Nucleic Acids Research, 1998, Vol. 26, No. 2 623–630

Green fluorescent protein as a scaffold for intracellular presentation of peptides

Majid R. Abedi, Giordano Caponigro and Alexander Kamb*

Ventana Genetics Inc., 421 Wakara Way, Salt Lake City, UT 84108, USA

Received August 21, 1997; Revised and Accepted November 6, 1997

ABSTRACT

Peptide aptamers provide probes for biological processes and adjuncts for development of novel pharmaceutical molecules. Such aptamers are analogous to compounds derived from combinatorial chemical libraries which have specific binding or inhibitory activities. Much as it is generally difficult to determine the composition of combinatorial chemical libraries in a quantitative manner, determining the quality and characteristics of peptide libraries displayed *in vivo* is problematical. To help address these issues we have adapted green fluorescent protein (GFP) as a scaffold for display of conformationally constrained peptides. The GFP–peptide libraries permit analysis of library diversity and expression levels in cells and allow enrichment of the libraries for sequences with predetermined characteristics, such as high expression of correctly folded protein, by selection for high fluorescence.

INTRODUCTION

A flurry of recent experiments has focused on the discovery of proteins and peptides that bind specific targets. The primary impetus for such experiments is the desire either to isolate peptide-based inhibitors of particular targets or to establish the normal binding partners of known proteins in the cell. Methods for ligand discovery that incorporate genetic techniques have enjoyed increased popularity, primarily because they are more rapid and less labor intensive than methods that rely exclusively on biochemical purification.

Two general types of genetic procedure have been applied, one designed to find peptide-based ligands *in vitro* (phage display), the other tailored to elucidation of protein–protein interactions *in vivo* (two hybrid analysis). Phage display involves the presentation of peptide sequences on the surface of filamentous phage particles such as M13 and f1 or on lytic viruses such as λ and T4 (1,2). Phage display of peptides or proteins is well suited to biochemical enrichment of ligands *in vitro* (and genetic amplification *in vivo*), but it is not appropriate for genetic selections that involve intracellular processes. To define peptide or protein ligands by experiments *in vivo*, expression systems that operate inside cells must be employed. For example, construction of a catalog of all interacting proteins and their binding partners (an ‘interaction map’ or ‘proteome map’) is presently best addressed by the two hybrid system using expression in yeast of whole proteins or protein domains (3,4). For this application proteins or relatively large protein fragments are required. But for other applications it is advantageous not to be restricted to natural protein sequences. To identify novel proteinaceous ligands for specific targets, it may be simpler to generate and examine a chemically diverse library of low molecular weight compounds based on peptides.

Though a potentially powerful tool, intracellular expression of peptide libraries suffers from several limitations. First, it is often difficult to know what the expression level of specific peptides or peptide fusions is; in many cases even an average measure of expression level is difficult to obtain. Second, the diversity of the library is not easily estimated, for example, a particular expression system may efficiently present only a small subset of possible peptide sequences. Third, it is typically impossible to follow the expression of peptides in particular cells, for example, to know whether or not a specific cell is expressing a member of the library. Fourth, it is not generally feasible to manipulate the library to alter its average properties once the library has been generated, for example, to bias the library toward sequences compatible with high expression. Fifth, efforts to restrict conformational freedom (in order to promote higher binding energies) by inserting the peptides into the interior of protein sequences may compound problems discussed above; such insertion libraries are likely to perturb the function and stability of the fusion partners in ways difficult to predict and measure.

Here we present a method for construction of peptide or protein fragment libraries using the autofluorescent polypeptide green fluorescent protein (GFP). These libraries contain sequences inserted within the GFP coding region and are suitable for experiments that require intracellular expression. The properties of the library can be easily and quantitatively monitored. Furthermore, the individual members of the library can be followed while they are expressed in cells using instruments such as a flow sorter (e.g. a FACS machine) and low and high expressors (corroborated by subsequent biochemical experiments) can be identified. This permits screening of the expression library on a cell-by-cell basis and enrichment for library sequences that have predetermined characteristics.

MATERIALS AND METHODS

Yeast strains and media

yVT12 (*MAT**a**, HML**a**, HMR**a**, sst2*Δ, *mfa1*Δ::*hisG*, *mfa2*Δ::*hisG*, *ade2-1*, *leu2-3*, *lys2*, *ura3-1*, *STE3*::*GAL1-STE3*::*HIS3*) was derived from JRY5312 (5) (kindly provided by Dr J.Rine). Yeast transformations were performed using the lithium acetate method

*To whom correspondence should be addressed. Tel: +1 801 581 1049; Fax: +1 801 581 9555; Email: kamb@vengen.com

(6) and transformants were selected and maintained on standard synthetic medium lacking uracil.

Plasmids

Construction of pVT21. pVT21 was obtained by manipulation of pACA151 (kindly provided by Dr J.Rine). pACA151 is a 6.7 kb 2μ yeast shuttle vector which carries the *URA3* and β-*Lactamase* genes; in addition it contains a *GFP* expression cassette made up of the *GAL*1/10 UAS, the coding region of a red shifted (S65T) *GFP* gene and the phosphoglycerate kinase (*PGK1*) 3′-end. To construct pVT21 the *Eco*RI site in pACA151 was converted into a *Bgl*II site. In addition, the *PGK1* 3′-end fragment of pACA151 was replaced with a 700 bp fragment (containing *Nar*I and *Bgl*II ends) which contained a modified *PGK1* 3′-end with termination codons in three reading frames.

Construction of pVT22–pVT31. pVT21 was used as the parent vector for pVT22–pVT31. In order to construct pVT22, pVT21 was used as template in two separate PCR reactions using primer pairs oVT329, oVT307 and oVT330, oVT317. Recombinant Pfu polymerase (Stratagene, La Jolla, CA) was used in all PCR amplifications. The termini of the resulting fragments contained *Xho*I/*Eco*RI and *Bam*HI/*Eco*RI restriction sites respectively. These two fragments were digested with *Eco*RI (New England Biolabs, Beverly, MA), ligated using T4 DNA ligase (Boehringer Mannheim, Germany) and PCR amplified using primers oVT329 and oVT330. The resulting 2 kb fragment contained the *GAL*1 UAS and *PGK1* 3′-UTR, as well as a *GFP* gene with a 6 codon insert corresponding to *Xho*I/*Eco*RI/*Bam*HI recognition sequences. pVT22 was obtained by digesting this 2 kb fragment with *Pst*I and *Hin*dIII and inserting it into the pVT21 backbone (also digested with *Pst*I and *Hin*dIII). pVT23–pVT31 were constructed using an identical cloning strategy except that instead of oVT307 and oVT317 the following primers were used: pVT23, oVT308 and oVT318; pVT24, oVT309 and oVT319; pVT25, oVT310 and oVT320; pVT26, oVT311 and oVT321; pVT27, oVT312 and oVT322; pVT28, oVT313 and oVT323; pVT29, oVT314 and oVT324; pVT30, oVT315 and oVT325; pVT31, oVT316 and oVT326 (see below).

Construction of GFP–peptide libraries

Random 20 amino acid libraries. One picomole of APT1 was annealed to 1 pmol APT2 and the second strand synthesized using Klenow fragment (Promega, Madison, WI). The resulting double-stranded oligonucleotide was digested with *Bam*HI and *Xho*I, inserted into *Bam*HI/*Xho*I-cut vector (pVT27, pVT28 or pVT29) and transformed into *Escherichia coli* DH5α.

Random 15 amino acid library in pVT27. Library construction was performed as in Cwirla *et al*. (1). Briefly, APT3 (15 pmol), APT4 and APT5 were mixed in a molar ratio of 1:50:50 and annealed in 20 mM Tris–HCl, pH 7.5, 2 mM $MgCl_2$, 50 mM NaCl by heating to 70°C for 5 min. The solution was allowed to cool to room temperature and ligated to 20 μg *Bam*HI/*Xho*I-cut pVT27 using 40 U T4 DNA ligase (Boehringer Mannheim, Germany). The ligated DNA was concentrated and desalted using Ultrafree MC columns (Millipore, Bedford, MA) and transformed into DH5α cells.

Fluorescence scanning of GFP constructs in yVT12

yVT12 cells containing the appropriate plasmid or library were plated onto selective medium supplemented with 2% dextrose or a combination of 2% galactose with 2% raffinose. Following incubation at 30°C, yeast derived from a single colony (or in the case of a library from a patch of cells) were transferred into selective liquid medium supplemented with the appropriate carbon source. These cultures were grown with shaking at 30°C until mid-log phase. The yeast were pelleted, resuspended in phosphate-buffered saline and scanned on a FACStarPLUS (Becton-Dickinson, San Jose, CA) scanner with excitation at 488 nm. Fluorescence emission was measured with a 515/40 nm band pass filter. Ten thousand events were recorded in each fluorescence scan. All scans were repeated in independently cultured cells in triplicate. Unless stated otherwise, fluorescence scans were performed using yeast grown under inducing conditions (2% galactose, 2% raffinose).

Immunoblot analysis

Protein samples for SDS–PAGE were obtained by extracting total cell protein from 0.5 ml late log culture yeast. Briefly, yeast were pelleted and boiled in the presence of glass beads and 20 μl sample buffer (50 mM Tris–HCl, pH 6.8, 2% SDS, 20% glycerol). The solution was vortexed for 1 min and made up to 100 μl with sample buffer. Following centrifugation the supernatant was kept and protein concentrations determined by the bicinchoninic acid method (Pierce, Rockford, IL). SDS–PAGE was carried out using the Tris buffer system (7), with 10 μg protein loaded per well. Gel transfer was performed using a Genie electrophoretic blotter (Idea Scientific, Minneapolis, MN). Following blotting the membrane was incubated successively with rabbit antisera containing polyclonal anti-GFP antibodies (Clontech, Palo Alto, CA) and peroxidase-conjugated anti-rabbit IgG (Santa Cruz Biotechnology, Santa Cruz, CA) and the bands visualized with the peroxidase substrates diaminobenzadine and hydrogen peroxide.

DNA sequence analysis

Yeast colonies containing library clones were picked from plates and subjected to whole colony PCR using oVT320 and oVT313 primers. These PCR fragments were purified using QiaQuick spin columns (Qiagen, Chatsworth, CA) and their DNA sequences were determined according to standard procedures (8) on an ABI373A automated sequencer (Applied Biosystems Division, Perkin-Elmer Inc., Foster City, CA). The sequences were edited and analyzed using the Genetic Data Environment (9).

P values for the observations of specific amino acids were calculated for each position in the set of 53 peptide sequences in pVTAPT2 and for the total set of amino acids using the formula

$$P' = [N!/k!(N-k)!]\,[f^k\,(t-f)^{N-k}]/t^N$$

where N is the total number of sequences (or total number of residues), k is the number of observations of a specific residue and P' is the probability of observing k residues given that f is the number of codons encoding the residue and t is the total number of codons in the peptide set. This P value (P') was corrected for the number of independent samplings (e.g. 20 amino acids and 15 positions) by applying the formula

$$P = 1 - (1 - P')^{20j}$$

where $j = 1$ for the composition computations and $j = 15$ for the position-by-position calculations.

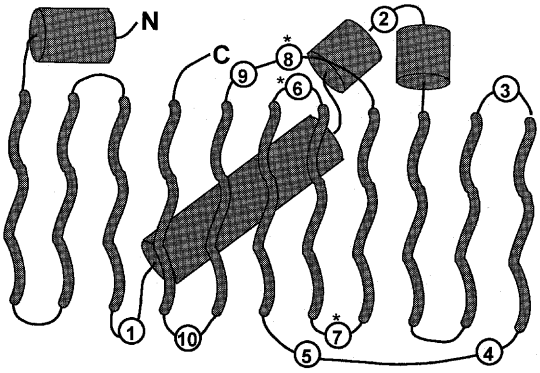
Oligonucleotides

oVT307, TGAGAATTCCTCGAGAGTGCAAATAAATTTAAGGGTAAG;
oVT308, TGAGAATTCCTCGAGCATATGATCTGGGTATCTTGAAA;
oVT309, TGAGAATTCCTCGAGACCTTCAAACTTGACTTCAGC;
oVT310, TGAGAATTCCTCGAGTCCATCTTCTTTAAAATCAATAC;.
oVT311, TGAGAATTCCTCGAGTTTGTGTCCAAGAATGTTTCCATC;
oVT312, TGAGAATTCCTCGAGTTGTTTGTCTGCCATGATGTATAC;
oVT313, TGAGAATTCCTCGAGTTCAATGTTGTGTCTAATTTTGAAG;
oVT314, TGAGAATTCCTCGAGGCCAATTGGAGTATTTTGTTGAT;
oVT315, TGAGAATTCCTCGAGAAGGACAGGGCCATCGCC;
oVT316, TGAGAATTCCTCGAGTTCGTTGGGATCTTTCGAAAG;
oVT317, TGAGAATTCGGATCCACTGGAAAACTACCTGTTCCATGG;
oVT318, TGAGAATTCGGATCCAAACGGCATGACTTTTTCAAGAG;
oVT319, TGAGAATTCGGATCCGATACCCTTGTTAATAGAATCG;
oVT320, TGAGAATTCGGATCCAACATTCTTGGACACAAATTGG;
oVT321, TGAGAATTCGGATCCTTGGAATACAACTATAACTCACAC;
oVT322, TGAGAATTCGGATCCAAGAATGGAATCAAAGTTAACTTC;
oVT323, TGAGAATTCGGATCCGATGGAAGCGTTCAACTAGC;
oVT324, TGAGAATTCGGATCCGATGGCCCTGTCCTTTTACC;
oVT325, TGAGAATTCGGATCCTTACCAGACAACCATTACCTG;
oVT326, TGAGAATTCGGATCCAAGAGAGACCACATGGTCC;
oVT329, GTTAGCTCACTCATTAGGCACCC;
oVT330, CGGTATAGATCTGTATAGTTCATCCATGCCATGTG;
APT1, GGCCTAGGATCC;
APT2, TGACTCGAG[NN(G/C/T)]$_{20}$GGATCCTAGGCC;
APT3, TCGAGAGTGCAGGT[NN(G/C/T)]$_{15}$GGAGCTTCTG;
APT4, ACCTGCACTC; APT5, GATCCAGAAGCTCC.
Restriction sites are underlined.

RESULTS

The first step in the process of creating an autofluorescent protein scaffold suitable for display of peptides involves identification of sites in the protein that accommodate peptide insertions without seriously disturbing protein function. However, it is important that sites be found which not only accept a single small inserted sequence but which also accept a wide variety of different sequences. Such sites are by definition robust to chemical perturbation and satisfy one of the prerequisites for a useful scaffold. Autofluorescent proteins provide a ready assay for determination of appropriate insertion locations. Because the activity of the protein (and perhaps its expression level) can be monitored quantitatively using a flow scanner, it is simple to assay many independent insertions either sequentially or in a bulk population. One simply generates mutant proteins by manipulating the DNA sequence such that a variety of different insertions are produced and examined by flow cytometry. Variants identified in this fashion reveal the nature of sites within the protein suited to display of foreign sequences. In the case of GFP it is likely that few sites other than the two protein termini will accept insertions. This prediction is based on the three-dimensional structure of GFP, which reveals a compact molecule with only a few loops exposed to the solvent (10).

A.



B.

Construct	Insertion Site
pVT22	Thr49-Thr50
pVT23	Met78-Lys79
pVT24	Gly116-Asp117
pVT25	Lys140-Leu141
pVT26	Gly134-Asn135
pVT27	Gln157-Lys158
pVT28	Glu172-Asp173
pVT29	Leu194-Leu195
pVT30	Gly189-Asp190
pVT31	Glu213-Lys214

Figure 1. (**A**) Model of GFP showing sites of peptide insertion. Numbers 1–10 correspond to insertion sites in pVT22–pVT31 respectively. (**B**) Sites of insertion within the *GFP* gene of pVT22–pVT31 of an 18 nt fragment coding for the hexapeptide Leu-Glu-Glu-Phe-Gly-Ser. Amino acid numbering is accordng to the wild-type *GFP* gene. *Sites which were most permissive to insertion of the hexapeptide.

Preparation and testing of GFP scaffold candidates

Using the crystal structure of GFP as a guide, 10 positions on the protein which fall within exposed loops were chosen (10; Fig. 1A). These 10 positions are located inside eight β-turns that connect the β-strands of the GFP protein. Linker sequences composed of recognition sequences for *Bam*HI, *Eco*RI and *Xho*I restriction endonucleases were introduced into the corresponding regions of the *GFP* gene (Fig. 1B). This yielded 10 chimeric *GFP* genes, each of which contained six additional codons that included the restriction

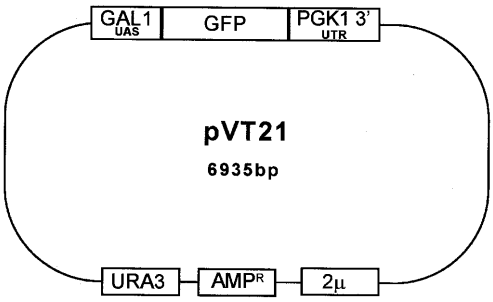


Figure 2. Map of the plasmid pVT21. The GFP gene is flanked by the upstream activation sequence of the GAL1 gene (GAL1 UAS) and the 3′-untranslated region of the phosphoglycerate kinase gene from yeast (PGK1 3′ UTR). The plasmid also contains the URA3 and AMP^r genes for selection in yeast and *E.coli* and 2 micron (2 μC) sequences for replication in yeast.

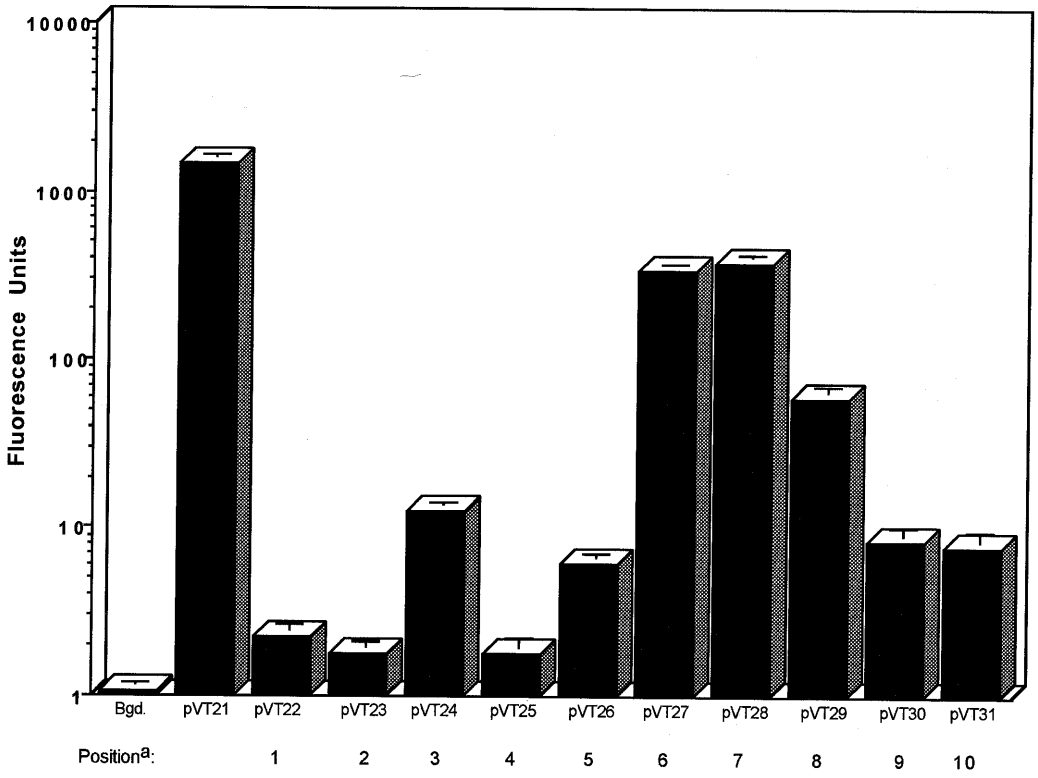


Figure 3. Mean fluorescence intensities of cell populations harboring GFP scaffold candidates and various controls. [a]Position numbers as in Figure 1A.

sites. These chimeric constructs were cloned into a yeast expression vector (pVT21, Fig. 2) that permitted induction of GFP expression in the presence of galactose.

Yeast were transformed with the pVT22–pVT31 plasmids and grown under inducing conditions (i.e. galactose-containing medium) to drive expression of the GFP hybrid proteins and analyzed by flow cytometry to gauge the levels of GFP fluorescence. Of the 10 insert sites examined, three GFP scaffold candidates (pVT27, pVT28 and pVT29) had mean fluorescence intensities in yeast cells which were much greater than 10-fold above the background fluorescence of cells containing the parental pVT21 plasmid under conditions that repress GFP expression (Fig. 3). Two of these candidates (pVT27 and pVT28) had mean fluorescence intensities that were nearly one quarter the value of wild-type GFP expressed from pVT21 under inducing conditions. Thus pVT27, pVT28 and pVT29 were considered candidates for encoding scaffolds based on fluorescence of their products.

Preparation of peptide display libraries

To test the ability of pVT27–pVT29 to accommodate a variety of sequences, DNA oligonucleotides coding for random peptides 20 amino acids long were synthesized and inserted into the *Xho*I and *Bam*HI sites of the three significantly fluorescent GFP scaffold candidates. These oligonucleotides consisted of *Bam*HI and *Xho*I sites flanking 60 bases of biased random sequence (see Materials and Methods). GFP–peptide libraries (designated by the suffix APT) in each of the three scaffold candidates were created by transformation and amplification in *E.coli*. A total of ~2000 individual clones were selected from each library for testing purposes. For each set of scaffold candidates, 20 random clones were examined to determine the percentage of insert-bearing clones. All three had insert frequencies of at least 90% (data not shown).

Evaluation of peptide display libraries

The amplified libraries from *E.coli* were transferred into yeast cells and the individual yeast libraries were grown separately under inducing conditions. Yeast cells from each (2000 member) library were examined by fluorescence microscopy. Though the absolute fluorescence levels of different cells varied, the fluorescence appeared to be uniformly distributed throughout the cytoplasm of the cells (excluding vacuoles and nuclei), not concentrated in clumps or subcellular compartments (data not shown). This result suggested that the GFP–peptide hybrid proteins were soluble in yeast.

To determine which of the three sites within GFP can best accommodate peptides comprising 20 residues of diverse sequence, fluorescence scans on a flow cytometer were carried out. Mean fluorescence intensities and the fraction of cells in specific fluorescence intensity windows were determined for yeast cell populations containing the libraries (see Table 1). The results suggested that two candidates (pVT27 and pVT28) provided a site suitable for library expression using GFP as a scaffold, according to the method of scaffold design pursued in these experiments. The other scaffold–peptide library (pVT29APT) had a significantly lower mean fluorescence intensity that was close to the background level. The library species in pVT27APT and pVT28APT each generated a collective mean fluorescence intensity that was ~10% of the construct containing the linker sequence alone. A fluorescence window was set to determine whether pVT27APT and pVT28APT clones generally produced low fluorescence intensities or whether

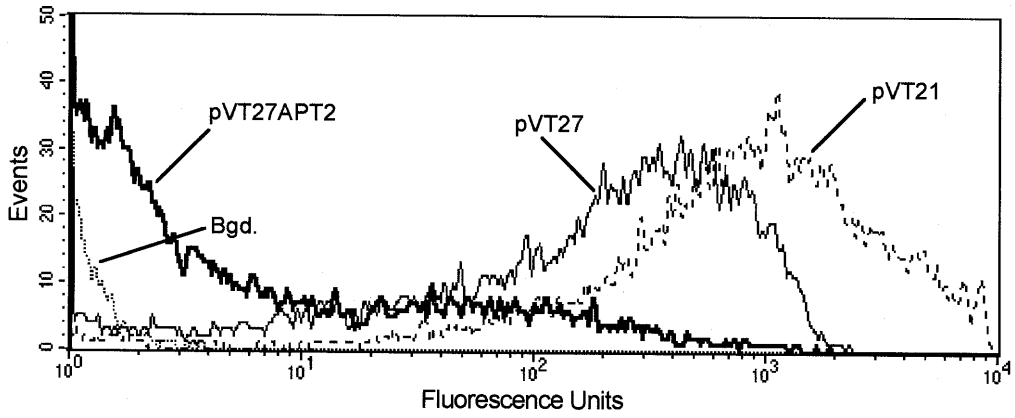


Figure 4. Fluorescence intensity scan of pVT21, pVT27 and pVT27APT2. Bgd., pVT21-containing yeast grown under repressing conditions (dextrose).

there was a wide range of intensities. At an intensity cut-off 10-fold above the background (cells without GFP), where 95% of the control GFP-expressing yeast (with pVT21) were above the threshold, nearly 15% of the pVT27APT- and pVT28APT-containing cells were also positive. This suggested that: (i) the pVT27APT and pVT28APT clones encoded proteins that were for the most part either expressed at lower levels than wild-type GFP produced by pVT21 or were less fluorescent, and (ii) there was significant variability in fluorescence among the individual library clones.

We chose pVT27 as a scaffold candidate to build a large GFP–peptide library. To facilitate this an oligonucleotide coding for a biased random 15 amino acid peptide (flanked by three constant amino acids on either end) was synthesized and cloned into pVT27 (see Materials and Methods). The resulting library contained 1.5×10^6 members and was designated pVT27APT2. A portion of yeast harboring pVT27APT2 clones did not fluoresce when grown under inducing conditions (Fig. 4). These dim yeast may have lacked fluorescence due to termination codons in the random peptide, improper folding of the full-length chimeric GFP protein, or for other reasons. Based on the biased random DNA sequence encoding the peptides, 27% of library members were expected to contain termination codons by chance, resulting in a truncated and non-fluorescent GFP protein (11). From the fluorescence intensity profiles, we estimated that ~60% of the library sequences produced non-fluorescent proteins. The difference (60% – 27%) may reflect the proportion of incorrectly folded and/or unstable GFP proteins in the library. In addition, it is possible that some of the dim library members result from chimeric GFPs in which fluorophore formation occurs at a slower rate than in wild-type GFP. Indeed, when the dim 60% yeast fraction was sorted from the pVT27APT2 library and grown under inducing conditions for a longer period (two additional doublings) a portion of the population became more fluorescent (data not shown).

To further explore the question of the folded state of GFP–peptide molecules produced by the pVT27APT2 library, the fluorescence properties of 10 individual clones were examined in detail. These yeast were obtained by collecting a subpopulation of the pVT27APT2 yeast library which was fluorescent at a level above that of uninduced cells. The sorted yeast clones were grown under inducing conditions and fluorescence emission spectra at 515 nm were measured. Wild-type GFP protein has excitation and emission maxima at 395 and 509 nm respectively. pVT21 and its derivatives produce a red shifted GFP variant (12) (S65T) which has an excitation maximum at 490 nm but also emits at 509 nm. Fluorescence analysis of these 10 clones with excitation at 488 nm revealed a broad distribution of mean fluorescent values (Fig. 5A). This result may be indicative of different GFP–peptide protein levels and/or of GFP–peptides in which the fluorophore is structurally perturbed.

Table 1. Mean fluorescence intensities of cell populations harboring pVT27APT, pVT28APT, pVT29APT and parent constructs

GFP construct	Fluorescence > 1× Bgd[a]		Fluorescence > 10× Bgd	
	Percent total population	Mean (FU[b])	Percent total population	Mean (FU)
pVT21 (Dex.)	1	3	0	
pVT21	96	1545	95	1565
pVT27	89	378	81	414
pVT27APT	39	41	15	99
pVT28	86	428	78	471
pVT28APT	42	28	13	78
pVT29	77	71	59	90
pVT29APT	32	7	2	37

Fluorescence gates were set either at background (Bgd) or at a value 10-fold higher than background (10 × Bgd).
[a]Background is defined as the minimum fluorescence intensity value which is larger than the fluorescence value of 99% of non-induced cells.
[b]FU, fluorescence units. Note FU are arbitrary measures of fluorescence and cannot be compared between experiments (e.g. between Figs 3 and 5).

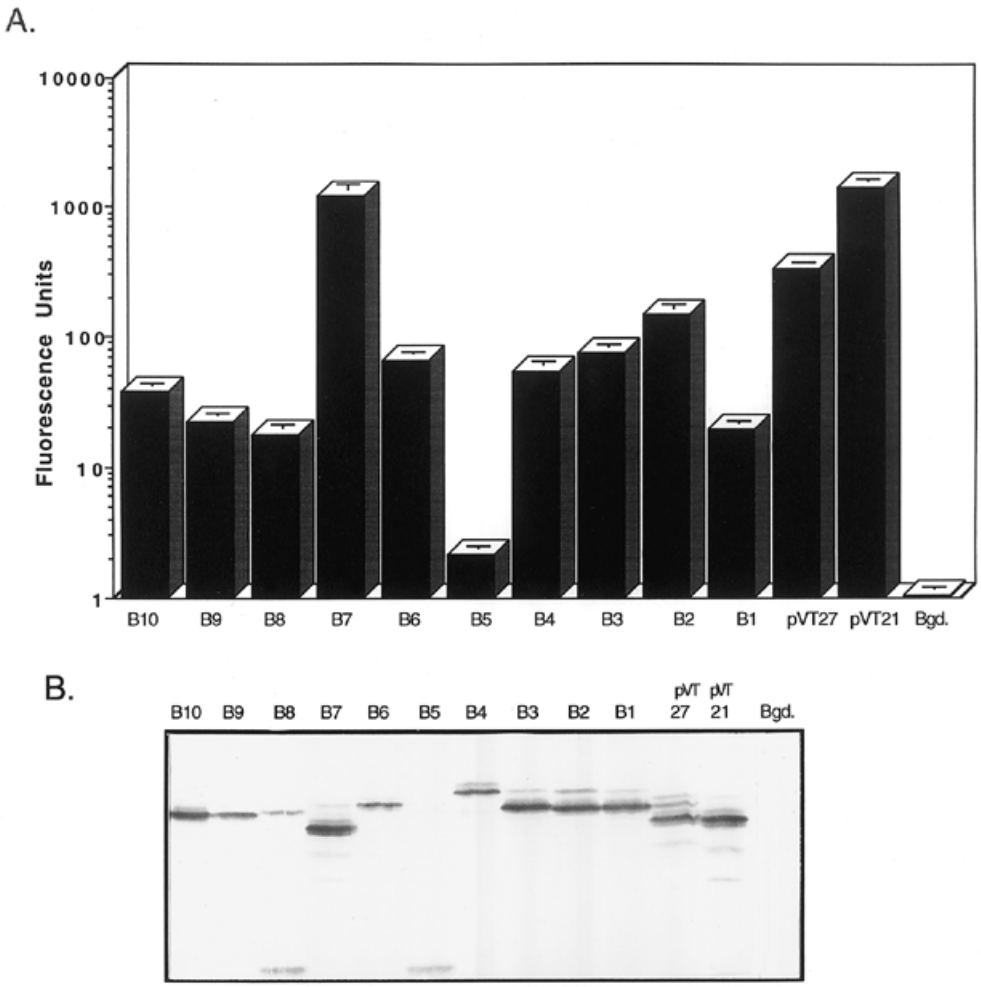


Figure 5. (**A**) Mean fluorescence intensities of 10 sorted pVT27APT2 yeast clones (B1–B10). (**B**) Western blot analysis of chimeric GFP proteins from 10 pVT27APT2 yeast clones. Aliquots of 10 µg total cell protein were loaded per well. The variability and band position may reflect differences in sequence composition among proteins encoded by different clones because none of the clones had multiple inserts or differences in insert sequence length.

A Western blot of proteins extracted from yeast cells harboring these 10 clones was also prepared to provide an estimate of GFP–peptide levels in these cells (Fig. 5B). There was a rough correlation between expression and fluorescence levels. For example, clone B5 produced the least fluorescence of any of the 10 clones examined (see Fig. 5A), >100-fold below the parental pVT27 construct. The protein level revealed by Western blot analysis was also the lowest of the set (Fig. 5B). Clones B5 and B8 (also associated with low fluorescence) produced a small fragment near the bottom of the gel that stained with the anti-GFP antibody that may represent a degradation product.

The possibility of bias in the peptide sequences capable of display by the pVT27 GFP scaffold was examined by sequence analysis of 53 independent clones from the pVT27APT2 library (Table 2). These clones were selected from the subset of pVT27APT2 sequences that generate fluorescent proteins by selection using a flow sorter. Analysis of the amino acid distribution of these peptides revealed some statistically significant bias. Glycine, lysine and threonine were over-represented, compared with their expected frequency of occurrence, while leucine and glutamate were under-represented. Glycine was one of the most dramatic outliers and this may reflect a preference for small, flexible residues in protein loops (13). Indeed, overabundance of glycine at position 12 in the peptide was the only statistically significant difference ($P < 0.005$) observed when analysis was performed position by position on the 15 residue peptide sequence (data not shown). However, analysis of the base composition of these 53 sequences revealed that the third codon position consisted of 40% guanine residues instead of the expected 33% (see Materials and Methods). This suggested that guanine was over-represented in the synthesis and, because glycine is encoded by GGN, may explain at least part of the bias toward glycine observed (data not shown). The reasons for over-representation of the other dramatic outlier in the analysis of overall amino acid composition, leucine, are unclear.

DISCUSSION

Peptide-based ligands are useful in a variety of contexts as probes of biological functions or as aids in the development of therapeutic compounds. Many techniques have been devised to isolate specific peptides from complex libraries which bind to

defined targets *in vitro*. Less attention has been focused on intracellular display of peptide or polypeptide libraries. The single exception is the yeast two hybrid system, a technique that involves expression in yeast cells of fusion protein libraries to identify proteins that interact with one another *in vivo* (4). A variant of the yeast two hybrid method has been developed in which the protein library consists of peptides of random sequence displayed on the surface of a thioredoxin-based scaffold (14). This approach permits isolation of peptides displayed on thioredoxin which bind specific probe sequences in yeast. For example, peptides capable of binding to and inhibiting the function of a cell cycle regulator (cyclin-dependent kinase 2) *in vitro* were identified through the application of this system (15). In principle it is also possible to use such peptide expression libraries to perform more traditional, broad-based genetic screens in which the expression library serves as a source of trans-dominant peptides that disrupt specific biochemical interactions in cells, thus conferring a phenocopy state on host cells within which they are expressed (Caponigro, in preparation).

Table 2. Statistical analysis of the amino acid composition of 53 random peptides selected from bright yeast

Amino acid	Expected no.	Observed no.	Observed/expected	*P*
Ala	48.7	46	0.95	0.68
Arg	64.9	66	1.09	0.18
Asn	32.5	34	1.05	0.75
Asp	32.5	36	1.11	0.68
Cys	32.5	28	1.11	0.68
Gln	16.2	15	0.93	0.87
Glu	16.2	28	0.86	0.04
Gly	48.7	92	1.89	<<0.01
His	32.5	24	0.74	0.38
Ile	32.5	40	1.23	0.43
Leu	64.9	8	0.12	<<0.01
Lys	16.2	33	2.03	<0.01
Met	16.2	27	1.67	0.07
Phe	32.5	25	0.77	0.46
Pro	48.7	43	0.88	0.59
Ser	81.3	66	0.81	0.18
Thr	48.7	69	1.42	0.02
Trp	16.2	27	1.67	0.07
Tyr	32.5	20	0.62	0.10
Val	48.7	52	1.07	0.65

A useful scaffold must effectively display a range of peptides *in vivo*. Thus it must be relatively resistant to degradation by proteases within the cell and it must be soluble, even when joined to a wide variety of foreign peptide sequences. It is helpful if the scaffold has other properties, for example, the peptide should be exposed on the protein scaffold in a conformationally constrained configuration (16). In addition, judicious choice of scaffold protein may allow the quantitative performance of the scaffold to be measured by its ability to display peptides and maintain high level expression in cells.

The recently solved crystal structure of GFP reveals that this protein assumes a barrel-like structure and has 10 solvent-accessible loops, two of which connect the helical chromophore segment to the rest of the protein. The remaining eight loops connect the β-strands of the barrel to one another. These loops are candidate sites for insertion of random peptides. By inserting peptides into the β-turns in GFP, loops can be identified by flow cytometry which accommodate random peptides while allowing GFP to retain fluorescence. Although GFP is known to accept N- and C-terminal fusions, there are two reasons for preferring internal sites for peptide display. First, conformational freedom is reduced by tethering the two ends of the peptide to rigid components of the structure; for peptides located at the protein termini it is only possible to tether one end. Second, peptides at either terminus will be charged, which limits the range of chemical/structural possibilities encompassed by the library. In the analysis reported here two loops were identified that could serve as sites for peptide insertion. As a natural peptide scaffold immunoglobulins (Igs) provide a useful analogy. The tertiary structure of the variable domain of an Ig subunit consists of a β-barrel together with two exposed loops that form hypervariable regions (17). These loops comprise the antigen binding site of Ig and can accommodate a vast number of different sequences. Presumably the stability of the β-barrel structure facilitates presentation of exposed loops such that the peptide backbones in different sequences assume rigid conformations. Thus a functional GFP scaffold might be thought of as a fluorescent Ig molecule.

One of the principal advantages of an autofluorescent scaffold is that once suitable sites for insertion in the protein are discovered, it is possible to quantitatively monitor the behavior of the individual scaffolds that are chosen. A flow sorter again serves as the appropriate tool for such analysis. A set of peptides is inserted into the scaffold at a predetermined position and the fluorescence properties of the ensuing expression library are examined. Quantitative values such as mean fluorescence intensity can be determined from the fluorescence intensity profile of the library population. This permits an estimate of the percentage of library sequences that do not lend themselves to expression in this context and, hence, an estimate of library complexity. In the one insertion library that we examined nearly half of the chimeric GFP clones produced fluorescent protein (excepting those predicted to have stop codons).

A flow sorter can be used not only as a screen to examine the properties of the generated expression libraries but also as a tool to manipulate and bias the libraries in potentially useful ways. For example, in certain cases it may be helpful to select from the expression library those sequences that express the highest levels of fusion protein in cells. Alternatively, it may be desirable simply to exclude all library constructs that do not express scaffold levels above background; many of these negative or 'dim' cells may harbor expression constructs that produce truncated or misfolded proteins that are degraded or do not function as soluble peptide display scaffolds. Typically libraries of more than a few million clones are difficult to construct and screen *in vivo*. Thus in some cases a premium may be placed on ensuring that the maximum number of library sequences express stable proteins. A flow sorter permits such selections to be carried out with extraordinary efficiency because cells can be sorted at a rate of 10 000 000–100 000 000 cells/h (18).

In the experiments described here only two sites in GFP (apart from the N- and C-termini) were found to display a variety of peptides in a manner compatible with autofluorescence. One of these sites (corresponding to pVT27) is located within one of the smaller loops of the protein (Ala154–Gly160). It is noteworthy that main chain atoms in this loop have the highest temperature factors of any backbone atoms in the structure, as high as the solvent-exposed N-terminus (10). This suggests that the insertion

site is more mobile than other loops and, as such, may not be an integral part of the structure. It is curious that GFP is so sensitive to structural perturbations, even in β-turns. The sites within GFP found to best accommodate peptide insertions (e.g. the site used to construct pVT27APT2) do not generate chimeric GFP with fluorescence levels as high as the wild-type. Presumably this extreme sensitivity betrays highly tuned chromophore chemistry and structural parsimony owing to rigorous evolutionary pressures on the protein and the organism (*Aequorea victoria*) that employs it.

The cause of the lower mean fluorescence intensity in pVT27APT2-containing yeast is not certain. Based on spectral analysis of a subset of pVT27APT2 clones, it seems unlikely that the differences in the majority of clones can be explained by shifts in the excitation and/or emission maxima of the chimeric proteins (Abedi, unpublished observations). More likely are the possibilities that quantum yield and/or protein levels are primarily responsible for the differences. A Western blot demonstrated low protein levels in the least fluorescent of the 'bright' selected clones. However, it did not explain why GFP insertion chimeras are significantly less fluorescent on average than wild-type GFP. Further biophysical experiments are necessary to test if, for example, the chimeric proteins undergo intramolecular cyclization to produce active fluorophore more slowly than the wild-type protein (19). It is also unclear why a significant fraction of the library (beyond the fraction expected to contain stop codons in the peptide encoding sequence) is non-fluorescent as judged by comparison with yeast that do not express GFP. It is possible that some of these dim clones are composed of inserts shorter than unit length and out of frame, because a few instances of deletions were observed during sequencing efforts (Abedi, unpublished observations). However, most of these dim clones may encode proteins that are degraded by cellular proteases, are incapable of folding, or fold into a conformation that does not permit autofluorescence. Some of these non-fluorescent chimeras may have useful aptamer properties in specific contexts, assuming they are not entirely absent within the cell.

Statistical analysis of sequences obtained from randomly selected bright clones suggests some modest skewing in representation of specific amino acids, notably glycine and leucine. However, it seems unlikely that there is a dramatic bias in the structural chemical properties encompassed by the peptide library in terms of charge or hydrophobicity, because no systematic preference for or avoidance of residues of specific chemical types was observed. Thus the GFP insertion library described here likely represents a relatively unbiased collection of peptide entities when expressed in yeast cells.

The experiments reported here utilized *Saccharomyces cerevisiae*. However, there is no reason to believe that other cell types, including mammalian cells, would not be suitable for evaluation of GFP–peptide expression libraries. Once the library was constructed in *E.coli* it could be transferred into mammalian cells and bright cells could be collected on a flow sorter as the first step in enrichment of the library for expressing clones. Alternatively, it may be preferable to characterize the library and, if desired, select a subset of library sequences by analysis in *E.coli* or yeast prior to introduction into mammalian cells. The assumption is that sequences that encode proteins with low autofluorescence are either unstable or misfolded. These sequences can be winnowed out in microbial cells more easily than in mammalian cells.

ACKNOWLEDGEMENTS

We thank Dr J.Rine for providing yeast strains and plasmids, W.Judd for assistance with DNA sequencing and C.Wang for computer programming assistance.

REFERENCES

1 Cwirla,S.E., Peters,E.A., Barrett,R.W. and Dower,W.J. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6378–6382.
2 Cortese,R., Monaci,P., Nicosia,A., Felici,F., Galfre',A., Tramontano,A. and Sollasso,M. (1995) *Curr. Opin. Biotechnol*., **6**, 73–80.
3 Fields,S. and Song,O. (1989) *Nature*, **340**, 245–247.
4 Chien,C.T., Bartel,P.L., Sternglanz,R. and Fields,S. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 9578–9582.
5 Boyartchuk,V.L., Ashby,M.N. and Rine,J. (1997) *Science*, **275**, 1796–1800.
6 Gietz,R.D. and Schiestl,R.H. (1995) *Methods Mol. Cell. Biol*., **5**, 255–269.
7 Laemmli,U. (1970) *Nature*, **277**, 680–685.
8 Sambrook,J., Fritsch,E. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
9 Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P.M. (1994) *Comput. Appl. Biosci*., **10**, 671–675.
10 Ormo,M., Cubitt,A.B., Kallio,K.K., Gross,L.A., Tsien,R.Y. and Remington,S.J. (1996) *Science*, **273**, 1392–1395.
11 Dopf,J. and Horiagon,T.M. (1996) *Gene*, **173**, 39–44.
12 Heim,R., Cubitt,A.B. and Tsien,R.Y. (1995) *Nature*, **373**, 663–664.
13 Edwards,M.S., Sternberg,J.E. and Thornton,J.M. (1987) *Protein Engng*, **1**, 173–181.
14 LaVallie,E.R., DiBlasio,E.A., Kovacic,S., Grant,K.L., Schendel,P.F. and McCoy,J.M. (1993) *Biotechnology*, **11**, 187–193.
15 Colas,P., Cohen,B., Jessen,T., Grishina,I., McCoy,J. and Brent,R. (1996) *Nature*, **380**, 548–550.
16 Ladner,R. (1995) *Trends Biotechnol*., **13**, 426–430.
17 Edmundson,A., Ely,K., Schiffer,M. and Panagiotopolous,N. (1975) *Biochemistry*, **14**, 3953–3961.
18 Shapiro,H. (1995) In *Practical Flow Cytometry*. Wiley-Liss, New York, NY, pp. 217–228.
19 Heim,R., Prasher,D.C. and Tsien,R.Y. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 12501–12504.