Recommendations on the implementation of Genetic Algorithms for the Directed Evolution of Enzymes for Industrial Purposes

DOI:
10.1002/cbic.201700013

Document Version
Accepted author manuscript

Citation for published version (APA):

Published in:

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Title: Recommendations on the implementation of Genetic Algorithms for the Directed Evolution of Enzymes for Industrial Purposes.

Authors: Mark Howard Barley, Nicholas Turner, and Roy Goodacre

This manuscript has been accepted after peer review and appears as an Accepted Article online prior to editing, proofing, and formal publication of the final Version of Record (VoR). This work is currently citable by using the Digital Object Identifier (DOI) given below. The VoR will be published online in Early View as soon as possible and may be different to this Accepted Article as a result of editing. Readers should obtain the VoR from the journal website shown below when it is published to ensure accuracy of information. The authors are responsible for the content of this Accepted Article.

To be cited as: ChemBioChem 10.1002/cbic.201700013

Link to VoR: http://dx.doi.org/10.1002/cbic.201700013
Recommendations on the implementation of Genetic Algorithms for the Directed Evolution of Enzymes for Industrial Purposes.

Mark H Barley, Nicolas J Turner and Royston Goodacre

School of Chemistry, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

Abstract
Within Directed Evolution the assessment of candidate enzymes and their modification is essential. In this study we investigate Genetic Algorithms (GAs) within this context and conduct a systematic study of the behavior of GAs on 20 Fitness Landscapes (FLs) of varying complexity, which allowed the tuning of the GA to be explored. From this study recommendations for the best GA settings to use for a GA-directed High Throughput Experimental program (where populations and the number of generations is necessarily low) are reported. The FLs were based upon simple linear models and were characterized by the behavior of the GA on the landscape as demonstrated by stall plots and the footprints and adhesion of candidate solutions which highlighted Local Optima (LO). In order to maximize progress of the GA and reduce the chances of being stuck in a LO it was best to use:- 1) high number of generations, 2) high populations, 3) remove duplicate sequences (clones), 4) double mutation, 5) high selection pressure (two best individuals go to the next generation), 6) and consider using a designed sequence as the starting point of the GA run. We believe that these recommendations may be appropriate starting points for studies employing GAs within Directed Evolution experiments.

1. Introduction
The pharmaceutical industry has a long history in High Throughput Experimentation (HTEx) methods for rapid screening. These were initially developed where robotic rigs were used to test thousands of compounds against biochemically important targets. Subsequently these have become widely used within the specialty chemicals industry\cite{1}, biomedical research\cite{2} and materials discovery\cite{3}.

With the increasing use of HTEx methods within both biology and chemistry the need for an efficient method of exploring large numbers of potential experiments (large search spaces) becomes paramount. Whether the problem is optimizing a catalytic system or modifying an enzyme so that it provides improved specificity under industrial conditions the use of HTEx methods allows the researcher to explore far more variables and do orders-of-magnitude more experiments than is possible for a team of chemists/biologists using conventional laboratory techniques. This means that the numbers of potential experiments can be huge, particularly in molecular biology. For example to find a short length of DNA (an aptamer of 30 nucleotides) with optimal binding to a specific target (a protein) requires a search through $4^{30}$ or approximately $1 \times 10^{18}$ possible strands of DNA.\cite{4} If each of these assessments were to take a mere second to conduct, without parallel experiments, an exhaustive evaluation would not be possible as the lifetime of the Universe is a mere $10^{17}$ seconds.\cite{5}. Therefore an alternative strategy must
be used and in response to the requirement for efficient searching of such large search spaces many researchers have been using evolutionary computationally-based algorithms (including GAs) to direct their HTEx programs.

A Genetic Algorithm (GA) is a well established method of finding a very good solution (usually not the optimum solution) to certain types of problems that require a search through a large set of possible solutions. In a GA a small set of candidate solutions (called a “population”), where each candidate solution is coded onto a “chromosome”, are allowed to “evolve” over many “generations” using principles derived from natural selection and evolution. The main requirements for the use of a GA are that the problem to be solved can be expressed mathematically and that all possible candidate solutions can be assessed, and hence ranked, using a “fitness” function or something similar. The algorithm can then evolve improved solutions over many generations to approach a very good solution to the original problem indicated above.

GAs are particularly useful for finding an optimum solution to a complex function or model where the model can be run many times with different parameter settings at minimal cost. This is because GAs are very data hungry; they require significant “populations” to evolve over many “generations” requiring a lot of model runs to reach the optimum solution. Despite this GAs have been used to improve a process (though rarely to get the optimum solution) where the acquisition of the fitness value for the GA requires real physical experiments to be run (for a summary of the literature see Supporting Information Section S1).

Directed Evolution (DE) is a well documented strategy for engineering proteins. While there are many methods of applying DE to an enzyme the research group of Reetz has had considerable success with sequentially applying saturated mutagenesis at selected sites (referred to as Iterative Saturated Mutagenesis; ISM) to improve the activity or stereoselectivity of an enzyme over a number of generations. However, the order in which sites are selected for ISM is potentially important; not all the potential pathways will give improved enzymes and so it is possible to end up in dead ends of the search space. A number of methods have been used to facilitate Directed Evolution in the hope of improving progress and avoiding dead ends. Some methods work by identifying and selecting key mutations that result in improved enzymes. These include the protein sequence activity relationship algorithm (ProSAR) which uses a statistical model to identify both positive and negative mutations and then creates libraries with enhanced numbers of good mutations and fewer poor mutations. The adaptive substituent reordering algorithm (ASRA) has a similar goal but unlike ProSAR does not make any assumptions about the nature of the fitness landscape to be explored (it is model free). Alternative methods use evolutionary based algorithms. For example a GA has been used to improve the experimental evolution of trypsin inhibiting hexapeptides; and an algorithm, similar to a GA, has facilitated the experimental evolution of peptide ligands (containing 16 amino acids) to improve binding to a model target glycostructure (ganglioside GM1).

As discussed above GAs generally rely on large populations and run for many generations. Thus a significant experimental resource is required to
implement a GA directed HTEx program. It is therefore important to establish the optimal selection of the key “tuning” parameters used within a GA. These include: the number of generations, population size, selection pressure employed, as well as the type of crossover and mutation rate. Despite the fact that the danger that a sub-optimal GA design could pose to an experimental program has been recognized,[17] little work to date has been reported on optimizing GA settings for a GA directed HTEx program. By their very nature GAs are stochastic so that multiple convergences have to be run before any statistically significant conclusions can be made about the resulting solution. Moreover, to achieve appropriate convergence a GA requires a fitness function or fitness model, which is needed to assess the population as it evolves. In order to assess enzyme amino acid sequences with enzyme rates and/or enantiomeric selectivity, biochemical tests have to be made and whilst laborious these are appropriate fitness functions. However, it is usually impossible (from arguments made above) to assess the whole search space and for DE applications no such complete datasets currently exist. Therefore to establish the optimum GA tuning parameters an approximation of the fitness function is required, derived from a subset of the total search space; in this study using data from Reetz and colleagues[12].

One of the few systematic studies on the tuning of a GA for use with a HTEx program (where the numbers of samples and generations are severely constrained) is the work by Clerc, Pereira and colleagues[17,18] using a benchmark model for the activity of metal catalysts based upon experimental results. When this function was used as a test bed to optimize key GA parameters the authors found that the best results were obtained with large populations, selection of parents by tournament selection and using elitism (always retaining the best individual from the previous generation- stops the GA going backwards). They also tested four methods of gene crossover but found no statistically significant differences.

The objective of this study is thus to provide some rules or guidelines for the best parameters to use for a GA guided HTEx. However it is anticipated that the guidelines may be useful for any GA directed HTEx program where the number of samples and/or the number of generations are severely constrained. The first model used is based upon experimental directed evolution data for epoxide hydroxylase[12]. From this model 19 further models of varying complexity are derived in order to test the generality of the GA tuning parameters. These optimum parameters for the GA were found by a series of designed experiments which investigated both the best values for the key GA parameters and the interactions between them based upon the fitness values provided by the full set of 20 models. A schematic explaining the workflow described in this study is shown in Figure S1 (Supporting Information).

2. Results and Discussion
2. 1. Deriving Test Functions for the Tuning of a GA
Only a brief summary of the derivation of the 20 models (Fitness Landscapes) used to test the GA will be described here. Full details are available in Section S2.
2.1.1. The Experimental Data
Experimental enantiomeric excess values (E-values) for 141 mutants at eight sites in the sequence of the enzyme epoxide hydroxylase have been published by Reetz’s group[12] using NDT degeneracy[19] (using 12 rather than 20 amino acids). Principal Component Analysis (PCA) of the sequences of these 141 mutants compared to 150 diverse random mutants on these eight sites demonstrates that the Reetz set only covers a tiny fraction of the overall search space; Section S2.1 and Figure S2

2.1.2 Multi-Linear Regression on the Reetz data
The raw E-values were divided through by 4.6 (the activity of the wild type) to give relative E-values (or activity) normalized with respect to the WT. For each amino acid position a change in amino acid at that position is quantified using one of the Principal Property scales of Hellberg[20]; with the value for the WT amino acid being set at zero for each scale. Multiple Linear Regression (MLR) was then used to build a model for the activity of the 141 sequences. The resulting model gave a $R^2$ of 0.799 with 7 significant linear terms, three interactions and one squared term; see Section S2.2 and Figure S3.

2.1.3. Modifying the MLR Model to derive Fitness Landscapes to test the GA.
The MLR model had a coefficient of zero for AA350 and hence the optimal sequence was not unique. To avoid this problem a coefficient of -0.721 (rounded up from the coefficients of one the interactions) was assigned to this 8th amino acid position. This gave a model with a unique global optimum (sequence = GCGCFCDF, relative activity = 127.2787) which could be “doubled up” to give a model for mutation at 16 amino acid positions with 6 interactions; optimal sequence = GCGCFCDFGCGCFCDF, relative activity = 253.5575; Table S1. The coefficients of this “Base Fitness Landscape” were then randomly reorganized to give 19 further models with 0, 2, 4 or 6 interactions; Section S2.3 and Table S2. In this work we will refer to these models that provide a fitness value for any 16 amino acid sequence as “Fitness Landscapes” (see next section) to distinguish them from other models (e.g. Experimental designs) that may be referred to in the text. These fitness landscapes should not be confused with “fitness-pathway landscapes” discussed in Gumulya et al.[12] where the pathways from the WT to improved mutants are represented by a decrease in experimental free energy values ($ΔΔG^*$), and minima in these landscapes are associated with dead ends or sequences that do not provide improved mutants.

2.1.4. The Complexity of the Fitness Landscapes
The idea of a fitness landscape was originally suggested by Wright[21] and has been more recently developed into a tunable fitness landscape called an NK model.[22] In this model $N$ is the number of “genes” (in the context of this work it would be the number of amino acid positions mutated) and hence defines the search space; and $K$ is a measure of the complexity of the landscape. For $K=0$ the fitness landscape looks like a single isolated peak (Mt. Fuji) and the higher the $K$ value the more convoluted the fitness landscape becomes with more peaks (local optima) and troughs. Though a wide range of values for $K$ have been suggested for biological systems (see Section S2.4 for some examples) there is some evidence that for the engineering of enzymes, providing the selected amino acid sites are well separated, additive ($K=0$) or...
near additive fitness landscapes (K is very low) work well\textsuperscript{[23]}. The estimated K value for each of the fitness landscapes is included in Table S2.

2.2. Characterization of the Models (Fitness Landscapes- FL)

2.2.1. The “Standard GA parameters”
The “Standard GA parameters” are the initial settings for the GA which we aim to improve in this work. While the selected values are to some extent arbitrary they are based upon previous work reported in the literature\textsuperscript{[17,16]}. Figure S4 is a schematic showing the design of a typical GA. In this work the number of generations and the population will be defined for each run but the Standard GA parameters are the default settings that should be assumed unless otherwise specified. A more complete description of the GA is given in Section S3.1 and only a brief summary will be given here.

For this work the “standard” initial population is the complete set of single point mutants of the wild type sequence, from which the fittest 3 individuals go forward to the start of the first generation. Tournament selection is used to select parents and children are formed by single point crossover (the same random location on both chromosomes) with 100% crossover rate. The number of children is set to be the same as the population number and clones (identical sequences) are allowed. Children are mutated (Section S3.1.2) once (Mutation Level- ML=0.5, if no decloning; ML=1 if mutation is combined with decloning) and then assessed for fitness with the best sequences (the Transfer Population or TP) selected to become parents for the next generation. For the Standard GA parameters TP=10 which corresponds to low selection pressure; Section S3.1.3. Elitism is used to stop the GA going backwards.

2.2.2. Characterization of the Fitness Landscapes by using Stall Profiles

The fitness landscapes were characterized by identifying sequences that hold-up or “stall” the progress of the GA. The GA is run 1500 times for 100 generations with a small population (Pop=15) to identify where the fitness of the best sequence does not change between generations. This is where the GA progress is stalled for one or more generations and the number of stalls (out of the 1500x100=150,000 total) can be plotted against the fitness values to give a “stall profile” of the progress of the GA across the landscape. The stall plot for FL-12 using Standard GA Parameters is shown in Figure 1A and the complete set of stall plots for all the fitness landscapes is shown in the Supporting Information (Section S5). In general, any sequence with stalls >1000 is probably stuck in a Local Optimum (LO) or has reached the Global Optimum (GO). Figure 1A shows that for FL-12 very few GA runs (using the standard parameters) get through to the GO at 172.1115. Most of the runs are caught in the LOs with fitness values at 124.9041 and 150.3016. From the full set of stall profiles it can be seen that: (1) for all the fitness landscapes the approach to the GO or any LOs causes a substantial build-up in stalling before the key sequence is reached; (2) some sequences that are not LOs still cause peaks in the stall profile: e.g. Feature at 92.7961 in FL-20 (see Section S5); (3) the complexity of the stall profile only relates very loosely to the number of interactions (and K value); (4) despite the simplicity of the underlying models most of the stall plots (particularly those with more than 2 interactions) are unique.
2.2.3. Identification and Characterization of Local and Global Optima

From the stall plots the GO and LOs for each fitness landscape could be identified and these are summarized in Table 1. In this study we define a LO as a sequence that is not a GO and has no single point mutants (of which there are typically 12x16=176) with a higher fitness value. Many of the LOs were identified from the stall plots. However some of the LOs listed in Table 1 have very small (even zero) stall totals so these were identified by other methods. For example the LO at 126.3001 in FL-12 (LO 34 in Table 1, absent from the stall plot) was identified by studying where a GA starting from the SPMs of LO 36 (104.4903) ended up (see next section). The GA was directed to the sequence with a fitness of 126.3001 by 13 SPMs of LO 36, and when the 126.3001 sequence was tested it was found to be a LO. Although the GA does not find this LO when starting from the best 3 SPMs of the WT (or from the Designed Sequence-see below) if the starting points of the GA are more widely spread it can be a very important LO. Hence starting the GA from a sequence with isoleucine in all 16 positions gives a stall plot with about 26,000 stalls for LO 34 (126.3001) and only 50 stalls for LO 35 (124.9041). These observations also show that Table 1 cannot be considered a comprehensive list of the LOs for these fitness landscapes; there may be others hidden within the landscape that can only be found by running the GA from many random starting sequences.

From the data collected for the stall plots it was possible to find the number of GA runs that found a specific GA sequence and then stalled. This we term the “stall footprint” and dividing the number of stalls for this sequence by the stall footprint gives the average “adhesion” or the number of generations the GA is held at that sequence. Hence for each sequence:- number of stalls=stall footprint x adhesion. The stall footprints (expressed as a percentage) and adhesion for each of the LOs are included in Table 1.

2.2.4:-- How progress towards the GO is limited by LOs and related sequences

To investigate the role of the LOs and related sequences with high stall rates the onward progress of the GAs from the GO and the 5 LOs for FL-12 (see Table 1) were assessed. For each sequence the 176 single point mutants (SPMs) were generated and used as the starting point for a short GA run (10 generations) with a large population (100) to find which SPMs allowed access to the GO. Multiple (200) short GA runs were performed for each SPM to collect statistical data on the final destination of the GA. In Table S4 the final destination of the GA from the 176 SPMs derived from the GO and the LOs for FL-12 are provided along with those derived from a selection of other sequences. These sequences were initially selected based upon their high stall rates but it soon became clear that they were grouped into “families” based upon the probability of the GA ending up at the GO or different LOs. Table S4 shows that each LO which seriously impedes the GA is the fittest member of a series of sequences with increasing stall rates that direct the GA towards the LO. In Figures 2 and S5 the results for the LOs found for FL-12 are summarized by plotting the fitness of the 176 SPMs of the LOs against their rank with the symbol color coded for the final destination of the GA. Hence for the set of sequences that are associated with the LO at 150.3016...
(B0-B7:- see Table S4) the top 108 SPMs all direct the GA towards the LO (B0). Although there is a 5% nominal probability for a GA starting from one of the set B sequences to get to the GO, in practice the probability is almost zero as such a low ranked SPM is very unlikely to be selected as a parent for the next generation when using small populations. This explains why the adhesion values for the LOs within FL-12 (see Table 1) are so high.

In summary if LOs are present in a fitness landscape and they trap the GA they are highly likely to hold up the GA for an unacceptably large number of generations. Alternative strategies such as selecting a less fit mutant to act as a template for a new round of saturated mutagenesis as suggested by Gumulya et al.\textsuperscript{[12]} may work for some of the LOs but for the more obstructive LOs such as sequence B0 discussed above the researcher would have to be very lucky to select one of the mutants that allowed progress past the LO.

One of the key objectives of this work is to find parameters for running the GA that help it avoid LOs (reduce the stall footprint) and/or to escape rapidly from LOs once caught (reduce adhesion).

2.3. Tuning the GA

2.3.1. Some Basic Results for FL-12

In figures S6-S9 the effect of changing some of the GA parameters on the progress of the GA over FL-12 is shown. In Figure S6 a box-whisker plot of the best fitness values (that is the fitness value for the best sequence in the population) of 200 runs over 20 generations with two populations (30 and 90) is shown. Figure S7 shows the same data expressed as a mean with 95% confidence intervals as error bars with a big improvement in clarity. The GA initially makes rapid progress but as it approaches the first LO (124.9041) progress slows and the line flattens. About one third of the runs get past the LO but these are excluded from these plots to improve clarity. Figure S7 clearly shows the effect of using a larger population in speeding up the GA although it does not obviously affect the proportion of runs getting past the LO (68 for Pop=90, 64 for Pop=30). Figure S8 shows the effect of changing the Selection Pressure from $TP=10$ (low) to $TP=2$ (high) to $TP=1$ (very high- see Section S3.1.3). There is little difference between low and high selection pressure on the progress of the GA but using $TP=1$ gave somewhat poorer results. However low selection pressure is clearly advantageous when it comes to avoiding the LO as the number of runs getting past 124.9041 was 60 for $TP=10$, 48 for $TP=2$ and 8 for $TP=1$. Figure S9 shows the effect on the progress of the GA upon introducing decloning and mutation; Section S3.1.2. Decloning is an important way of retaining the diversity of the population as the GA converges as it approaches the GO (or LOs), and if there was an experimental step would avoid pointless repetition. There is a big improvement in GA progress between the standard parameters ($ML=0.5$: no decloning, single mutation) and single mutation ($ML=1$ corresponding to decloning plus single mutation) and a smaller improvement with extra mutation. However the decloning/mutation level has a big effect on the number of runs getting past the LO with 80 for the lowest level ($ML=0.5$), 165 for $ML=1$ and 189 (out of 200 runs) for $ML=2$.

2.3.2. Looking for Trends across all 20 Fitness Landscapes
Given the diversity of the fitness landscapes described above it is important to identify trends in GA convergence, and how these trends vary with different parameter settings. The GA will progress at very different rates on different fitness landscapes. To counteract this effect and to bring the results for each fitness landscape into line the population required for each fitness landscape to reach a specific target was determined. The target was that 50% of the GA runs should reach 85% of the fitness of the GO in 12 generations (see Section S3.2.2 for details). The population used for each generation varied between 13 for FL-20 to 61 for FL-4 and 122 for FL-12; Table S5. Using these populations for each model means that the results obtained as the GA parameters are changed are comparable across all the fitness landscapes and deviations from the mean for each model can be analyzed to find trends (see Section S3.2.3 for a more complete description of how the figures below are made).

**2.3.3. Selection Pressure and Mutation across all Fitness Landscapes.** Figure 3A shows the effect of selection pressure and the target selected for the GA averaged across all 20 fitness landscapes. As noted earlier a lot of effort would be required to get the GA to complete convergence so better targets for a GA directed experimental program would be a high percentage of the fitness of the GO (eg. 75%, 85% or 95% as seen in Figure 3A). The results show a consistent and significant effect of selection pressure at these target levels and throughout the evolution of the GA with TP=2 always giving the best progress. TP=1 is always lower as are TP=3 and TP=10. However the small effect of selection pressure reduces to insignificance when 90+% of the runs are getting to target. Figure 3B shows a similar plot where both the mutation level and selection pressure are changed for a constant target (85% of GO fitness). The behavior of the GA towards selection pressure is exactly as seen in panel A and is the same for the three mutation levels. Increasing mutation from Standard Parameters ($ML=0.5$) to $ML=1$ significantly improves the GA progress; but increasing to $ML=2$ only gives slightly better progress at low generations and becomes a disadvantage close to GA convergence. Figure S10 shows similar plots for a target of 75% and 95% of GO fitness and confirm that $ML=2$ improves GA progress when the GA is far from convergence (as seen in Figure S10 for X=75) but becomes a liability when close to convergence (for X=95, $ML=2$ gives significantly worse results than $ML=1$). Figure 4 shows the effect of higher levels of mutation and the effect of switching to a lower level of mutation three quarters of the way through the GA evolution. Note that the results shown in Figure 4 are based upon 19 FLs. The results for FL-12 were excluded as the behavior of the GA on this FL was very different from the other FLs with increasing levels of convergence even with high constant mutation rates; Figure S11. For the 19 FLs with a constant mutation level it is clear that $ML=1$ (on average) gives the best progress for the GA over the 16 generations, although $ML=2$ is not significantly worse, unlike $ML=3$ and all higher levels of mutation. If the mutation level is reduced after 11 generations there is a significant improvement for $ML=2$ (error bars do not overlap) and much bigger improvements for the higher mutation levels, but even with this change of mutation level there is no advantage in going to higher mutation levels than $ML=2$. 
All these results have used Standard Mutation (for each round of mutation the nominal probability of a sequence being mutated is one). An alternative mutation method uses the nominal probability of a gene being mutated =1/L (L=number of sites mutated; 16 in this case) and gives a distribution of mutations that follows the Poisson distribution (and hence will be referred to in this work as Poisson Mutation). Figure S12 reproduces Figure 4 with the equivalent results for Poisson mutation superimposed. Poisson Mutation gives poor results if decloning is not used (ML=0.5) but for ML=2 gives very similar results to constant Standard Mutation. At higher mutation levels Poisson Mutation gives much better results than constant Standard Mutation but not as good results as switching to a lower mutation level (see Section S3.1.2 for more information on the different methods of mutation).

2.3.4. The effect of Increased Mutation on the Stall Plots for FL-12.
Returning to Figure 1; panel A shows the stall plot for FL-12 using Standard Parameters for the GA with a population of 15, no decloning and one round of mutation (ML=0.5). The upper panels show the results for no decloning, the lower panels have decloning switched on; left hand panels have one round of mutation, right hand panels have 2 rounds of mutation. Figure S13 shows the equivalent results for a population of 90. Using the Standard Parameters with Pop=15 gives very poor results with almost all the runs getting stuck on the LOs at 124.904 and 150.3016 and only 10-12 runs (out of 1500) getting to the GO (Table S6). Increasing the population does not alter this result. Increasing the mutation level (panel B no decloning, 2 rounds of mutation; panel C decloning and one round of mutation-ML=1) greatly reduces the number of runs getting stuck on 124.904, however for small populations few runs are getting through to the GO (61 for ML=1, panel C) with most of the runs getting stuck between 124.904 and the GO. For the large populations (90) the results are much better with more than 95% of the runs (ML=1) getting through to the GO. For ML=2 the results are not so good. Compared to ML=1 the runs getting through to the GO are greatly reduced (from 61 to 2 for Pop=15; from 1442 to 351 for Pop=90) with most of the runs getting stuck between the LO at 150.3016 and the GO. These results show that the form of a stall plot (and hence the progress a GA can make across the fitness landscape) can change drastically with changes to the settings of the GA.

2.3.5. Using a “Designed Sequence(DS)” as the Initial Population(IP).
Up to this point the initial population used to start each GA run has been the best 3 SPMs derived from the starting sequence (WT). An alternative approach is to use the fitness values for the 176 SPMs to identify the best amino acid in each of the 16 positions and build a sequence (Designed Sequence or DS) with these amino acids in the appropriate positions. For some of these fitness landscapes (particularly FL14-20 where K=0.25 or below) the use of DS as a starting point for the GA is so close to the GO (because of the simplicity of the fitness landscapes) that convergence to the GO is trivial. For one model (FL-8) the DS is a LO (Sequence 24 in Table 1) on which the GA gets stuck. For all the other models the use of the DS as the initial population significantly speeds up the progress of the GA as shown in Figure 5. The variation of GA progress with selection pressure and mutation...
level seen in Figure 5 closely follows the trends seen in previous figures and described above. The big difference seen in Figure 5 between the progress of the GA with the two IPs is partly a reflection of the linear nature of the fitness landscapes. Real systems with more complex underlying features partially obscured by random noise are unlikely to give such good results for a DS as IP. However for some systems the use of a DS as the IP could be important—particularly for those HTEX programs exploring a relatively simple fitness landscape (approximates to a linear model) as big improvements could be made within a small number of generations.

2.3.6. Effect of using the “Designed Sequence” as the Initial Population on the stall plots for FL-12.
It has already been noted that starting the GA from different starting sequences can be used to identify new LOs (eg. Sequence 34—see Table 1) and hence it would be expected that the stall plots for FL-12 will be different depending upon whether the initial population is the best 3 SPMs or the DS. Four sets of stall plots can be compared:- Figure 6 (Pop=15, IP=DS) and Figure 1 (Pop=15, IP=3SPMs); against two set of stall plots in the Supporting Information:- Figure S13 (Pop=90, IP=3SPMs) and Figure S14 (Pop=90, IP=DS). Comparing the low population stall plots (Figure 6A vs. Figure 1A) the big differences are that in Figure 6A no runs get past 150.3016 and there are clear peaks at 124.904 and 104.4903. After 100 generations with Standard GA parameters (Figure 6A) only 28 runs get through to 150.3016; 286 runs finish on 104.4903 and the remainder get stuck on 124.904; Table S6. The results shown in Figure 6A are obtained with TP=10. If they are repeated with TP=2 then none of the runs get beyond 124.904. However with decloning and/or higher levels of mutation these results improve (panels B-D in Figure 6). The major difference with Figure 1 is that the peak at 104.4903 is still a major feature in panels B and C, and just visible in panel D. However few of the runs (<5) finish on 104.4903, most finish between the LOs and the GO (as seen in Figure 1 panels B to D) although for ML=1 (Figure 6C) 189 runs finish on 150.3016 and 37 runs finish on 124.904. Clearly changing the starting point of the GA significantly alters the stall plot as the footprints of the different LOs change. This is illustrated over a wider range of fitness landscapes in Table S7 which is similar to Table 1 except that the data for each of the 40 LOs were obtained from stall plots with IP=DS. Those LOs that are substantially changed from Table 1 are indicated by a star.

2.3.7. Can Increased Mutation Levels reduce the Footprint of Key LOs?
Local Optima with stall footprints >20% when using either IP= 3SPMs or IP=DS were selected. If an LO had substantial footprints with both IP then the IP with the largest footprint was used. However for practical reasons, where there was more than one LO for each fitness landscape, all the LOs for that fitness landscape had to use the same IP. The 14 selected LOs are listed in Table S8. To find a footprint value a stall plot had to be obtained; so to get a representative set of footprint values for specific GA settings all the stall plots were done as 5x replicates. The experimental design was very simple with 3 levels of mutation (ML=0.5, 1.0 and 2.0) all with TP=2 and a fourth setting with ML=1 and TP=10. Part of the raw results are seen in Table S8 and (after conversion to footprints expressed as a percentage) are plotted in Figure S15.
These results show that $ML=2$ is effective at reducing the footprints of all 14 LOs to insignificant levels (<5%).

### 2.3.8. If a LO has a low Footprint does it stop blocking the GA?

Reanalysis of the raw data used to derive the footprints of the 14 sequences showed that this was not always the case. The proportion of GA runs successfully getting past the fitness value of the 14 key LOs is seen in Figure 7. Analysis of the distribution of the final fitness values (after 100 generations) found that the behavior of the GA depended very much on the nature of the LO. For some sequences (e.g. Sequence 3) increasing the mutation level dramatically reduced the footprint and massively increased the proportion of GA runs that avoided the LO (see Figure 7). For other sequences (e.g. Sequence 12) increasing the mutation reduced the footprint (Figure S15) but the GA was being stopped on sequences with a fitness value just below that of the LO (probably a sequence which is one of the “family” associated with the LO- see Section 2.2.4) and very few GA runs are successfully getting past the LO (see Figure 7). From Figure 7 it is clear that increasing population and increasing mutation both help the GA to avoid being caught by the LO but the degree of success depends very much on the nature of the LO sequence.

### 2.3.9. Can Increased Mutation levels reduce the Adhesion of Key LOs?

The adhesion of the same set of 14 LOs was investigated by determining how many generations were required before the GA moved on from the specified LO (used as the IP in 1500 GA runs of 100 generations each). Two population sizes (15 and 90) and 4 levels of mutation ($ML=1$ to 4) were investigated and the average adhesion values are plotted in Figure 8 (note scale change on Y axis). A similar set of runs investigated the role of selection pressure on adhesion and found that there was no consistent pattern between $TP=2$ and $TP=10$ (data not shown). There are no GA settings that reduce the Adhesion of all the sequences to levels acceptable for a HTEx program (2-3 generations at the most). Although it is clear that large populations and high levels of mutation help there are many sequences that still show very high levels of Adhesion. It is clear that the GA finds it very difficult to get out of a LO once it has been caught.

### Section 2.4:- Comparison of a GA with other methods of facilitating Directed Evolution (DE).

Using a GA to facilitate DE of an enzyme has the advantage that it systematically explores a substantial proportion of the total search space. It does not assume any underlying form of the fitness landscape it is exploring (a potential disadvantage of ProSAR) and it is scalable to very large search spaces. By contrast, although ASRA gave good results when two amino acid sites were mutated (140 out of 400 possible mutants were assessed so total coverage of the search space was high), it is unclear how this method would work if 16 amino acid sites were being mutated with the resulting reduction in coverage of the search space. The disadvantage of a GA is that specific sequences have to be synthesized and assessed and this is not readily compatible with the present methods (such as ISM) used for DE.
3. Conclusions
It is clear from this work that the behavior of the GA on a Fitness Landscape (FL) is very dependent upon the features within that landscape. In this work the FLs have been characterized by running GAs across the landscape multiple times and finding sequences that impeded the GA (as shown in the stall plots). This allowed Local Optima (LO) to be identified and their properties to be assessed. Despite the simplicity of the FLs used here it is clear that for those FL with $K>0.25$ the behavior of the GA on the landscape is quite unique (see Stall plots in the Supporting Information). This, along with the complex interactions between the different GA parameters and how they change as the GA approaches convergence has made it difficult to find trends and rules that are applicable to all FL. The idea that a GA can be usefully trained on a model that might in some way be related to the experimental system under investigation is not supported by this work. Although general rules have been suggested (see below) for GA settings to be used when maximum progress of the GA is required while the number of generations and population size is seriously constrained there is no guarantee that any specific FL will support these rules. The rules described below should be used with caution. For example, while in general it is best to use two rounds of mutation ($ML=2$) to maximize GA progress and to reduce the chance of getting caught by a LO, if the search space is relatively small and the aim of the experimental program is to find the true GO then $ML=1$ would be a better option; despite the increased chance of getting caught in a LO.

The use of a Designed Sequence (DS) as the starting point for a GA run (rather than the best 3 SPMs) had a very strong effect on the progress of the GA. The use of a DS is equivalent to assuming a simple linear model (no interactions) for the fitness landscape and demonstrates that the use of modeling to suggest improved individuals to be added into the population between generations could potentially greatly improve the rate of convergence of the GA. For many HTEx applications the experimental data may be fitted to some form of response surface[24] that will improve as more data points are measured. However the modeling of the activity of a protein as a function of its sequence to help in the GA directed evolution of an enzyme towards a target is less well developed. This will be the subject of a future paper.

4. Recommendations
In this study a statistical analysis of the behavior of a GA on some simple fitness landscapes ($K<1$) has shown that GAs are not good at consistently avoiding Local Optima (LOs) when using the small populations and small numbers of generations appropriate for a GA directed HTEx program. No GA parameter settings were found that allowed the GA either to avoid or to escape consistently from all the LOs investigated. Despite these caveats the following GA settings would be the initial recommendation for any GA directed HTEx program where it is important to minimize the amount of experimental effort.

1) Determine the number of variables (sites for mutation in the case of Directed Evolution) and hence the size of the search space. Whilst
the researcher will not *a priori* know the Global Optimum (GO), they need to decide what an acceptable optimum should be.

2) A major constraint will be the number of generations the GA can be run. This needs to be as large as possible—preferably equal or larger than the number of variables (sites to be mutated).

3) As large a population as possible should be employed, based on a cost/benefit decision. This will reduce stalling (Figures S12, S13) and minimize the chances of getting caught in a LO (Table S6, Figures 7 and 8).

4) After each generation the two fittest individuals should be selected and allowed to go forward to the next generation (high selection pressure) (Figure 3).

5) Identical sequences should be always removed (decloning) throughout the evolution of the GA (Figures 1, S7).

6) To minimize the chance of being trapped by a LO use decloning plus two rounds of mutation (ML=2) (Figure 7).

7) Consider using a Designed Sequence as an alternative to starting with the initial population based on the best of the Single Point Mutants (SPMs).

5. Computational methods

A series of multi-level full factorial experimental designs\(^{[24]}\) were used to investigate the impact of changes to the GA parameters on the progress of the GA across the 20 FLs. For any set of GA parameters on a specific FL the GA was run 100x to obtain the number of runs reaching a target, this process was then repeated 15 times to obtain an accurate mean and this was repeated for all 20 FLs. The results are presented as mean values with associated 95% confidence intervals (\(= 2T\sigma/\sqrt{N}\) where \(N=\)number of values contributing to the mean, \(\sigma=\)their standard deviation and \(T=\)the value from T-Tables with \(N-1\) degrees of freedom, \(\alpha=0.05\)). Hence running the GA to collect the data used significant computer time (up to several days on a 3.5GHz Windows desktop machine); each symbol in Figure 4 required 28500 GA runs, the full figure required about 1.14 million GA runs. The results were analysed on a 2.6GHz Apple MacBook Pro laptop with all calculations done in Matlab© 8.1 (R2013a). More details on the methods used are provided in section S3 in the Supplementary Information.

**Abbreviations**

- AA Amino acids
- DE Directed Evolution
- DS Designed Sequence
- FL Fitness Landscape
- GA Genetic Algorithm
- GO Global Optimum
- HTEx High Throughput Experimental
- IP Initial Population
- LO Local Optimum
- ML Mutation Level
- Pop Population
SPM  Single Point Mutant  
TP Transfer Population  
WT Wild Type  

Acknowledgements  
This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and Glaxo-SmithKline (GSK) under the Strategic Longer and Larger (sLoLa) grant initiative ref BB/K00199X/1.

Keywords: - genetic algorithms; fitness landscapes; tuning; high throughput experimental; optimization

Table 1: The Local and Global Optima Identified for Each Fitness Landscape

<table>
<thead>
<tr>
<th>Model (Fl)</th>
<th>LO Number</th>
<th>Fitness</th>
<th>Frac(2)</th>
<th>Sequence</th>
<th>Final Value(%) (3)</th>
<th>Stalls</th>
<th>Footprint(%) (4)</th>
<th>Adhesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GO</td>
<td>253.5575</td>
<td>0.910</td>
<td>&lt;1</td>
<td>GCGCDFGGCFCFDFCF &gt;99</td>
<td>282</td>
<td>&lt;1 &lt;1</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>200.6256</td>
<td>0.910</td>
<td>&lt;1</td>
<td>GCGCDFGGCFCFDFCF</td>
<td>189</td>
<td>&lt;1 &lt;1</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td>200.6256</td>
<td>0.910</td>
<td>&lt;1</td>
<td>GCGCDFGGCFCFDFCF</td>
<td>189</td>
<td>&lt;1 &lt;1</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>2 GO</td>
<td>275.7971</td>
<td>0.917</td>
<td>&lt;1</td>
<td>RCGRDGFCGCGCFDFCF</td>
<td>67</td>
<td>&lt;1 &lt;1</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>218.9218</td>
<td>0.897</td>
<td>&lt;1</td>
<td>RCGRDGFCGCGCFDFCF</td>
<td>185</td>
<td>&lt;1 &lt;1</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>218.9218</td>
<td>0.897</td>
<td>&lt;1</td>
<td>RCGRDGFCGCGCFDFCF</td>
<td>185</td>
<td>&lt;1 &lt;1</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>3 GO</td>
<td>147.1037</td>
<td>0.917</td>
<td>&lt;1</td>
<td>GGCFCDFGRRDCDFDF</td>
<td>76</td>
<td>&lt;1 &lt;1</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>143.5924</td>
<td>0.976</td>
<td>&lt;1</td>
<td>GRLCFRDFGRRDCDFDF</td>
<td>13</td>
<td>5138 13 13</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>4 GO</td>
<td>161.4147</td>
<td>0.879</td>
<td>&lt;1</td>
<td>GRCCDCDRCGCGCFDFCF</td>
<td>45</td>
<td>&lt;1 &lt;1</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>4a</td>
<td>149.2764</td>
<td>0.825</td>
<td>&lt;1</td>
<td>GRCCDCDRCGCGCFDFCF</td>
<td>28</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>5 GO</td>
<td>183.812</td>
<td>0.897</td>
<td>&lt;1</td>
<td>GCRCFCDFGRRDCDFDF</td>
<td>24</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>5a</td>
<td>173.0279</td>
<td>0.957</td>
<td>&lt;1</td>
<td>GCRDCDRCGCGCFDFDF</td>
<td>24</td>
<td>1185 24 24</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>6 GO</td>
<td>153.1960</td>
<td>0.897</td>
<td>&lt;1</td>
<td>RRRCDDDDRCGCDRDCDF</td>
<td>12</td>
<td>&lt;1 &lt;1</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>6a</td>
<td>153.1960</td>
<td>0.897</td>
<td>&lt;1</td>
<td>RRRCDDDDRCGCDRDCDF</td>
<td>12</td>
<td>&lt;1 &lt;1</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>7 GO</td>
<td>103.9629</td>
<td>0.77</td>
<td>&lt;1</td>
<td>GCRCDFDGRRCDFDFDF</td>
<td>50</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>7a</td>
<td>97.0540</td>
<td>0.934</td>
<td>&lt;1</td>
<td>GCRCCSFRRDRDCDFDF</td>
<td>39</td>
<td>2124 47 47</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>8 GO</td>
<td>156.9155</td>
<td>0.861</td>
<td>&lt;1</td>
<td>GRRGRDDGGRCCDFDFDF</td>
<td>43</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>8a</td>
<td>157.5483</td>
<td>0.861</td>
<td>&lt;1</td>
<td>GRRGRDDGGRCCDFDFDF</td>
<td>43</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>9 GO</td>
<td>14932</td>
<td>0.897</td>
<td>&lt;1</td>
<td>GRCCDCDRCGCGCFDFCF</td>
<td>50</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>14932</td>
<td>0.897</td>
<td>&lt;1</td>
<td>GRCCDCDRCGCGCFDFCF</td>
<td>50</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>10 GO</td>
<td>169.4052</td>
<td>0.897</td>
<td>&lt;1</td>
<td>GRCCDCDRCGCGCFDFCF</td>
<td>50</td>
<td>&lt;1 &lt;1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>10a</td>
<td>153.9282</td>
<td>0.890</td>
<td>&lt;1</td>
<td>RRGRDRDFGRCCDFDFDF</td>
<td>31</td>
<td>14932 31 31</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>11 GO</td>
<td>177.2177</td>
<td>0.755</td>
<td>&lt;1</td>
<td>GCHRDFDDDRRCRCCDFDF</td>
<td>86</td>
<td>&lt;1 &lt;1</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>11a</td>
<td>177.2177</td>
<td>0.755</td>
<td>&lt;1</td>
<td>GCHRDFDDDRRCRCCDFDF</td>
<td>86</td>
<td>&lt;1 &lt;1</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>12 GO</td>
<td>157.3016</td>
<td>0.873</td>
<td>&lt;1</td>
<td>RRCCFDGDRGRDCDFDFDF</td>
<td>22</td>
<td>10981 21 21</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>12a</td>
<td>146.7139</td>
<td>0.852</td>
<td>&lt;1</td>
<td>RRCCFDGDRGRDCDFDFDF</td>
<td>22</td>
<td>10981 21 21</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>13 GO</td>
<td>149.7765</td>
<td>0.784</td>
<td>&lt;1</td>
<td>RCGRDRDFGRCCDFDFDF</td>
<td>32</td>
<td>1230 2 2</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>13a</td>
<td>132.6814</td>
<td>0.784</td>
<td>&lt;1</td>
<td>RCGRDRDFGRCCDFDFDF</td>
<td>32</td>
<td>1230 2 2</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>14 GO</td>
<td>172.1115</td>
<td>0.957</td>
<td>&lt;1</td>
<td>GCHRDFDDDRRCRCCDFDF</td>
<td>84</td>
<td>&lt;1 &lt;1</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>14a</td>
<td>121.9377</td>
<td>0.957</td>
<td>&lt;1</td>
<td>GCHRDFDDDRRCRCCDFDF</td>
<td>84</td>
<td>&lt;1 &lt;1</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>15 GO</td>
<td>125.6362</td>
<td>0.957</td>
<td>&lt;1</td>
<td>GCHRDFDDDRRCRCCDFDF</td>
<td>100</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>15a</td>
<td>135.5818</td>
<td>0.957</td>
<td>&lt;1</td>
<td>GCRCCDFGRCDRCDRCD</td>
<td>100</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>16 GO</td>
<td>119.3987</td>
<td>0.910</td>
<td>&lt;1</td>
<td>RCGRDFDFGRCCDFDFDF</td>
<td>63</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>16a</td>
<td>145.8370</td>
<td>0.910</td>
<td>&lt;1</td>
<td>RCGRDFDFGRCCDFDFDF</td>
<td>63</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>17 GO</td>
<td>178.1557</td>
<td>0.910</td>
<td>&lt;1</td>
<td>GCRCDFDGGCDFDFDFDF</td>
<td>99</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>17a</td>
<td>178.1557</td>
<td>0.910</td>
<td>&lt;1</td>
<td>GCRCDFDGGCDFDFDFDF</td>
<td>99</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>18 GO</td>
<td>103.6071</td>
<td>0.910</td>
<td>&lt;1</td>
<td>RRRCFDGRDRDRCDRDFDF</td>
<td>100</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>18a</td>
<td>103.6071</td>
<td>0.910</td>
<td>&lt;1</td>
<td>RRRCFDGRDRDRCDRDFDF</td>
<td>100</td>
<td>N/A</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Note 1: Standard Parameters (TP=10, no declination, MI=0.5, IE=3 Best SPMs, Pop=15. GA run 1500 times/100 generations
Note 2: Fraction=LO Fitness divided by GO Fitness
Note 3: Final value refers to the percentage of runs that finish on specified value.
Note 4: Footprint given as the percentage of runs stalling on the specified sequence
Figure 1: Stall profiles for the progress of the GA over FL-12 with population of 15; 1500 runs of 100 generations. IP=Best 3 SPMs. Panel A: Standard GA Parameters (No decloning, one round of mutation, $ML=0.5$). See Supporting Information for similar stall profiles for the full set of 20 fitness landscapes. Panel B: No decloning, two rounds of mutation; Panel C: decloning with one round of mutation ($ML=1$); Panel D: decloning with two rounds of mutation ($ML=2$). See Figure S11 (Supporting Information) for the corresponding figure with a population of 90.
Figure 2: The eventual destination of the GA starting from the SPMs(176) of six sequences (five LOs and the GO; Table S4) for FL-12. The sequence directs the GA towards the nearest LO for the vast majority of SPMs, so those SPMs that take the GA elsewhere in more than 10% of runs are colour coded: open symbols 10-50% of runs go to specified destination; solid symbols 50-100% of runs go to the specified destination. Note that the fitness values for the SPMs derived from the LOs at 104.4903 and 126.3001 were increased by 12 units and that for the LO at 150.3016 was increased by 8 units to improve figure clarity. A similar plot covering the last 77 SPMs for each sequence is shown in Figure S4.
Figure 3: The variation of the degree of convergence of the GA averaged over the full set of 20 fitness landscapes, and how it depends upon target, selection pressure, mutation level and number of generations. Each set of 4 results goes from most selective ($TP=1$, left hand symbol) through intermediate selectivity ($TP=2,3$, middle two symbols) to the lowest selection pressure ($TP=10$, right hand symbol). To bring the results for the models into line the populations used (see Section 2.3.2) change between models but is constant for a given model. Error bars give 95% confidence intervals for the mean (N=20). Panel A: Variation of GA progress with selected target and selection pressure. Panel B: Variation of GA progress with selection pressure and mutation level ($X=85\%$ for all symbols). See section S3.2 for more detail on how this figure was constructed.
Figure 4.- The variation of GA progress over 16 generations with mutation level (ML) and selection pressure averaged over 19 FL (all except FL-12, see text). The contrast between retaining a high level of mutation for the full 16 generations and switching to a lower mutation level after 11 generations is shown. Each set of 4 results goes from most selective ($TP=1$, left hand symbol) through intermediate selectivity ($TP=2,3$, middle two symbols) to the lowest selection pressure ($TP=10$, right hand symbol). To bring the results for the models into line the populations used (see Section 2.3.2) changes between models but is constant for a given model. Error bars give 95% confidence intervals for the mean ($N=19$).
Figure 5: A comparison of using a Designed Sequence as the Initial Population compared to the Standard Initial Population (best 3 SPMs). Results are averaged across 11 Fitness Landscapes (FL). These are all the FL with $K>0.25$ except FL-8 (because using IP=DS takes the GA directly to a LO -Sequence 24-see Table 1-where it sticks); and FL-12 (which shows anomalous behavior as previously noted). The results show the percentage of runs reaching the target after the specified number of generations. For each set of 8 results (2 sets of 4) the first set refers to ML=1, the second to ML=2. Each set of 4 results goes from most selective (TP=1, left hand symbol) through intermediate selectivity (TP=2,3, middle two symbols) to the lowest selection pressure (TP=10, right hand symbol). The error bars are 95% confidence intervals on the means calculated across the 11 FL. See Section 2.3.2 for the populations used for each FL.
Figure 6: Stall profiles for the progress of the GA over FL-12 with population of 15 and starting from a ‘Designed Sequence’ as the initial population (1500 runs of 100 generations each, TP=10). Panel A: Standard GA Parameters (No Decloning, one round of mutation, ML=0.5). Note that if the stall plot in Panel A is repeated with TP=2 no runs make it past the LO at 124.9041. Panel B: No Decloning, two rounds of mutation; Panel C: Decloning with one round of mutation (ML=1); Panel D: Decloning with two rounds of mutation (ML=2). See Figure S12 (Supporting Information) for the corresponding figure with population=90.
Figure 7: The Effect of Mutation Level and Population on the Percentage of GA runs that get past 14 Key Sequences (LOs). Sequences selected because of their high footprints (>20%) when using a GA with Standard Parameters. Red symbols: Population=15; blue symbols: Population=90. The sequence numbers refer to LO numbers in Table 1, and are summarized in Table S8 (see Supporting Information). Note the scale change in the Y axis.
Figure 8: The effect of Mutation level and Population size on the Adhesion of 14 key sequences selected because of their high footprints (>20%) when a GA is using Standard Parameters. The sequence numbers refer to LO numbers in Table 1, and are summarized in Table S8 (see Supporting Information). Note the scale change in the Y axis.