Functional genomics, an integrated approach

Final report for the Trieste component

For period July 15 1998 - July 14 1999

Andrew Bradbury

Los Alamos National Laboratory, Los Alamos, New Mexico

And

International School for Advanced Studies (SISSA), Via Beirut 2-4, Trieste, Italy

November 1999

Prepared for the US Dept. of Energy, award number: DE-FG02-98ER62647

DOE Patent Clearance Granted
MP Dvorscak
Mark P. Dvorscak
(630) 252-2393
E-mail: mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office
July 27, 2001
Date

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Phage display offers the possibility of selecting polypeptides (and the genes which encode them) from libraries of 10^{10} or more different polypeptides on the basis of their abilities to bind target proteins and subdomains. This diversity far surpasses the estimated number of total genes in the human genome. The application of this technology to the Human Genome Project will powerfully accomplish a central goal: the derivation of ligands that recognize the protein products of all human genes, such ligands being either antibodies, or protein fragments.

The general principle of phage display relies on the coupling of phenotype and genotype in the phage (a protein displayed on the surface of the phage is encoded by a gene found within the phage). The selection of a phage on the basis of its binding properties will lead to the simultaneous isolation and cloning of the gene responsible for those binding properties.

Where the recognition ligands derived from this relatively new technology are antibody binding regions (single chain Fv - scFv are usually used) they can be used in the same way as traditional antibodies. As such, they can play essential roles in assigning gene function, by characterising spatiotemporal patterns of protein expression, elucidating potential protein-protein interactions, and aiding in protein purification. Where the recognition ligands are protein fragments, they can be considered to be potential protein-interaction partners for the immobilised polypeptide and a starting point for further biochemical studies. Furthermore, recognition ligands with inhibitory activities will facilitate the elucidation of protein function(s), both *in vitro* and *in vivo*.

Of relevance, this technology is readily amenable to automation and allows the application of high throughput methods, a situation which is very desirable in the Human Genome Project where thousands of genes must be considered in parallel rather than just a few, as occurs in most scientific research.

Specific Aims of this pilot project:

1. To produce a large phagemid antibody library.
2. To develop a general method for deriving antibodies against the protein products of cloned genes.
3. To develop technologies which permit the display of gene fragment libraries on phage.

The work described in this report was carried out under my direction in both Trieste (International School of Advanced Studies) and the Los Alamos National Laboratory, with the bias of the work carried out as follows:

large phagemid antibody library creation - Trieste

filter vector - Trieste

peptide selections - LANL

Creating a large phagemid antibody library

The use of phage antibody libraries has been far less extensive than would have been expected from the publicity they have received, and has led to the belief, in some quarters, that the technology does not work. It is true to say that the power of the technology was probably overstated initially, and it is only recently that some of the many problems are being resolved. Even now, good antibody phage libraries can be found in only two laboratories in the US, and no more than seven worldwide. The technology works best in those laboratories, and the experience of libraries sent out to inexperienced laboratories remains bad. This is likely to be due to a number of factors, including the difficulties in using the libraries and the low quality of the libraries sent out (probably due to a slow proteolysis of the displayed antibodies occurring upon prolonged storage, a problem which needs to be resolved). Furthermore, such libraries are available for the selection of antibodies against individual specified antigens only, and not for all antigens as would be required for a genome project.

Previous Accomplishments:

Before the start of this project, we had developed a new phage display vector (pDAN5) incorporating a number of design improvements, including a new antibody tag (SV5), a his tag and the use of restriction sites which are not present in human V regions. These are however, relatively minor improvements. In addition, two lox recombination signals, loxP wild type and loxP 511, were included in the plasmid. One of these (loxP511) was included within the scFv linker, while the other was at the 3' end of gene 3.

Results achieved:

The goal of creating this plasmid was to develop an in vivo recombination system which would allow the creation of far larger libraries. Although this was not funded by this project, it is directly relevant to it, and will prove invaluable in the future prosecution of this project. The first step was the validation of the vector described, and in particular, the use of the lox site as a protein linker, as opposed to a DNA signal, between the heavy and light chain variable genes. The efficiency of this was tested in a number of ways, including the cloning and expression of a number of monoclonal antibodies into the pDAN5 vector and the creation of small libraries using this linker and others to test the levels of display. In all tests, the lox linker appeared to function as well as any of the other linkers previously used.

A test of the capacity of this vector to mediate in vivo recombination was carried out, by creating two scFvs each of which contained either the heavy chain gene, or the light chain gene, of a functional scFv (D1.3, which recognises lysozyme) in association with irrelevant partner chains, X and Y (VL/X-VH/D1.3 and VL/D1.3-VH/Y). Phagemid containing these scFv genes were allowed to infect *E. coli* expressing cre recombinase at a multiplicity of infection of 20:1. After successful recombination, each bacteria should contain four different scFv genes (VL/D1.3-VH/D1.3; VL/D1.3-VH/Y; VL/X-VH/D1.3 and VL/X-VH/Y). Phagemid were rescued from such bacteria and coupled to the appropriate scFv protein by passage through normal *E. coli* infected at multiplicity of infection of one. When infection was carried out in bacteria expressing cre recombinase, recombination was demonstrated in 25% of phagemid by PCR and 17% by ELISA, whereas in bacteria not expressing cre, no recombination was found. These results indicated that recombination induced by cre recombinase could be used to shuffle heavy and light chain genes between different phagemids and so give rise to new specificities. Furthermore, the recombination reaction appeared to go to equilibrium, since VL/D1.3-VH/D1.3 comprised approximately 25% of the total.

In order to make a large scFv library, a small primary with a diversity of 7×10^7 was created in the pDAN5 vector. In the library, each heavy chain V gene is associated with a different light chain V gene. The total potential diversity which can be created from such a primary library is the product of the number of heavy and light chains, approximately 5×10^{15} in the case of this library. Many of the most recently published libraries have harvested this potential diversity by cloning. However, this is a relatively inefficient procedure, and no library larger than 10^{10} has been described. A far more efficient means of coupling the different heavy and light chain combinations is to use the in vivo recombination method developed. This primary library was used to infect cre recombinase expressing bacteria at a multiplicity of infection of 200:1. By sequencing the different V genes found within a single bacteria, it was found that at least nineteen different phagemid can enter each bacterium, and extensive shuffling occurs between the light and heavy chains by virtue of recombination induced at the lox sites. By exploiting this method, a library, with a nominal diversity of 10^{11} , was made. This has been used to derive antibodies against at least sixty five different antigens. The affinities of the antibodies derived from this library were found to be equivalent to those selected from the libraries created using cloning.

A paper describing this work is in press in *Nature Biotechnology*.

The primary library described above was made using PCR assembly to couple VH and VL chains. This can suffer from problems due to loss of diversity caused by the preferential amplification of a small subset of V regions. As a result the true diversity may be lower than that described above. Initially, it was planned to make another library using cloning rather than PCR assembly. However, this has proved to have taken longer than expected. Furthermore, the library made using PCR assembly functions very well after recombination. The main problem with this, and all libraries made by cloning, has been the expression levels of the antibodies in the library, which do not usually exceed 1 mg / L . It is planned to make a new library, which will overcome this problem by preselecting antibodies for high expression prior to recombination. This is likely to be very important in the application of this technology to the human genome project, where high expression levels in the microtitre format are essential.

The development of a general method to derive antibodies that recognise the protein products of cloned genes

A prerequisite for the derivation of antibodies against the protein products of all human genes is access to selectors which correspond to those genes. A number of different methods can be foreseen to derive such selectors:

- 1 Proteins produced from full length cDNAs encoding the full length open reading frames (ORFs) in recombinant hosts.
 - 2 Protein fragments produced from fragments of ORF cDNAs in recombinant hosts. These may be either defined fragments, such as domains or functional units, or random fragments created by fragmentation of full length ORF cDNAs.
 - 3 In vitro translated proteins.
- All the above can be considered to be cDNA based selection methods.
- 4 Synthetic peptides derived directly from gene sequences.
 - 5 Biochemical fractions derived from natural complex sources (cell or tissue extracts, serum etc.) - proteome based selection
 - 6 Selection by infection strategies, in which bacterial infection relies upon a specific antibody-antigen interaction, where the antibody is displayed on the phage and the antigen on the surface of bacteria.

Each of these methods have their advantages and disadvantages. In general, the use of polypeptide selectors derived from cDNAs have the advantage that antibodies recognising conformational epitopes can be selected. Such selectors are closest to the natural form of the antigen, although they lack any post-translational modification. They also have no problem with gene identification, the gene encoding the selector is known in advance. The same is true of synthetic peptides. These, however, have the disadvantage that many peptides are known not to be immunogenic. In fact, much work has been carried out in the past in an attempt to predict which peptides can act as good surrogate immunogens for native proteins. The rules governing peptide 'immunogenicity' are likely to be different for animals with complete immune systems, when compared to in vitro systems such as phage display.

The best antigen source is likely to be biochemical fractions derived from complex natural source. These will contain all the antigenic features of the original, including post-translational modifications, and processing events, however, these suffer from two big disadvantages: after selection, the antigen used must be identified, and an appropriate tissue source for all antigens must be found and appropriately fractionated to give high enough concentrations to allow selection to occur.

Of the methods described above, methods 1, 2 and 4 have been attempted in this project.

Results achieved:

Selecting against protein antigens

A number of proteins produced in recombinant hosts were used to select antibodies using two phage antibody libraries (pDAN and one supplied by Prof. J. Marks, UCSF). Selection was carried out in two different ways. In the first method, which can be considered to be the 'classic' method, proteins are adsorbed onto the surface of an immunotube (a 4ml tube made of Maxisorb, a specially treated polystyrene with high capacity for protein adsorption, produced by Nunc). Selection is then carried out by adding the library to the tube containing the protein selector, incubating, washing and eluting. This is repeated two to three times. Using this method antibodies against 19 different proteins, comprising 187 different monoclonal scFvs, were selected. One of the proteins was ubiquitin, against which it has proved extremely difficult to derive antibodies by immunisation. 13 different scFvs were selected against ubiquitin and some of these were shown to recognise native ubiquitin in a number of different assays. Antibodies selected using this method were shown to be functional in ELISA (this is the screening method used), Western blotting, functional enzyme inactivation and functional inhibition of binding. Affinities were measured for four antibodies from the pDAN library. These were found to be between 15 and 82nM, values similar to those obtained for traditional monoclonal antibodies. In general, antibodies selected using this method appear to have the same advantages and disadvantages as traditional monoclonal antibodies - different antibodies are functional in different assays. Some are good Western blot reagents, others will inhibit function, some will only recognise the native protein. Initially it was hoped that polyclonal populations of antibodies could be stored and used, so avoiding the problems associated with the identification of individual monoclonals. However, it was found that polyclonal populations of antibodies are relatively unstable when no selective pressure is applied, and it appears that a

single cycle of growth after freezing is sufficient for the culture to be overtaken by non-expressing antibodies. For this reason, individual monoclonals need to be identified for each antigen..

As this method is not particularly amenable to automation, a microtitre based method was also developed, using polystyrene pins. Each pin is able to adsorb up to 300ng of protein. Pilot experiments were carried out using 12 different proteins bound to 12 pins selected on the two libraries available. It was found that antibodies were selected against 10/12 proteins with the JM library and 11/12 with the pDAN library, with the missing proteins being different for the two libraries. These experiments were extended to a total of 45 different antigens, with antibodies against all antigens being derived in at least one selection experiment. Present experiments are underway to address the question of how many different monoclonal scFvs can be derived using this method, whether the monoclonals derived have the same fingerprints as those derived by classic selection (and so are the same) and whether they are functional in the same assays.

The ability to select antibodies against protein antigens using the microtitre format lays open the way to high throughput automation of the procedure, once a method to produce large numbers of protein selectors has been developed, one of the limiting factors in this project. The amount of protein required to perform selection and subsequent ELISA screening is currently less than 100ug, but can probably be reduced further.

Selecting against peptides

One problem with the use of recombinant proteins, is the need to individually clone, express and purify the individual proteins. This represents a highly labour intensive task which could be avoided by the use of synthetic peptides derived from open reading frame (ORF) sequences. This would have a number of advantages: cloning and protein purification could be avoided and suitable peptides could be identified directly from the sequence, without human input (once appropriate selection criteria have been developed). In order to determine whether peptides could be used to select phage antibodies, a total of 188 overlapping scanning peptides were synthesised for five different proteins (albumin, ubiquitin, Chk1, Cdk2 and transglutaminase) on two microtitre trays. These were synthesised bound to pins at their C terminus (Chiron mimeotopes) and used to select phage antibodies. The advantage of the pin format, is the ease of use in selection. Three to four rounds of selection into individual wells were performed for each peptide and the final phage were tested for their ability to recognise the peptide which was used for selection as well as the native protein from which the peptide was derived. Some peptides (5-30% of peptides tested) were able to select antibodies which could recognise the full length protein.

A numerical value was given to each peptide on the basis of a number of biophysical and predictive properties. All the scanned peptides for each protein were then ranked for each of the different properties, and the position at which the first two peptides which gave rise to antibodies which recognised the native protein were determined. On this basis, solvent accessibility appears to be the best predictor of peptide "antigenicity" in phage antibodies. It is important to note that this method attempts to identify any peptide which can be considered antigenic, and the best peptides were not necessarily identified.

Using the algorithm developed, peptides were predicted for two different proteins (zABC1, p53 and delta subunit of the GABA receptor). Peptides for the three or four most solvent exposed peptides were synthesised and used to select antibodies. In each case, antibodies were selected which were able to recognise the full length protein (in either ELISA or Western blot) from which the peptide was derived. In one case, zABC1 (a transcription factor over-expressed in breast cancer), it has proved impossible to derive antibodies using traditional approaches, and the antibodies selected against the peptides are presently being used in collaboration with Joe Gray in UCSF to assess the levels of expression of this protein in different breast cancers.

In addition to using the internal pin bound peptide approach, C terminal peptides, which have been traditionally used to derive polyclonal antisera in rabbits, were also used. These were biotinylated and selection was carried out in the soluble phase, with capture of peptide and attached phage antibody occurring using magnetic streptavidin and avidin beads. X different proteins () were tested, and antibodies which were able to recognise the full length protein in ELISA were derived for Y of these. The antibodies derived against Ku80 were examined in greater depth. It was shown that these antibodies could recognise Ku80 both in Western blots as well as in preliminary immunoprecipitation experiments, appearing to co-immunoprecipitate Ku70 as well. If further extended, these results indicate the enormous potential of this technique in the analysis of multi-protein complexes using gene based phage antibodies for immunoprecipitation.

These studies are very promising for the application of this technology to the selection of an enormous panel of monoclonal antibodies against all gene products. However, certain caveats apply: in one case, antibodies which recognised Ku80 were also shown to recognise myosin, this probably being due to the identity of 5 out of 15 amino acids in the peptide used. It will be impossible to filter out this degree of similarity in advance. As a result, verification of specificity becomes an important part of the use of such antibodies, and it is recommended that at least four different antibodies directed against at least two of each of two peptides are used to provide verification by overlapping patterns of reactivity. This requires a great deal of work once antibodies are selected. This in combination with the fact that the affinity of the antibodies (although not yet formally tested) is likely to be lower than those selected against real proteins or protein domains, indicates that it is probably better to address a high throughput protein expression system, to provide antigen sources, rather than peptides. In this case, there is more work upstream, but the quality of the antibodies derived is likely to be higher, and less work will be required for antibody verification.

Gene fragment libraries on phage

The yeast two hybrid system has been very successful in the identification of potential protein-protein interactions. However, while it is relatively easy to use for the study of the interactions of a single gene product, the number of transformations required would make it extremely difficult to apply to 100,000 different genes on a genome-wide scale. Furthermore, selection occurs intracellularly, so it may not be feasible for detecting interactions between secreted or membrane proteins.

An alternative method is the use of phage display. The use of this technique to study protein-protein interactions are just beginning. It has the advantage that one could create a single general library of open reading frames or open reading frame fragments which could then be used to screen any potential interaction: other proteins, small molecules or pharmaceuticals.

The combination of all fragment libraries of all genes would represent the complete coding potential of the genome: a complete genomic cDNA display library. Such a display library would not need to be very large (10^3 - 4 fragments from 10^5 genes would be $<10^9$ different clones) and could be normalised, so that each gene is equally represented and not dominated by highly expressed proteins. The use of fragments is likely to overcome problems due to the poor display of full length proteins, and will provide sufficient structural information for immunological selections (using monoclonal or polyclonal antibodies) and interactions due to small domains, but it may fail to identify interactions due to longer polypeptide domains.

One problem with the isolation of DNA fragments encoding open reading frames, is that of frame. Only 1/18 random fragments derived from an open reading frame cloned randomly will be in the correct frame. A genetic method to isolate open reading frames would be the most effective, and the creation of a polycloning site upstream of an antibiotic resistance gene (ampicillin) could, in principle allow the selection of only those pieces of DNA which created open reading frames.

Results achieved

A filter vector, able to filter out open reading frames from non-coding DNA has been designed, constructed and tested. This contains chloramphenicol resistance in the backbone and an ampicillin resistance upstream of gene 3 (the filamentous phage protein usually used for phage display). The ampicillin resistance gene is flanked by two homologous lox sites. By cloning a functional scFv, and one with a frameshift, it was shown that the discrimination in favour of the open reading frame was greater than 100:1. By passaging the phagemid through bacteria expressing cre recombinase, the ampicillin was removed from the vector. This was demonstrated both at the DNA and the protein level. Furthermore, prior to removal of the ampicillin gene, the scFv was very poorly displayed in the phage, but once the ampicillin gene was removed, display was as good as display in standard vectors.

This vector was used to create a small library of fragments from a plasmid containing a single gene - human transglutaminase, an enzyme intimately involved in the pathogenesis of celiac disease. The titre of this library when plated on chloramphenicol plates (which is the resistance gene contained in the backbone of the filter vector) was approximately 100,000 clones, whereas on ampicillin, (which is the antibiotic resistance used to filter out the ORFs) only 2000 clones were obtained. This represents a 50 fold reduction, and corresponds to what would be expected if this vector was effectively selecting fragments of

DNA which encoded open reading frames. Analysis of the clones in this library, to determine whether ORFs have been selected and whether they represent legitimate ORFs, is presently underway.

One possibility which will be explored is that the filter vector is able to select ORFs which form some kind of functional, structural or folding domain. This would be related to the observation that the function of a protein can be inhibited if fused to a poorly folding or aggregating domain. If this was shown to be the case, such a vector could be used to select folding domains from either normalised cDNA libraries or the human genome. Following selection, such domains could then be displayed in phage following removal of the ampicillin gene. The level of expression of such domains in *E. coli* will be explored, with the potential to use this vector to produce protein fragments in a high throughput manner which can then be used to select phage antibodies as described above.

The PI believes this represents a patentable invention, but has not initiated patent proceedings until it has been satisfactorily reduced to practice.

Future directions

At present, with a good library, such as the one described above, the selection of antibodies against any antigen is not a problem. A number of bottlenecks still exist however. The first, mentioned above, is that of creating or deriving the selectors corresponding to gene products. Peptide selectors have a number of advantages, in that one can envisage automated selection systems where open reading frames from genes are identified, the protein sequence scanned according to algorithms developed within the context of this proposal, and potentially 'antigenic' peptides identified, and automatically synthesised on microtitre pins. Selection of polyclonal antibodies against each pin can be automated, and selected antibodies can be stored for future reference. The antibodies selected against peptides in this project have been shown to be functional in ELISA, Western blotting and immunoprecipitation. Phage antibodies have been used in immunohistochemistry, and there is no reason why antibodies selected against peptides could not be used in the same way.

While work on the peptide selection scheme will continue, despite its attractions, this scheme has a number of problems. The first is related to the possibility (demonstrated here) that some selected antibodies may recognise other proteins by virtue of common sequence elements. The second is related to the poorer quality of antibodies selected against peptides as opposed to native protein domains, and the last is related to the need for extensive antibody characterisation (at least two antibodies against each of at least two peptides) following selection, as a result of the occasional binding to similar, but different, proteins.

This problem may be solved by the development of a high throughput protein domain synthesis effort. Should the filter vector prove to be effective in the selection of protein domains, this could be used to derive an extensive library of protein fragments suitable for a wide range of different interests. This could be either done on a random - genome or cDNA - basis, or alternatively on a gene by gene basis, with the creation of small libraries of thousands of clones from individually isolated and identified cDNA clones. The filter vector described above, may provide the means to synthesise many such protein domains in a high throughput manner. Alternative scaffolds, such as GFP or intracellular antibiotic resistance genes (such as chloramphenicol), may provide alternative means to select for different sets of protein domains (e.g. those expressed intracellularly may be more appropriately selected by an intracellular antibiotic resistance). The creation of many sequenced characterised protein domains using this vector could provide the means to select antibodies against true protein domains rather than poorly folding peptides. Such domains would also be useful for structural and protein interaction studies, as well as provide potential ligands for protein chips and biochemical or enzyme studies. The identification of protein sequence tags (PSTs) would also prove to be extremely useful in the further mapping and functional analysis of the genome.

A second bottleneck relates to the downstream use of the antibodies. Presently, they can be produced as scFv either in the bacterial strain in which phage were produced, or at higher levels in strains which are not suppressor strains. Passage from one strain to another can be done by a simple infection step. Unfortunately the levels of production are not high 0.1-10mg for each litre of culture, with most antibodies expressing at levels less than 1mg/litre. The creation of libraries containing V genes which are biased towards high production levels may well overcome this problem. In fact, some scFv can be expressed at levels as high as 70mg/litre, which would be sufficient for most scientific purposes. By preselecting scFvs for high expression from the recombined library, and performing a second recombination step, all

potentially high expressing individual VH and VL genes should be retained, since each will be present with a number of different VH and VL genes, at least one of which will lead to selection. An alternative way around this problem is the use of phage themselves rather than purified scFv. This is presently under investigation.

The last bottleneck is related to the validation of the antibodies in tests other than ELISAs. This would be best carried out as a collaborative effort, with a web based database of genes and antibodies being made available into which researchers using antibodies can write observations on the utility of the various antibodies for different purposes. It is likely that at a certain stage, the requirements of this project will no longer be sustainable in an academic/government setting, and it will need to be transferred to a commercial venture of some sort. Antibody validation is likely to be more labour intensive for antibodies selected on peptides than those selected on proteins or protein domains.

Whatever the selectors used, and whatever the final form of the antibodies used, the selection of antibodies against all gene products is an ambitious project with long time scales: the selection of antibodies against 400 gene products per month will take twenty years to cover the whole genome! To reduce this to a more reasonable 4 years will require selection against 2000 gene products per month. It is clear that automation is essential to complete this work within a reasonable time scale. Initially antibodies produced by this method will find use in scientific research, later, however, as protein chip technologies advance, these antibodies will prove to be very useful as the recognition ligands in such chip devices. The potential importance of this project has been recognised in the commercial sector, where a company initiated a strong effort to derive antibodies against all gene products (using peptide selectors), with a view to selling the antibodies as well as access to the database of antibody binding specificities and tissue distributions. However, they have recently given up this approach, finding that there is insufficient commercial interest in the database, and that the quality of the data is hard to keep at the highest level. This does not detract from the potential value of such a project, which can only really be contemplated within the context of a national publically well-funded organisation.

Value to the community

The value of a source of antibodies against all gene products will be invaluable. In the interim period, antibodies can be derived against proteins produced by researchers in a collaborative venture, with samples of antibodies being maintained at LANL for further distribution and amplification.