as might be expected from theory (36). Individual lineages vary greatly in the region and amount of the sequence space they explore (Supplementary Figure 10), though for lineages that move away from their starting sequence, the average movement is generally towards the nearest main peak (Figure 5C).

Machine learning model captures the aptamer landscape

A more stringent test of our understanding of the landscape is binding prediction. A Random Forest (28) is a machine learning model capable of non-linear regression without over-fitting, and so is suitable for characterizing multimodal landscapes such as that of Figure 5. We summarized the sequences tested in terms of 302 explanatory variables (including the base at each position, the frequencies and average positions of single, doublets and triplets of base pairs, longer motifs and structural predictions, Supplementary Table 3). These variables and the observed binding scores for all the aptamers were used to train a Random Forest to predict binding.

To test the effectiveness of the resulting model in describing the landscape, we created an entirely independent set of test sequences. First, the random forest was used to predict scores for 10^6 sequences, each base independently selected uniformly at random. 5500 of these sequences were then picked across the range of predicted bindings. This set of sequences was synthesized and assayed *in vitro*. The test set included sequences with both high and low aptameric binding (Figure 7A, Supplementary Table 4). The scores predicted by the random forest explained 70% of the observed variance in binding to APC in this wholly separate test set (Figure 7A and Supplementary Table 4). This is a highly effective prediction for sequences that are entirely unrelated to those on which the model was trained. One could expect even better prediction for variations based more closely on the evolved sequences and indeed, the model, having a mean-square error of 0.115 (the 'out-of-bag error') explains 92% of the variance in the evolved sequences on which it was developed.

The importance of the different explanatory variables in the prediction is shown in Figure 7B and Supplementary Figure 13. It is very possible that features of the sequence beyond the 302 used here could explain more of the variance [other studies have used different DNA features for modelling sequence-function relationships e.g. (3)]. However, the most important features identified here are identified by other modelling approaches (see Materials and methods section) or indeed inspection (Supplementary Table 2). It is notable that none of the metrics derived from secondary structure predictions has a

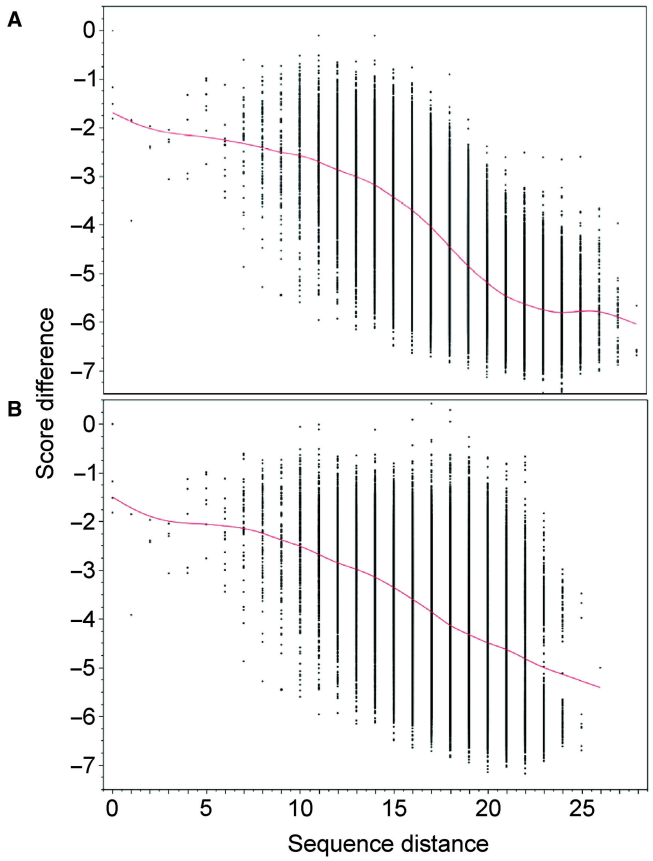

Figure 6. Multimodality and ruggedness in the sequence-fitness landscape. Relationship between sequence distance and score difference for all aptamers (**A**) Relative to the highest scoring sequence in the complete experiment (G9.3415), (**B**) Relative to the closer (in sequence space) of the top scoring sequence in the two major peaks in Figure 5B and C (G9.3415 — ATCCCCCCCTCCCCTTATGTG CACCCGCAT and G8.5150 —ATAGGGGTTGGTTTGGGGGGG GGGATACTC). In each case the line is a cubic spline, stiffness $\lambda = 100$. The line in (B) is straighter than that in (A) indicating that these two peaks account for much of the multimodality in the sequence-fitness landscape.

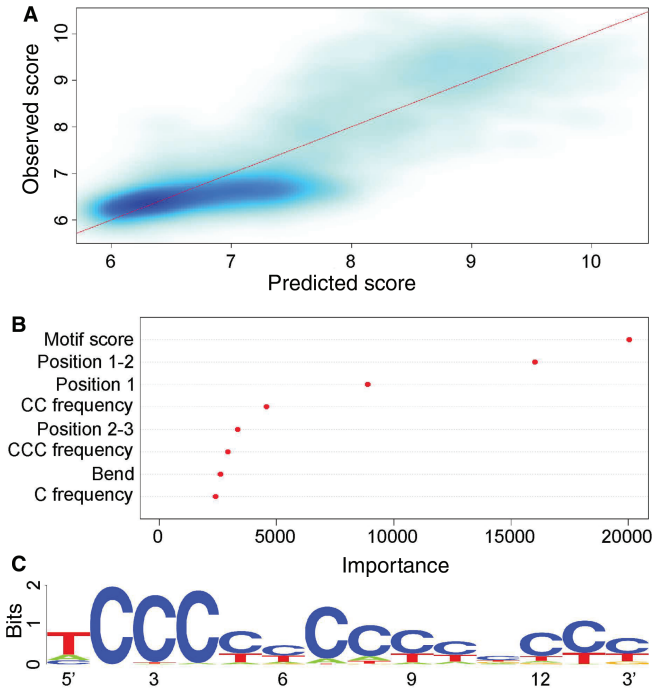

Figure 7. Prediction of aptamer binding. (**A**) Observed versus predicted binding for 5500 test sequences. Colour intensity indicates the density of points. Red line indicates predicted = observed. Pearson correlation = 0.87. (**B**) Importance of the most explanatory variables in the random forest predictor of aptamer binding. Importance is defined as the decrease in node impurity attributable to these variables across the forest. The motif reflects a contiguous sequence found anywhere within the 30mer. (**C**) Information content logo of the motif referred to in (B). The overall height at each position represents the sequence conservation, within which the height of each symbol indicates the relative frequency of each nucleotide.

high importance. It is not clear whether this is due to specific secondary structure truly being unimportant for these aptamers or an inability of the prediction approach used to identify secondary structures correctly in the chip-surface environment.

Prior knowledge can be incorporated in CLADE aptamer design

There is much prior knowledge that may be relevant to aptamer design, e.g. in terms of known structure-function relationships such as nucleic acid motifs that bind specific amino acids (37) or particular structural features with desirable properties (38), or complete existing aptamer sequences. Such knowledge may lead to particularly informative or interesting regions of the sequence-fitness landscape. Of the 500 control aptamers included as parents for the creation of G2, 255 were specifically generated using prior knowledge. 129, (the 'designed' set) were generated via an algorithm which balanced 40 structural and sequence features found in small molecule aptamers (Supplementary Figure 14, Supplementary 'Methods' and C.G. Knight *et al.* unpublished data), while 126 sequences (the 'quadruplex' set) were random sequences apart from the nucleotides required to form a guanine (G) quadruplex (39). A G-quadruplex is a structural feature dependent on appropriately positioned G residues, and is a characteristic of a known DNA aptamer that binds the protein thrombin (11,12).

There was a striking contrast between fates in the evolution of the 'designed' and 'quadruplex' sets of sequences. The 129 'designed' set aptamers in G1 gave rise to 3707 sequences in G9. In contrast, from the 126 'quadruplex' set in G1, only a single aptamer remained in the final generation and that was predicted to have lost its quadruplex structure. The quadruplex aptamers perform poorly even in relation to random sequences—there being 2.3% as many quadruplex as random sequences in G1, but only 0.13% by G9. The poor performance of the quadruplex aptamers is consistent with the random forest model results in which the presence of a quadruplex sequence has a very low importance (299th of 302 explanatory variables) and the variables of high importance are very unlike G-quadruplex sequences (Figure 7B). This result suggests that while important in DNA aptamers

(11,12), the G-quadruplex structural feature may not be of universal utility for protein-binding DNA aptamers and/or that there are additional features required to make a quadruplex-containing aptamer that could not easily be achieved from the starting set of sequences we used. The good performance in the evolution of the designed sequence set relative to the random and quadruplex sets suggests that there may be aspects of aptamer design that are common across very different substrates. The success of designed sequences is partly attributable to many of them already being close to the left-hand fitness peak (Figure 5B and C) in the first generation.

DISCUSSION

These experiments reveal a fitness landscape that is complex (Figure 5B and C) yet very well captured by our predictive model: the correlation of observed versus predicted fitness for an entirely independent set of test sequences is remarkably high (0.87). The experiments are strictly limited to a single molecule (APC) interacting under very specific conditions with a very constrained form of molecule (DNA 30 mers). Even so, the effectiveness of the model is striking considering that even with nine generations of evolution we covered only $\sim 4 \times 10^4$ sequences, which is a tiny proportion of available sequence space ($\sim 10^{18}$ sequences). Further, protein-aptamer interactions are complex, potentially inducing structural changes within both molecules (12).

The way this surprising predictability is achieved, and hence clues to the nature of the genotype-fitness relationship in this system, is revealed by looking into the workings of the model. This also provides hypotheses for the physical basis of aptameric binding. The most important features that determine aptamer binding, as described by the model, concern C bases in the middle of sequences, such as the match to the motif shown in Figure 7C. Such sequential Cs are not liable to π-stacking in single-stranded DNA which may aid flexibility of the sequences (27), a feature that is considered explicitly by the 'bend' variable that is also important in the random forest model (Figure 7B). However, the evolution is by no means directed towards a C homopolymer. Firstly, the motif shown in Figure 7C is not an approximation to a string of Cs—if a C homopolymer motif of the same length is scored, and the random forest refitted using the homopolymer motif instead of that in Figure 7C, the homopolymer motif has only a quarter of the motif's importance in the model (data not shown). Secondly, the complexity of the landscape means that runs of Cs are neither necessary nor sufficient for high-scoring aptamers. For instance, sequence CACCGGGCCCCCCCCCCCCCCCACCAC GCC (G2.3596) contains a string of 15 Cs (there is only one aptamer in the whole evolution with more), yet scores in the bottom 10% of aptamers overall. Conversely, the sequence ATCACACCAAACCTTTCTTCGATTAA GTCG (G6.5628), with no more than two adjacent Cs, scores in the top 10%. The other most important features in determining binding relate to the three bases at the 5′-end of the sequence (the 'position' variables). Even among these relatively simple variables, the model captures complex effects. For instance, a C base in position 2 can give good binding, but *only* in combination with 'A' in position 1 and 'C' or 'T' in position 3. The CLADE procedure used here thoroughly explores these variables, meaning that, by G9, most aptamers start with 'ATC', the sequence with the highest marginal effect (Supplementary Figure 13). This 5′-motif is physically positioned away from the surface of the chip, potentially representing a particular recognition motif. This dependence of binding on specific sequence motifs rather than merely nucleotide composition or other general features are consistent with the known association of tight aptameric binding with sequences of high information content (40).

The insight into the sequence-fitness landscape (Figure 5) and the ability to predict binding (Figure 7) derive directly from the CLADE approach. However, this is not the only advantage to CLADE when compared with fully *in vitro* techniques such as SELEX and similar approaches (41). Specific and noteworthy features of CLADE include: unlike fully *in vitro* approaches, CLADE is not biased towards sequences favourable for PCR; there is no immobilization of the target ligand, which risks selecting aptamers which bind the immobilization matrix; different affinity ligands are produced from the same experiment [this is possible but uncommon in SELEX (42)] providing affinities suited to a chosen target's concentration range; there is complete control of population sizes, selection and mutation techniques, which allows one to deal with (and indeed study) issues of founder effects, elitism and other pitfalls of evolutionary dynamics. The last of these offers scope to use sequence-changing operators that are not accessible to fully *in vitro* techniques, e.g. the use of arbitrary knowledge-based transformations, such as the introduction of particular motifs or palindromic sequences. Indeed external knowledge may be incorporated at any stage of CLADE. We have demonstrated the efficacy of incorporating diverse prior knowledge in the design of the initial population. Because CLADE does not use enzymes means that it is not confined to standard nucleic acid bases, nor to bases that are substrates for nucleic acid polymerization. Aptamers could equally easily be designed to include locked nucleic acids, peptide nucleic acids or non-biological bases, or indeed any other monomers, which may also be less subject to biological degradation if taken into an *in vivo* system. In this study, we have used a single binding assay. However, CLADE could equally be applied to the evolution of other traits such as enzymatic activities in ribozymes or for DNA computing (43). Different forms of multiplex assay, e.g. by mass spectrometry, may in the future allow the evolution of aptameric binding to different targets simultaneously, an approach that will be aided by increasing spot densities on custom microarrays and improvements in mass spectrometric imaging (44). None of these advantages come at the cost of speed since CLADE is quick: the automated sequence generation meant that we were able to perform one generation of CLADE, including the design, synthesis and analysis of microarrays, in 24 h.

The understanding we have gained of an aptameric landscape is a form of 'Quantitative Structure (or Sequence)-Activity Relationship' (QSAR, a term used in chemometrics). The CLADE approach could be extended in the way that other QSAR approaches have been, to use the connection between sequence and fitness, captured in our case by a random forest, in the process of the aptamer evolution itself. The use of models to accelerate the search process is well known in evolutionary computation (45) and has been used successfully in developing desirable enzymatic activities (46). Conversely, the model could be used to minimize binding when generating aptamers to a different target without recourse to *in vitro* counter-selection. In this regard, it is very important that the CLADE method 'sees' the low-binding sequences—compare the uniformity in colour (binding score) of Figure 5A, which shows only sequences that make it through to the final generation (as might be assayed in an *in vitro* technique), with the broad colour/score range of Figure 5B and C which include all the sequences assayed by CLADE. The model we developed is therefore particularly good at predicting, and hence generating, low-affinity binding sequences, as well as 'good' aptamers (Figure 7A). Such minimization of cross-reactivity is a crucial factor if aptamer arrays for proteomics and metabolomics and their advantages of principle over much better established antibody technologies (20,21) are to be realized. More generally, a number of targets for positive or negative selection can be handled simultaneously using *multiobjective* evolutionary algorithms (47,48).

Mapping sequence-function landscapes for particular biomolecules as we have done here is of wide relevance, notably in the emerging discipline of synthetic biology (49). However, mutational landscapes and landscape concepts of sequence-fitness relationships extend much further among multi-molecule systems. For instance genomic landscapes of cancer (50) and landscapes of whole organism inclusive fitness (51), even 'landscapes of change' (52). For whole organisms, the landscape metaphor has conceptual limitations (53), is rarely adequately tested (54) and seldom extends back to the scale of the effects of specific genome changes (55). However, progress is being made in the theory of sequence-fitness relationships in whole organisms (56). At the same time there is an unprecedented ability to work with mutations generated at unknown locations in complete genomes via high throughput sequencing (57) and array based technologies (58). Thus it may soon be possible to apply similar approaches to those developed here to understanding the vast, multimodal sequence-fitness landscapes experienced by the complete genomes of evolving biological organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at *NAR* Online.

ACKNOWLEDGEMENTS

We would like to thank Andy Hayes, Sanjay Nilapwar, Steve Oliver, Mark Santa Ana, John Cooper, Axel Stover, Dominic Suciu and Karl Maurer for discussions and assistance in the microarray work and Julia Handl for analytical assistance.

FUNDING

Biotechnology and Biological Sciences Research Council; David Phillips fellowship from the BBSRC (to J.K.); Wellcome Trust (to C.K.). This is a contribution from the Manchester Centre for Integrative Systems Biology (www.mcisb.org) which is supported by a BBSRC/EPSRC grant. Funding for open access charge: BBSRC.

Conflict of interest statement. None declared.

REFERENCES

1. Waddington,C.H. (1957) *The Strategy of the Genes; A Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin, London.
2. Poelwijk,F.J., Kiviet,D.J., Weinreich,D.M. and Tans,S.J. (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, **445**, 383–386.
3. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
4. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. III and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
5. Kaufmann,S. and Levin,S. (1987) Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, **128**, 11–45.
6. Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. London,B*, **255**, 279–284.
7. Cowperthwaite,M.C. and Meyers,L.A. (2007) How mutational networks shape evolution: Lessons from RNA models. *Annu. Rev. Ecol. Evol. Syst.*, **38**, 203–230.
8. Fogel,D.B. (2000) What is evolutionary computation? *IEEE Spectrum*, **37**, 28–32.
9. Ellington,A.D. and Szostak,J.W. (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
10. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
11. Macaya,R.F., Schultze,P., Smith,F.W., Roe,J.A. and Feigon,J. (1993) Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proc. Natl Acad. Sci. USA*, **90**, 3745–3749.
12. Padmanabhan,K., Padmanabhan,K.P., Ferrara,J.D., Sadler,J.E. and Tulinsky,A. (1993) The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. *J. Biol. Chem.*, **268**, 17651–17654.
13. Beaudry,A.A. and Joyce,G.F. (1992) Directed evolution of an RNA enzyme. *Science*, **257**, 635–641.
14. Davis,J.H. and Szostak,J.W. (2002) Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc. Natl Acad. Sci. USA*, **99**, 11616–11621.
15. Bartel,D.P. and Szostak,J.W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411–1418.
16. Asai,R., Nishimura,S.I., Aita,T. and Takahashi,K. (2004) *In vitro* selection of DNA aptamers on chips using a method for generating point mutations. *Anal. Lett.*, **37**, 645–656.
17. Ikebukuro,K., Okumura,Y., Sumikura,K. and Karube,I. (2005) A novel method of screening thrombin-inhibiting DNA aptamers using an evolution-mimicking algorithm. *Nucleic Acids Res.*, **33**, e108.
18. Stanlis,K.K. and McIntosh,J.R. (2003) Single-strand DNA aptamers as probes for protein localization in cells. *J. Histochem. Cytochem.*, **51**, 797–808.

19. Stadtherr,K., Wolf,H. and Lindner,P. (2005) An aptamer-based protein biochip. *Anal. Chem.*, **77**, 3437–3443.
20. Tomizaki,K.Y., Usui,K. and Mihara,H. (2005) Protein-detecting microarrays: current accomplishments and requirements. *ChemBioChem.*, **6**, 782–799.
21. Famulok,M., Hartig,J.S. and Mayer,G. (2007) Functional aptamers and aptazymes in biotechnology, diagnostics, and therapy. *Chem. Rev.*, **107**, 3715–3743.
22. Wood,S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.*, **99**, 673–686.
23. Huber,W., von Heydebreck,A., Sultmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18(Suppl 1)**, S96–S104.
24. Wedge,D., Rowe,W., Kell,D. and Knowles,J. (2008) *In silico* Modelling of Directed Evolution: Implications for Experimental Design and Stepwise Evolution. *J. Theor. Biol.*, (in press).
25. Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.
26. Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.
27. Sain,A., Chen,J.Z.Y. and Ha,B.Y. (2006) Persistency of single-stranded DNA: The interplay between base sequences and base stacking. *Physica. A*, **369**, 679–687.
28. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
29. Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth International, Monterey.
30. Koza,J.R. (1992) *Genetic Programming: On thePprogramming of Computers by Means of Natural Selection*. MIT Press, Cambridge, Mass.
31. SAS Institute Inc. (1989–2007). Cary, NC.
32. R development core team. (2008). R foundation for statistical computing, Vienna.
33. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
34. Silverman,S.K. (2008) Artificial functional nucleic acids: Aptamers, ribozymes and deoxyribozymes identified by *in vitro* selection. In Lu,Y. and Li,Y. (eds), *Functional Nucleic Acids for Sensing and Other Analytical Applications*. Springer, New York.
35. Fitter,S. and James,R. (2005) Deconvolution of a complex target using DNA aptamers. *J. Biol. Chem.*, **280**, 34193–34201.
36. Orr,H.A. (2006) The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J. Theor. Biol.*, **238**, 279–285.
37. Hoffman,M.M., Khrapov,M.A., Cox,J.C., Yao,J., Tong,L. and Ellington,A.D. (2004) AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res.*, **32**, D174–D181.
38. Lescoute,A. and Westhof,E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.
39. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
40. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–137.
41. Paegel,B.M. and Joyce,G.F. (2008) Darwinian evolution on a chip. *PLoS Biol.*, **6**, e85.
42. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
43. Stojanovic,M.N. and Stefanovic,D. (2003) A deoxyribozyme-based molecular automaton. *Nat. Biotechnol.*, **21**, 1069–1074.
44. Northen,T.R., Yanes,O., Northen,M.T., Marrinucci,D., Uritboonthai,W., Apon,J., Golledge,S.L., Nordstrom,A. and Siuzdak,G. (2007) Clathrate nanostructures for mass spectrometry. *Nature*, **449**, 1033–1036.
45. Jin,Y., Olhofer,M. and Sendhoff,B. (2002) A framework for evolutionary optimization with approximate fitness functions. *IEEE Trans. Evol. Comput.*, **6**, 481–494.
46. Fox,R.J., Davis,S.C., Mundorff,E.C., Newman,L.M., Gavrilovic,V., Ma,S.K., Chung,L.M., Ching,C., Tam,S., Muley,S. *et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.*, **25**, 338–344.
47. Knowles,J.D. and Corne,D.W. (2000) Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy. *Evol. Comput.*, **8**, 149–172.
48. Zitzler,E. (1999) *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. Shaker, Aachen.
49. Andrianantoandro,E., Basu,S., Karig,D.K. and Weiss,R. (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, 2:2006.0028.
50. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjoblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
51. Wright,S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Internat. Cong.Genetics*, **1**, 356–366.
52. Kwinter,S. and Boccioni,U. (1992) Landscapes of change: Boccioni's "Stati d'animo" as a general theory of models. *Assemblage*, **19**, 50–65.
53. Reiss,J.O. (2007) Relative fitness, teleology, and the adaptive landscape. *Evol. Biol.*, **34**, 4–27.
54. Belotte,D., Curien,J.B., Maclean,R.C. and Bell,G. (2003) An experimental test of local adaptation in soil bacteria. *Evolution*, **57**, 27–36.
55. Lenski,R.E., Ofria,C., Pennock,R.T. and Adami,C. (2003) The evolutionary origin of complex features. *Nature*, **423**, 139–144.
56. Orr,H.A. (2006) The population genetics of adaptation on correlated fitness landscapes: the block model. *Evolution*, **60**, 1113–1124.
57. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
58. Gresham,D., Ruderfer,D.M., Pratt,S.C., Schacherer,J., Dunham,M.J., Botstein,D. and Kruglyak,L. (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science*, **311**, 1932–1936.