Published online February 18, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 3 ***e32***
doi:10.1093/nar/gni031

Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: identification of novel zinc finger proteins

Marcus D. Hughes[1,3], Zhan-Ren Zhang[1], Andrew J. Sutherland[2], Albert F. Santos[4] and Anna V. Hine[1,3,*]

[1]School of Life and Health Sciences and [2]Chemical Engineering and Applied Chemistry, School of Engineering & Applied Science, Aston University, Aston Triangle, Birmingham B4 7ET, UK, [3]ProtaMAX Ltd, 55 Colmore Row, Birmingham B3 2AS, UK and [4]GE Healthcare, Cardiff Laboratories, Forest Farm, Whitchurch, Cardiff CF14 7YT, UK

Received December 7, 2004; Revised and Accepted January 27, 2005

ABSTRACT

We have successfully linked protein library screening directly with the identification of active proteins, without the need for individual purification, display technologies or physical linkage between the protein and its encoding sequence. By using 'MAX' randomization we have rapidly constructed 60 overlapping gene libraries that encode zinc finger proteins, randomized variously at the three principal DNA-contacting residues. Expression and screening of the libraries against five possible target DNA sequences generated data points covering a potential 40 000 individual interactions. Comparative analysis of the resulting data enabled direct identification of active proteins. Accuracy of this library analysis methodology was confirmed by both *in vitro* and *in vivo* analyses of identified proteins to yield novel zinc finger proteins that bind to their target sequences with high affinity, as indicated by low nanomolar apparent dissociation constants.

INTRODUCTION

Combinatorial chemistry is a well-established field that involves the high-throughput synthesis and subsequent screening of small organic molecules. Since Furka's initial demonstration of peptide libraries in the late 1980s (1–3), chemists have developed numerous synthetic and screening strategies (4–6). Key to all of these approaches is the identification of the active compound(s) from within a combinatorial library—a process termed deconvolution. Effective deconvolution invariably links synthesis with screening, the screening strategy employed dictating the synthetic approach used. For example, 'one bead one compound' libraries contain thousands of beads, each bearing multiple copies of a single library compound (7). The library is screened *en masse* for a desired activity and any active bead(s) isolated. The active compound borne by that bead(s) is then characterized either conventionally, for example, by Edman degradation of peptides (7) or via a chemical or radio frequency tag ('bar code') for other small molecule libraries (8). Alternatively, a reductive approach may be employed (9). After an initial 'active' library has been identified, a process of synthesis and screening of successively smaller libraries is performed until a single active compound is identified. Perhaps the most elegant approach to deconvolution is Houghten's positional fixing methodology, which intimately links the synthetic approach with the subsequent screening (10). In this approach a set of related libraries are generated before screening. Each library contains compounds with one specific residue 'fixed' as a single building block and the remaining residues fully randomized. Different libraries have a different building block at this 'fixed' position. Screening this set of libraries, for a desired activity, enables direct identification of the optimum building block at the 'fixed' residue. Other residues may be optimized in analogous fashion. This process may be carried out sequentially, optimizing one residue at a time (11), or alternatively all of the sets of libraries may be screened simultaneously (12,13) to allow the optimum library compound to be identified directly from a single round of screening.

Combinatorial protein libraries are similarly well-established and are generated at the genetic level through

*To whom correspondence should be addressed. Tel: +44 121 204 3961; Fax: +44 121 359 0733; Email: a.v.hine@aston.ac.uk
Present address:
Zhan-Ren Zhang, Millipore Bioprocessing Ltd, No. 1 Industrial Estate, Medomsley Road, Consett, Co. Durham DH8 6SZ, UK

site-directed gene randomization and/or gene shuffling (see below). In its simplest form, gene randomization involves the saturation mutagenesis of a single codon to encode each of the 20 amino acids, the resulting combinatorial library consisting of 19 new genes and 1 parental gene. Such libraries allow for the exhaustive exploration of a small, defined region of sequence space within a protein (typically 1–6 residues) and may be synthesized by a variety of methods [for a recent review see (14)]. Each method aims to randomize key residues involved in the function of the protein. Another commonly used approach to generate combinatorial libraries is gene shuffling (15), in which homologous genes, either naturally occurring or generated by error-prone PCR, are fragmented by nuclease digestion with DNAse. The gene fragments are then recombined using cycles of melting, annealing and extension to generate the randomized gene library. Screening the proteins produced from libraries generated by either approach may reveal variants with improved or novel activity.

A class of proteins that has been particularly amenable to site-specific gene randomization is the Cys_2-His_2 zinc finger domain. This is the most common DNA-binding motif found in eukaryotes with the zinc finger genes accounting for ~2% of the entire human genome, and is the second most common protein domain encoded by the human genome (16,17). The domains are ~30 amino acids long and comprise an anti-parallel β-sheet followed by an α-helix (18,19). Zinc finger proteins have a modular design in which each zinc finger domain recognizes a 3 base subsite (19) and proteins contain multiple adjacent domains that recognize a contiguous DNA sequence. Sequence recognition is mediated principally by amino acids in positions −1, 3 and 6 (numbered relative to their positions in the α-helix) which together contact the nucleotides within a 3 base subsite. Previous randomization studies have elucidated the zinc fingers capable of recognizing all 5′-ANN-3′ (20) and 5′-GNN-3′ (21) triplets within the context of a Zif268 derived protein. Such strategies have proven useful in developing novel transcription factors enabling the regulation of biologically important genes, such as repression/activation of *erbB-2* (22) and inhibition of viral replication (23,24) .

Regardless of the nature of the parental protein, the screening/deconvolution of combinatorial protein libraries is generally achieved by a process of biopanning, which encompasses aspects of both 'encoded one bead, one compound' and the reductive approaches employed in chemistry. Biopanning is generally applied to display libraries, such as phage (25), bacterial (26), ribosome (27), RNA (28) or on-bead display libraries (29), in which proteins are physically linked to their encoding sequence. This enables large libraries to be expressed and screened *en masse* for their ligand-binding properties. The library is screened by several rounds of biopanning, whereby the complexes are incubated with the desired target molecule, which is typically immobilized on a solid support. Subsequent washing removes the weakly binding proteins while retaining the desired proteins, which are then eluted for the next round of biopanning. Successively smaller numbers of displayed proteins are screened until 'active' proteins are isolated. The identity of the active protein borne on the display vehicle is established by sequencing the linked encoding sequence (DNA or RNA) of the carrier vehicle (phage, bacteria, ribosome or bead).

We wished to step away from this approach and be able to read the identity of the active protein directly from an initial screen of protein libraries in a manner similar to Houghten's positional fixing. While this approach has been applied routinely for deconvolution in combinatorial chemistry it is, to the best of our knowledge, yet to be applied to biological protein libraries. We aimed to generate and screen a series of overlapping protein libraries, such that the identity of active protein(s) may be interpreted directly from initial library screens. The encoding strategy is designed to allow simple and rapid deconvolution of the data to elucidate sequences that bind optimally to each DNA triplet, thereby eliminating the need for either purification or for a linked sequence tag. Our studies focus on the randomization of three principal DNA-contacting residues in the second finger of a three-zinc finger protein, ZFH, described previously (30), fused to green fluorescent protein (GFP) to generate His_6-ZFH-GFP (31). The principal DNA-contacting residues in fingers 1 and 3 were left unchanged to provide a constant framework within which to analyze binding of the second finger to any desired DNA triplet.

MATERIALS AND METHODS

Construction of pETcoco-LIB

For gene randomization, plasmid pETcoco-LIB was derived from plasmid pETcoco-ZFH-GFP (30). Briefly, a 37 bp cassette, encompassing the three codons to be randomized, was excised from pETcoco-ZFH-GFP by combined HindIII/BsiWI digestion. The cassette was then replaced with a 20 bp oligonucleotide cassette consisting of the hybridized oligonucleotides 5′-AGC TTC GTT CAC GTG ATG AC-3′ and 5′-GTA CGT CAT CAC GTG AAC GA-3′. This cassette was designed to introduce a PmlI restriction site between the HindIII and BsiWI sites, and also to cause a downstream frameshift that introduces a UAA termination codon into the zinc finger gene.

Gene library construction

Randomized DNA cassettes were constructed using 'MAX' randomization as described previously (32), with a 12-fold excess of selection oligonucleotides. This process permits non-redundant randomization, in which only the one favored codon for the expression of each required amino acid (all 20 or a subset thereof) is cloned. DNA cassettes were then ligated into 100 ng pETcoco-LIB vector, prepared by digestion with PmlI, dephosphorylated (to minimize subsequent self-ligation of any partially digested vector) and then restricted with BsiWI and HindIII. The ligations were extracted with phenol/chloroform and precipitated with ethanol overnight at −20°C. Recovered DNA was resuspended in 2 µl sterile dd H_2O. Meanwhile, electrocompetent Tuner (DE3) (Novagen) cells were prepared according to the manufacturer's instructions (Geneflow) and a 40 µl aliquot of the cells was added to the resuspended DNA. The cell/DNA mixture was incubated on ice for 1 min and then transferred into a chilled 2 mm electroporation cuvette. Electroporation was carried out at 2500 V in an Equibio Easyject Prima (Geneflow) and the cells were gently resuspended in 1 ml SOC media immediately after

electroporation. For assessment of transformation efficiency, a 20 µl aliquot of cells was plated onto Luria–Bertani (LB) supplemented with chloramphenicol (12.5 µg/ml) agar. The remainder of the transformed cells was used to generate a library, by inoculation into a 200 ml culture of LB supplemented with chloramphenicol (12.5 µg/ml) broth. Resulting libraries were accepted only if the accompanying plate contained a minimum of 120 colonies, which equates to 6120 individual clones within the liquid-culture library. Libraries were expressed and clarified lysates were prepared as described previously (31).

Library screening

Randomized, GFP-labeled protein libraries were screened against immobilized double stranded DNA of sequence 5′-T_{10}GGGXXXGCTT_{10}-3′ and 5′ biotinylated complement (where XXX refers to the appropriate DNA triplet) as described previously (31).

Yeast one-hybrid analysis

Reporter plasmid pLacZi (BD Biosciences, Franklin Lakes, NJ) was modified by insertion of the hybridized oligonucleotides 5′-AGC TTG AAT TCG AGC TCG GTA CCC GGG CAT CTA CAG ACC-3′ and 5′-TCG AGG TCT CTA GAT GCC CGG GTA CCG AGC TCG AAT-3′ between the HindIII and XhoI restriction sites to generate plasmid pLacZiX, which lacks the SalI restriction site and has an XbaI site within the MCS. Reporter constructs were then generated by ligating the hybridized oligonucleotides 5′-AAT TCG GGX XXG CTG GGX XXG CTG GGX XXG CTT-3′ and 5′-ACG GAA GCX XXC CCA GCX XXC CCA GCX XXC CCG-3′, where XXX represents the required codon, between the HindIII and XbaI sites of appropriately digested pLacZiX. This generates a reporter with three tandem copies of the DNA target upstream of the *lacZ* reporter.

Activation vectors were constructed in pGAD424 (BD Biosciences). Initially, the vector was modified to remove HindIII sites by standard mutagenic protocols. The modified zinc finger gene was then amplified from pETcoco-LIB with primers 5′-GGG GGA ATT CGA GAA ACT TCG TAA TGG TTC-3′ and 5′-GGG GGG ATC CTC ATT TCT TGT TCT GAT G-3′, which introduce 5′ EcoRI and 3′ BamHI restriction sites, respectively. The amplified gene was then inserted between similarly digested, modified pGAD424, to generate pGADLIB. DNA cassettes were constructed as described above and inserted between the HindIII and BsiWI sites in this vector.

Reporter plasmids were integrated into the genome of yeast strain YM4271 according to the manufacturer's instructions. Each of the resulting reporter strains was then transformed with its respective activation vector and the resulting transformants were recovered on selective media and then replica plated onto similar media containing X-gal, according to the manufacturer's instructions. Blue colonies were picked for further analysis.

Single zinc finger protein production

Individual zinc finger genes were constructed by inserting the hybridized oligonucleotides 5′-AGC TTT AGT XXX AGC GAC YYY TTA CAA ZZZ CAT CAG C-3′ and 5′-GTA CGC TGA TGT ZZZ TGT AAG YYY TCG CTC XXX CTA A-3′ (where XXX, YYY and ZZZ refer to the required codons at positions −1, 3 and 6, respectively) into HindIII/BsiWI digested pGEX-ZFH (30). The genes were expressed and protein purified essentially as described previously (33), except that protein production was on a 1/10th scale and relative yields were increased by employing 0.5 ml glutathione Sepharose 4B resin, a 30 min adsorption time and elution with 2 × 0.8 ml glutathione buffer.

RESULTS

Library construction

We aimed to generate libraries in which clonal representation was high and encoded protein bias was minimized. To achieve this goal, there were three principal considerations. First, the potential toxicity of clones to *Escherichia coli* must be minimized so that clones are not depleted prior to the induction of expression. Second, there must be no excess of the parental gene after randomization. Last and the most important, the randomization process itself must not be subjected to the inherent bias dictated by the genetic code.

Since zinc finger proteins bind DNA, those that bind tightly to sequences present within the *E.coli* chromosome are likely to have a detrimental effect on *E.coli* growth. Previous studies with a model 'tight' zinc finger protein (ZFH) and a deletion mutant of the same gene (the small deletion caused a frameshift during expression) had shown that the plasmid encoding the functional high-affinity protein transformed poorly when compared with its variant plasmid that encoded the non-functional protein. This problem occurred even in the absence of induction of protein expression and was ameliorated only partially by the use of a traditional T7 expression system (34), which is generally considered sufficient to circumvent toxicity problems. Such bias in transformation has important implications for library construction, since plasmids encoding the proteins with the 'best' activities (i.e. those that bind tightly to their target sequences) are likely to be selected against and therefore, to be absent or, at best, poorly represented in the resulting protein libraries. This potential problem of differential transformation was circumvented by cloning the parental *ZFH* zinc finger gene in the vector pETcoco-1. This vector is designed to exist as a single copy within the host strain thus minimizing the effects of any residual, leaky expression from the T7 promoter. The resulting construct, pETcocoH_6-ZFH-GFP transformed freely into *E.coli*.

To eliminate the potential for over-representation of the parental zinc finger protein, the parental *ZFH* gene was then modified. The resulting plasmid, pETcoco-LIB, contains both a frameshift mutation and a termination codon within its *ZFH* gene, so that self-ligation of the plasmid cannot result in the subsequent translation of parental zinc finger protein. Rather, intact zinc finger/GFP fusions can result only from the correct insertion of a randomized DNA cassette (see Materials and Methods). Since the randomization cassette must be dephosphorylated to prevent its concatamerization during cloning, the modified vector pETcoco-LIB also contains a PmlI site between the HindIII and BsiWI cloning sites. Digestion with PmlI and subsequent dephosphorylation of

the vector prior to HindIII and BsiWI digestion effectively prevents regeneration of the parental vector by ligation, thus maximizing the potential yield of recombinant clones.

Finally, to remove the effects of inherent bias of the genetic code, DNA cassettes were generated by 'MAX' randomization (32), which allows for non-degenerate encoding of the required amino acids at each randomized position. This technique permitted facile construction of a series of 60 overlapping DNA cassettes from one set of oligonucleotides. These cassettes were then cloned into plasmid pETcoco-LIB to generate 60 randomized gene libraries. In each library, one of the principal DNA-contacting residues was fixed, e.g. position −1 as alanine, while at positions 3 and 6 (the other two positions principally involved in DNA recognition) all 20 amino acids were encoded to give equal representation of every amino acid. The randomization strategy was repeated, fixing each of the 20 amino acids at each of the 3 principal DNA-contacting positions. Thus, each library encoded a mixture of 400 individual proteins (Figure 1). We wished to construct libraries that were as close to 100% complete as possible. Using the equation described by Patrick *et al*. (35) we calculated that for a 99.9% probability of the library containing every possible randomized sequence, we required ∼13-fold excess of clones. Our acceptance criteria for a library was therefore that it should contain sufficient clones for a 15-fold excess (i.e. 15 × 400 = 6000). In practice, the majority of our libraries exceeded this figure and typically libraries contained sufficient clones to equate to a >50-fold excess. The resulting gene libraries were expressed.

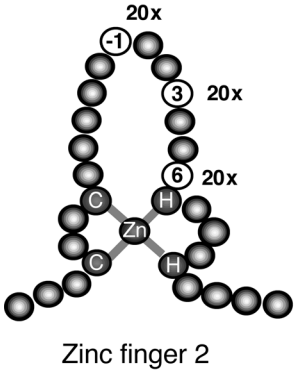
Encoded deconvolution: library screening and data processing

To determine whether the principle of positional fixing would be applicable to such complex mixtures of proteins, five DNA targets were selected. Rather than engaging in laborious protein purifications, we wished to be able to screen protein–ligand interactions directly from within the crude bacterial lysates. Accordingly, protein libraries were not purified; rather screening experiments were conducted directly, with clarified *E.coli* lysates containing libraries of zinc finger–GFP fusions. Total zinc finger protein content was estimated by measuring the fluorescence of the GFP tag fused to the zinc finger. The interaction of each of the 60 positionally fixed libraries was measured against each of the five target DNAs. Interactions were measured using a plate-based protein–DNA interaction assay, suitable for measuring mixtures of proteins (31).

Within the limits of the protein concentrations in our libraries, preliminary experiments demonstrated a direct relationship between the amount of GFP-labeled protein present and the signal generated (data not shown). The data from the library screening was therefore scaled according to the total amount of GFP fluorescence present in each library. To facilitate cross-comparison between individual library/DNA combinations, the data was then normalized (where the highest signal after three washes = 100%) and plotted (Figure 2). The putative identities of interacting proteins were then read directly from these graphs.

Encoded deconvolution: data interpretation

In most cases, the graphs suggest one or two possible candidate proteins for interaction with the target DNA sequence. For example, Figure 2A suggests clearly that the zinc finger protein for recognition of the triplet 5′-TCC-3′ will be either WNK or WYK while that for 5′-GCA-3′ (Figure 2B) is either TRR or WRR. In some cases, the picture is a little more complex. For example, binding data to the triplet 5′-GTT-3′, infers multiple amino acid possibilities (Figure 2C; T, W or Y at the −1 position, S, C or P at position 3 and R at position 6) suggesting perhaps that multiple zinc finger proteins are capable of recognizing 5′-GTT-3′ with similar affinity. We considered data from both three and four washes since although in most cases data from four washes shows very similar trends to that from three washes, there are some exceptions. For example, preliminary examination of Figure 2D suggests that protein RTR will best recognize triplet 5′-ACG-3′, but careful examination shows that the signal from R in position 6 diminishes to a level comparable with that from C after four washes.



Library	-1	3	6
1	A	MAX	MAX
2	C	MAX	MAX
3	D	MAX	MAX
4	E	MAX	MAX
5	F	MAX	MAX
6	G	MAX	MAX
7	H	MAX	MAX
8	I	MAX	MAX
9	K	MAX	MAX
10	L	MAX	MAX
11	M	MAX	MAX
12	N	MAX	MAX
13	P	MAX	MAX
14	Q	MAX	MAX
15	R	MAX	MAX
16	S	MAX	MAX
17	T	MAX	MAX
18	V	MAX	MAX
19	W	MAX	MAX
20	Y	MAX	MAX

Library	-1	3	6
21	MAX	A	MAX
22	MAX	C	MAX
23	MAX	D	MAX
24	MAX	E	MAX
25	MAX	F	MAX
26	MAX	G	MAX
27	MAX	H	MAX
28	MAX	I	MAX
29	MAX	K	MAX
30	MAX	L	MAX
31	MAX	M	MAX
32	MAX	N	MAX
33	MAX	P	MAX
34	MAX	Q	MAX
35	MAX	R	MAX
36	MAX	S	MAX
37	MAX	T	MAX
38	MAX	V	MAX
39	MAX	W	MAX
40	MAX	Y	MAX

Library	-1	3	6
41	MAX	MAX	A
42	MAX	MAX	C
43	MAX	MAX	D
44	MAX	MAX	E
45	MAX	MAX	F
46	MAX	MAX	G
47	MAX	MAX	H
48	MAX	MAX	I
49	MAX	MAX	K
50	MAX	MAX	L
51	MAX	MAX	M
52	MAX	MAX	N
53	MAX	MAX	P
54	MAX	MAX	Q
55	MAX	MAX	R
56	MAX	MAX	S
57	MAX	MAX	T
58	MAX	MAX	V
59	MAX	MAX	W
60	MAX	MAX	Y

Figure 1. Composition of the randomized zinc finger protein libraries 1–60. A schematic representation of zinc finger 2 is shown highlighting the positions of randomization. The identities of amino acid residues, encoded within each library, in the positions −1, 3 and 6 of the α-helix of zinc finger 2 are indicated, where 'MAX' denotes a mixture of all 20 amino acids (comprising one codon optimal for the expression of each amino acid in *E.coli*). Zinc finger gene libraries were constructed by 'MAX' randomization (32) of the *ZFH* gene as described in Materials and Methods.

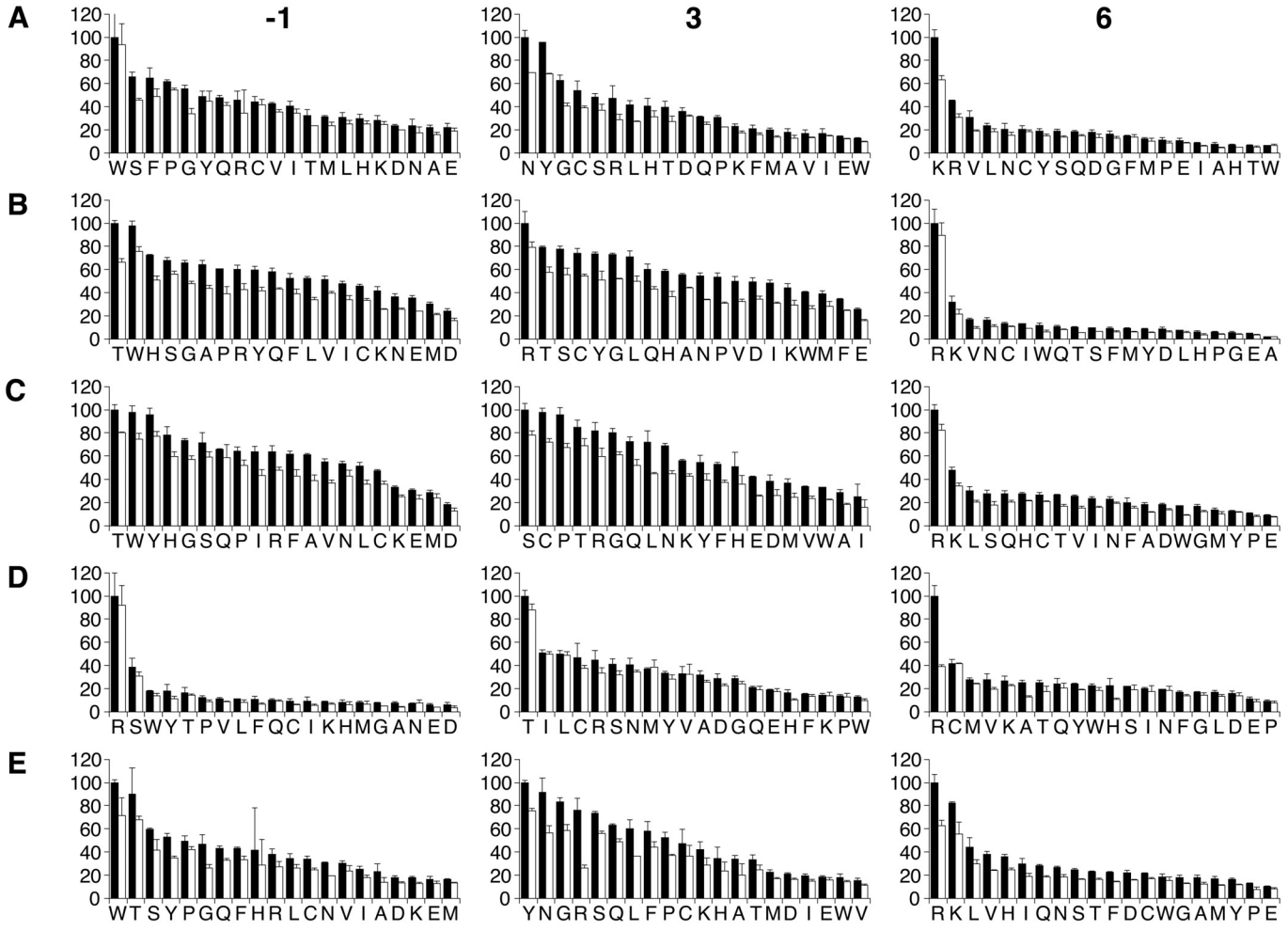


Figure 2. Processed screening data for zinc finger libraries. Randomized protein libraries 1–60 were screened against target DNA sequences containing the triplets (**A**) TCC, (**B**) GCA, (**C**) GTT, (**D**) ACG and (**E**) TAC. The resulting data were scaled according to the relative GFP fluorescence in each protein library, normalized (where 100% = the highest signal after three washes) and sorted by the highest signal. Numbers on the ordinate are normalized DNA binding (%) and letters on the abscissa are the identity of the fixed amino acid in positions −1, 3 and 6 (Figure 1). Filled bars represent data measured after three washes and open bars represent data obtained after four washes, as described in Materials and Methods. All experiments were performed in triplicate. Error bars represent 1 SD.

Thus, both RTR and RTC were investigated further. Data from the fourth wash may also be used to eliminate possibilities. For example, preliminary examination of Figure 2E suggests W or T at position −1, followed by Y or N at position 3 with R at position 6 for recognition of the triplet 5′-TAC-3′. However, the signal from N after three washes at position 3 has a relatively large error bar and is diminished in magnitude after four washes. Therefore, we elected to study only TYR and WYR for this triplet. The full list of proteins selected for further study is given in Table 1.

Table 1. Summary of the processed library screening data

Target DNA	−1	3	6
5′-TCC-3′	W	N/Y	K
5′-GCA-3′	T/W	R	R
5′-GTT-3′	T/W/Y	S/C/P	R
5′-ACG-3′	R	T	R/C
5′-TAC-3′	T/W	Y	R

Identities of the potential interacting residues at positions −1, 3 and 6 for each of the five target DNA sequences was derived from the normalized graphs illustrated in Figure 2, after consideration of data from both three and four washes.

Target confirmation: *in vivo*, cell-based assays

Would identities read from these graphs be biologically active? It is likely that many different zinc fingers can bind to the target sequences and it was the aim of our approach to attempt to find the 'best' protein from these possibilities. This is particularly important when multiple possibilities exist (e.g. Figure 2C). Are all of these combinations viable proteins or are some of them simply mathematical combinations that have no activity in nature? To examine these possibilities and to establish whether our engineered proteins bind to DNA *in vivo*, 'mini libraries' were constructed as defined in Table 1, by using 'MAX' randomization. These 'mini libraries' were then screened using the yeast one-hybrid assay (BD Biosciences). Here, the target DNA sequence is placed upstream of a reporter gene (β-galactosidase) while the protein of interest is fused to a yeast transcriptional activator domain. If the zinc finger recognizes its target DNA *in vivo* (i.e. within the yeast cell), β-galactosidase gene expression

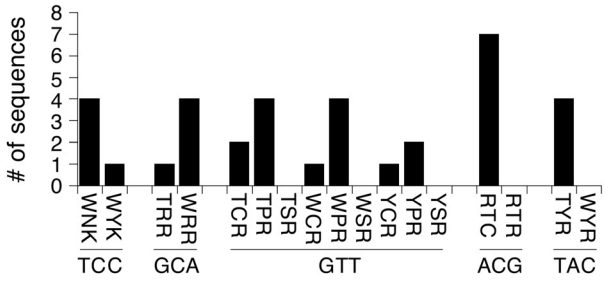


Figure 3. Histogram showing active zinc finger proteins identified by the *in vivo* (cell-based) assay. 'Mini libraries' encoding all potential candidate proteins identified from processed zinc finger library data (Table 1) were generated by MAX randomization and their interactions *in vivo* with their respective target DNAs assessed by yeast one-hybrid analysis as described in Materials and Methods. For each target DNA sequence, positively staining blue colonies were picked and sequenced to establish the identity of the amino acids encoded at the positions −1, 3 and 6 of zinc finger 2. On the abscissa, vertical lettering indicates amino acid identity and horizontal lettering indicates the target DNA sequence.

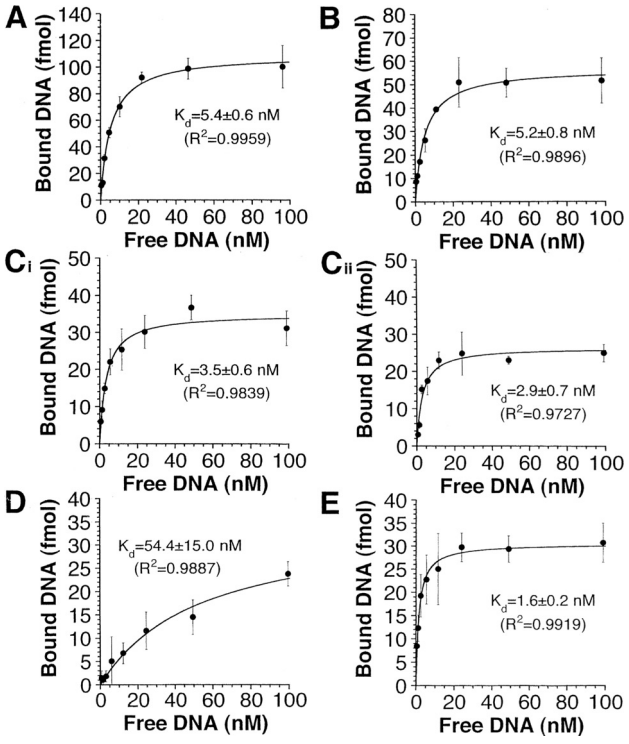


Figure 4. Measurement of apparent dissociation constants ($K_{d_{app}}$) between purified zinc finger proteins and their predicted target DNA sequences. $K_{d_{app}}$ values were estimated as described previously (24) with the exceptions that assays were performed in HE^{TM} microplates (Molecular Devices, Sunnyvale, CA) and that microplates were washed using a Multiwash Advantage (Tri Continent, Grass Valley, CA) plate washer. Zinc finger proteins and corresponding DNA sequences were: (**A**) WNK with 5′-TCC-3′; (**B**) WRR with 5′-GCA-3′; (**C_i**) TPR and (**C_{ii}**) WPR, both with 5′-GTT-3′; (**D**) RTC with 5′-ACG-3′; and (**E**) TYR with 5′-TAC-3′, where letters represent the amino acids at positions −1, 3 and 6 in the α-helix in the second zinc finger of protein ZFH (30). All experiments were performed in quadruplicate. Error bars represent 1 SD.

will be activated, resulting in the hydrolysis of X-Gal and consequent blue staining of the individual yeast colony. Alternatively, if the zinc finger fails to bind its DNA target sequence, no activation will occur and the colony will remain white. Approximately 5–10 blue colonies were picked from each mini library and their zinc finger genes sequenced (Figure 3) to generate qualitative data about zinc finger activity in a cellular environment. The majority of *in vivo* analyses give a clear conclusion regarding the identity of the 'best' interacting zinc finger protein and since we were interested in identifying an optimal zinc finger domain for each DNA triplet, the proteins that led most frequently to β-galactosidase activation were chosen for further *in vitro* study. However, as suggested by the original library assays, it would appear that multiple proteins are indeed capable of binding tightly to the triplet 5′-GTT-3′. The two proteins that occurred most frequently in the *in vivo* analysis of this triplet were both therefore selected for further study.

Target confirmation: *in vitro* assays

To confirm the DNA-binding activity of each identified protein, individual zinc finger genes were constructed, expressed and the resulting proteins were purified. The interactions of these proteins were then confirmed *in vitro*, as described previously (30). The apparent dissociation constant of each protein with its target DNA sequence was then estimated. Each protein bound to its target DNA sequence with an apparent K_d ($K_{d_{app}}$) in the low-nanomolar range (Figure 4). We therefore conclude that the process of direct deconvolution is indeed feasible for the direct analysis of protein libraries.

DISCUSSION

We have successfully read the identity of biologically active proteins directly from crude protein libraries, without recourse to multiple rounds of biopanning, protein purification or sequence tagging of clones. Screening of mixtures of proteins, in this case 400 per library (20 × 20 × 1), has allowed for the testing of a potential 40 000 interactions (a total of 8000 different proteins × 5 DNA targets) within only 300 assays (60 libraries × 5 DNA targets). The data obtained in this manner have been refined by *in vivo*, cell-based analysis and validated by subsequent *in vitro* analysis (of purified zinc fingers) to elicit novel proteins that bind to their ligands with low-nanomolar affinity. The randomization approach employed generates non-degenerate gene libraries in which individual proteins are encoded uniformly (32). High-quality randomization is playing an increasingly important role in protein engineering (36) and we believe that it plays an important role in the ability to link directly from the analysis of crude proteins to the identification of high-affinity 'hits'. Using conventional randomization at three positions, differences in the concentrations of individual proteins within a library can vary by >1000-fold (32) and it is unlikely that mass screening would generate true results under such conditions. The signal from a low-affinity protein present in excess would be likely to mask that of a high affinity, but poorly represented protein.

In every case, yeast one-hybrid analysis confirmed the activity of at least one candidate protein, thereby validating the screening approach employed in direct deconvolution. For the majority of our DNA targets, direct analysis of the library

screening data generated only two candidate proteins. Thus in future studies, it would be expedient to conduct *in vitro* analysis of both the purified candidates to determine their affinities for the target ligand. In contrast, where a small pool of candidate proteins results (e.g. Figures 2C and 4C_i, C_{ii}), an *in vivo* screening stage may be the most appropriate to eliminate both mathematical possibilities and lower affinity candidate proteins before *in vitro* analysis. Interestingly, the two candidate proteins identified by *in vivo* screening both bound to their target ligand with similar, high affinity (Figure 4). Thus, it is reasonable to speculate that multiple inferences from library screening data are likely to represent multiple active proteins with similar affinity for the target ligand. These may reflect a preference during the selection for residues that are tolerant to mutations at other randomized positions. Nevertheless, the nature of data generated by direct screening will sometimes generate mathematical combinations of residues that have either low or no biological activity and therefore, a final analysis of the activity of individual proteins should always be performed.

With specific reference to zinc fingers, we have generated novel proteins that bind to the triplets ACG, GCA, GTT, TAC and TCC, respectively. The first three of these sequences might be expected: zinc finger proteins that bind to the triplets 5′-ANN-3′ and 5′-GNN-3′, albeit within different contexts (different adjacent zinc finger domains) and within different zinc finger scaffolds, have been reported previously (20,21). Interestingly, these studies also generated multiple proteins rather than a single consensus for target 5′-GTT-3′, while a protein with residues RTD in the positions −1, 3 and 6 of the α-helix bound sequence 5′-ACG-3′ (20), which is similar to RTC identified in the present study (Figure 4D). Conversely, the proteins identified to bind to 5′-GCA-3′ were quite different, having residues QDR (21) and WRR (Figure 4B) and at the positions −1, 3 and 6 of the α-helix, respectively. We were more surprised to find zinc finger proteins so readily that bind to triplets commencing with a 5′-T, since reports of engineered zinc fingers that bind to 5′-TNN-3′ sequences are scarce (37). The results for 5′-TAC-3′ and 5′-TCC-3′ are interesting in that such closely related nucleotide sequences are targeted by two quite different proteins, TYR and WNK, respectively. The selection of tryptophan is of particular interest as it is encoded only once within the genetic code and is therefore likely to be poorly represented within the conventional NNN or NNG/T randomized libraries (32).

We were particularly surprised to find proline at position 3, in the middle of the α-helix. Proline is rarely found within the center of α-helices (38); since prolines tend to reside within the first turn of α-helices (39). Proline is well known to destabilize α-helices as it is unable to form main chain hydrogen bonds. Therefore, it is likely to have an effect on the structure and stability of this α-helix. However, previous studies have observed proline residues within the α-helix of randomized zinc finger proteins following selection (37) and occasionally within other proteins (40). In the context of the highly stable zinc finger domain, the location of proline is perhaps less surprising and structural analysis of this protein domain may elicit interesting new secondary structure.

In summary, use of positionally fixed protein libraries has enabled the discovery of novel zinc finger proteins with high affinities for their DNA targets. This approach allows for the direct identification of candidate proteins, without recourse to multiple rounds of biopanning, protein purification or sequence tagging of clones. As such, it offers advantages over the conventional randomization and biopanning strategies and may improve the quality of the 'hits' obtained from combinatorial libraries.

ACKNOWLEDGEMENTS

This work was supported by grant B14245 from the BBSRC to A.V.H. and A.J.S. We gratefully acknowledge Prof. J. D. Sutherland (University of Manchester) and Dr R. A. J. Darby (Aston University) for critical evaluation of the manuscript. A.V.H. and M.D.H. would like to declare a potential conflict of interest in that they both own shares in a University spin-out company, ProtaMAX Ltd, which commercializes 'MAX' randomization, a technique used in this paper. Funding to pay the Open Access publication charges for this article was provided by BBSRC Grant B14245.

REFERENCES

1. Furka,A., Sebestyen,F., Asgedom,M. and Dibo,G. (1988) Cornucopia of peptides by synthesis. In *Proceedings of the 14th International Congress of Biochemistry: Highlights of Modern Biochemistry*, VSP, Utrecht, The Netherlands, Vol. 5, pp. 47.
2. Furka,A., Sebestyen,F., Asgedom,M. and Dibo,G. (1988) More peptides by less labour. *Proceedings of the 10th International Symposium of Medicinal Chemistry*, Budapest, Hungary, pp. 288.
3. Furka,A. (1995) History of combinatorial chemistry. *Drug Dev. Res.*, **36**, 1–12.
4. Sanchez-Martin,R.M., Mittoo,S. and Bradley,M. (2004) The impact of combinatorial methodologies on medicinal chemistry. *Curr. Top. Med. Chem.*, **4**, 653–669.
5. Geysen,H.M., Schoenen,F., Wagner,D. and Wagner,R. (2003) Combinatorial compound libraries for drug discovery: an ongoing challenge. *Nature Rev. Drug Discov.*, **2**, 222–230.
6. Hoogenboom,R., Meier,M.A.R. and Schubert,U.S. (2003) Combinatorial methods, automated synthesis and high-throughput screening in polymer research: past and present. *Macromol. Rapid Comm.*, **24**, 16–32.
7. Lam,K.S., Salmon,S.E., Hersh,E.M., Hruby,V.J., Kazmierski,W.M. and Knapp,R.J. (1991) A new type of synthetic peptide library for identifying ligand-binding activity. *Nature*, **354**, 82–84.
8. Ede,N.J. and Wu,Z.M. (2003) Beyond Rf tagging. *Curr. Opin. Chem. Biol.*, **7**, 374–379.
9. Carell,T., Wintner,E.A., Sutherland,A.J., Rebek,J.,Jr, Dunayevskiy,Y.M. and Vouros,P.M. (1995) New promise in combinatorial chemistry: synthesis, characterisation, and screening of small-molecule libraries in solution. *Chem. Biol.*, **2**, 171–183.
10. Houghten,R.A., Dooley,C.T. and Appel,J.R. (2004) *De novo* identification of highly active fluorescent kappa opioid ligands from a rhodamine labeled tetrapeptide positional scanning library. *Bioorg. Med. Chem. Lett.*, **14**, 1947–1951.
11. Houghten,R.A., Pinilla,C., Blondelle,S.E., Appel,J.R., Dooley,C.T. and Cuervo,J.H. (1991) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature*, **354**, 84–86.
12. Dooley,C.T., Chung,N.N., Wilkes,B.C., Schiller,P.W., Bidlack,J.M., Pasternak,G.W. and Houghten,R.A. (1994) An all D-amino acid opioid peptide with central analgesic activity from a combinatorial library. *Science*, **266**, 2019–2022.
13. Dooley,C.T., Chung,N.N., Schiller,P.W. and Houghten,R.A. (1993) Acetalins—opioid receptor anatagonists determined through the use of synthetic peptide libraries. *Proc. Natl Acad. Sci. USA*, **90**, 10811–10815.
14. Neylon,C. (2004) Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res.*, **22**, 1448–1459.

15. Stemmer,W.P.C. (1994) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.
16. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
17. International Human Genome Consortium (2001), Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
18. Rhodes,D. and Klug,A. (1986) An underlying repeat in some transcriptional control sequences corresponding to half a double helical turn of DNA. *Cell*, **46**, 123–132.
19. Pavelitch,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å. *Science*, **252**, 809–817.
20. Dreier,B., Beerli,R.R., Segal,D.J., Flippin,J.D. and Barbas,C.F.,III (2001) Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, **276**, 29466–29478.
21. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F.,III (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
22. Liu,Q., Segal,D.J., Ghiara,J.B. and Barbas,C.F.,III (1997) Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc. Natl Acad. Sci. USA*, **94**, 5525–5530.
23. Beerli,R.R., Segal,D.J., Dreier,B. and Barbas,C.F.,III (1998) Toward controlling gene expression at will: specific regulation of the erB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc. Natl Acad. Sci. USA*, **95**, 14628–14633.
24. Reynolds,L., Ullman,C., Moore,M., Isalan,M., West,M.J., Clapohah,P.O., Klug,A. and Choo,Y. (2003) Repression of the HIV-1 LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. *Proc. Natl Acad. Sci. USA*, **100**, 1615–1620.
25. Hoess,R.H. (2001) Protein design and phage display. *Chem. Rev.*, **101**, 3205–3218.
26. Samuelson,P., Gunneriusson,E., Nygren,P.-A. and Stahl,S. (2002) Display of proteins on bacteria. *J. Biotechnol.*, **96**, 129–154.
27. Mossner,E. and Pluckthun,A. (2001) Directed evolution with fast and efficient selection technologies. *Chimia*, **55**, 325–329.
28. Takahashi,T.T., Austin,R.J. and Roberts,R.W. (2003) mRNA display: ligand discovery, interaction analysis and beyond. *Trends Biochem. Sci.*, **28**, 159–165.
29. Nord,O., Uhlen,M. and Nygren,P.-A. (2003) Microbead display of proteins by cell-free expression of anchored DNA. *J. Biotechnol.*, **106**, 1–13.
30. Zhang,Z.-R., Palfrey,D., Nagel,D.A., Lambert,P.A., Jessop,R.A., Santos,A.F. and Hine,A.V. (2003) Fluorescent microplate-based analysis of protein–DNA interactions I: immobilized protein. *BioTechniques*, **35**, 980–986.
31. Zhang,Z.-R., Hughes,M.D., Morgan,L.J., Santos,A.F. and Hine,A.V. (2003) Fluorescent microplate-based analysis of protein–DNA interactions II: immobilized DNA. *BioTechniques*, **35**, 988–996.
32. Hughes,M.D., Nagel,D.A., Santos,A.F., Sutherland,A.J. and Hine,A.V. (2003) Removing the redundancy from randomized gene libraries. *J. Mol. Biol.*, **331**, 973–979.
33. Woodgate,J.H., Palfrey,D., Nagel,D.A., Hine,A.V. and Slater,K.H. (2001) Protein-mediated isolation of plasmid DNA by a zinc finger-glutathione *S*-transferase affinity linker. *Biotechnol. Bioeng.*, **79**, 450–456.
34. Nagel,D.A. (2004) Development of the MAX randomization technique. PhD Thesis, Aston University, Birmingham, UK.
35. Patrick,W.M., Firth,E.M. and Blackburn,J.M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng.*, **16**, 451–457.
36. Lutz,S. and Patrick,W.M. (2004) Novel methods for directed evolution of enzymes: quality, not quantity. *Curr. Opin. Biotechnol.*, **15**, 291–297.
37. Wu,H., Yang,W.-P. and Barbas,C.F.,III (1995) Building zinc fingers by selection: toward a therapeutic application. *Proc. Natl Acad. Sci. USA*, **92**, 344–348.
38. MacArthur,M.W. and Thornton,J.M. (1991) Influence of proline residues on protein conformation. *J. Mol. Biol.*, **218**, 397–412.
39. Richardson,J.S. and Richardson,D.C. (1988) Amino-acid preferences for specific locations at the end of alpha-helices. *Science*, **240**, 1648–1652.
40. Rudresh, Jain,R., Dani,V., Mitra,A., Srivastava,S., Sarma,S.P., Varadarajan,R. and Ramakumar,S. (2002) Structural consequences of replacement of an alpha-helical Pro residue in *Escherichia coli* thioredoxin. *Protein Eng.*, **15**, 627–633.