Journal of Cellular Biochemistry 80:187–191 (2000)

From Genes to Proteins: High-Throughput Expression and Purification of the Human Proteome

Joanna S. Albala,[1]* Ken Franke,[2] Ian R. McConnell,[1] Karen L. Pak,[1] Peg A. Folta,[1] Brian Karlak,[2,†] Bonnee Rubinfeld,[2] Anthony H. Davies,[2] Gregory G. Lennon,[1,‡] and Robin Clark[2]

[1]Lawrence Livermore National Laboratory, Livermore, California 94550
[2]Onyx Pharmaceuticals, Inc., Richmond, California 94806

Abstract The development of high-throughput methods for gene discovery has paved the way for the design of new strategies for genome-scale protein analysis. Lawrence Livermore National Laboratory and Onyx Pharmaceuticals, Inc., have produced an automatable system for the expression and purification of large numbers of proteins encoded by cDNA clones from the IMAGE (Integrated Molecular Analysis of Genomes and Their Expression) collection. This high-throughput protein expression system has been developed for the analysis of the human proteome, the protein equivalent of the human genome, comprising the translated products of all expressed genes. Functional and structural analysis of novel genes identified by EST (Expressed Sequence Tag) sequencing and the Human Genome Project will be greatly advanced by the application of this high-throughput expression system for protein production.

A prototype was designed to demonstrate the feasibility of our approach. Using a PCR-based strategy, 72 unique IMAGE cDNA clones have been used to create an array of recombinant baculoviruses in a 96-well microtiter plate format. Forty-two percent of these cDNAs successfully produced soluble, recombinant protein. All of the steps in this process, from PCR to protein production, were performed in 96-well microtiter plates, and are thus amenable to automation. Each recombinant protein was engineered to incorporate an epitope tag at the amino terminal end to allow for immunoaffinity purification. Proteins expressed from this system are currently being analyzed for functional and biochemical properties. J. Cell. Biochem. 80:187–191, 2000.

The Human Genome Project has provided a wealth of biological data while fostering a multitude of technological advances for the development of high-throughput, genome-scale science. This massive, government-sponsored undertaking will be completed by 2003, resulting in a compilation of all human gene sequences. In recent years, the field of genomics has advanced the identification and characterization of genes and genomes from many experimental species. Concurrently, the generation of public cDNA databases has further accelerated novel gene discovery. The immediate challenge now lies in determining the function of newly identified genes.

Sequence information can provide insight toward understanding the function of a particular gene using comparative genomics approaches such as Blast analysis [Altschul et al., 1990]. However, for a gene to be fully characterized, it must be examined for functionality at the protein level [Cantor and Little, 1998]. The field of proteomics—the protein equivalent of genomics—has emerged as the next logical progression in deciphering cellular biochemistry beyond gene sequence [Wasinger et al., 1995]. Most proteomic studies to date have focused on protein profiling, which is analogous to differential gene expression, but at the protein level. Typically this is done by 2-D gel electrophoresis coupled with mass spectrometry. By these methods, it is possible to compare a variety of parameters applied to whole cells and dissect their effects on the abundance and modifications of hundreds of proteins [Humphery–Smith and Blackstock, 1997].

Like genomics, the study of the proteome requires a diversity of large-scale operations

†Current address: Molecular Applications Group, Palo Alto, CA 94304 ‡Current address: GeneLogic, Inc., Gaithersburg, MD 20878

Contract grant sponsor: U.S. DOE by LLNL; Contract grant number: W-7405-ENG-48.

*Correspondence to: Dr. Joanna S. Albala, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, 7000 East Avenue L-452, Livermore, CA 94550. E-mail: albala1@llnl.gov

Received 2 August 1999; Accepted 2 August 1999

for protein analysis using automated procedures and a bioinformatics system to analyze the extensive data generated. Several new technologies have been developed for the examination of proteins using genomic formats. Some of these methods include Proteinchips™ (Ciphergen, Inc.), Profusion™ technology (Phylos, Inc.) [Roberts and Szostak, 1997], high-density yeast two-hybrid analysis for protein-protein interactions [Hua et al., 1998], and development of species-specific protein maps [Bartel et al., 1996]. The need has become apparent for additional high-throughput strategies to facilitate the analysis of protein function and structure [Dove, 1999]. We have developed a novel proteomic approach whereby each individual protein encoded by the genome can be expressed and purified at levels sufficient for analysis of biochemical properties and structure on a true genomic scale. As such, this approach differs from the vast majority of proteomic strategies which primarily identify and quantitate proteins from whole cells or tissues.

From Genes to Proteins

Lawrence Livermore National Laboratory (LLNL) and Onyx Pharmaceuticals, Inc. have developed an automatable high-throughput system for protein expression and purification. This partnership combines the experience Onyx has amassed in baculoviral expression and protein purification with the miniaturization, bioinformatics, and high-throughput capabilities that LLNL has advanced as a major genome center. The system utilizes selected cDNAs contained within the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) [Lennon et al., 1996] collection as the basis for the production of an array of purified recombinant proteins.

The IMAGE Consortium was founded to array, sequence, and map a collection of cDNAs representing all human (and subsequently, mouse) genes and to distribute these sequences and clones to the public [Lennon et al., 1996]. cDNA libraries from a variety of sources are arrayed for distribution at LLNL. End sequence for each IMAGE clone is generated at Washington University and deposited immediately into the public sequence database, dbEST. Since the inception of IMAGE approximately 6 years ago, over 2.8 million human cDNA clones have been arrayed, making it the largest publicly available cDNA collection in the world. Greater than 2.3 million sequences from IMAGE clones are included in dbEST, comprising the majority of the human and mouse ESTs in the database. Of the approximate 6,000 known genes characterized to date, more than 90% are contained within IMAGE and of these, 25% are full length.

We have developed a novel method for high-throughput protein production. The greatest advantage of this approach is the ability to access the human proteome from any representative gene within the IMAGE collection. A key component for the design of this strategy was the employment of a versatile system for recombinant protein expression. Insect viruses from the family *Baculoviridae* have been engineered to generate a range of foreign proteins in large quantites in eukaryotic hosts [Summers and Smith, 1987]. Baculoviruses are harmless to humans, proteins can be easily expressed in milligram quantities, and insect cells grow at ambient temperature without the need for carbon dioxide [Kidd and Emery, 1993]. Furthermore, these eukaryotic cells accomplish most posttranslational modifications, including phosphorylation, N- and O-linked glycosylation, acylation, disulfide cross-linking, oligomeric assembly, and subcellular targeting, which may be critical for the biological integrity of recombinant proteins [Davies, 1994]. In contrast to the more commonly used bacterial expression systems, recombinant baculoviral expression generally produces soluble proteins without the need for induction or specific temperature conditions. Although *E. coli* is capable of producing large quantities of recombinant protein, as a prokaryote it does not possess the cellular machinery for post-translational modifications most likely necessary for accurate production of functional human proteins [Guengerich et al., 1993; Williams, 1999]. Moreover, the baculovirus system provides robust expression that is yet to be bettered by other eukaryotic systems. Considering these features, the baculovirus expression system was our technique of choice for converting arrayed cDNAs into protein.

Traditionally, foreign genes are introduced into the 130 kilobase viral genome by homologous recombination. This is accomplished by inserting the cDNA of interest into a plasmid transfer vector containing a suitable viral promoter, termination signal, and flanking viral DNA sequences. Typically, the strong poly-

hedrin promoter is used to drive protein expression and the flanking viral DNA regions allow for recombination at the polyhedrin locus of the virus. Replacement of the polyhedrin gene is tolerated since it is not essential for viral replication or production, and the recombinant mRNA is expressed to the high levels of the native gene [Davies, 1994].

Protein purification from baculoviral-infected cells is comparable to that of any other expression system. Many expression vectors have been developed to generate fusion proteins by the inclusion of a tag sequence onto the recombinant protein that facilitates purification. For instance, polyhistidine fusion proteins are purified by immobilized metal affinity chromatography; fusions to specific immunogenic partners, by immunoaffinity chromatography; and glutathione-S-transferase (GST) fusion proteins, by substrate affinity chromatography [Harlow and Lane, 1988; Smith and Johnson, 1988; Van Dyke et al., 1992]. The recombinant baculoviruses in our system incorporate a Glu-Glu epitope, which is recognized by an antipeptide monoclonal antibody [Grussenmeyer et al., 1985]. This antibody binds efficiently to tagged proteins in crude cellular lysates and yet allows for rapid elution by free peptide under nondenaturing conditions. The use of the Glu-Glu tag for protein purification has been previously described [Rubinfeld et al., 1991; Porfiri et al., 1995; Rubinfeld and Polakis, 1995]. In summary, this immunoaffinity purification scheme has been shown to be rapid and effective resulting in proteins of high purity [Bollag et al., 1993; Rubinfeld and Polakis, 1995]

Strategy for High-Throughput Production of Recombinant Proteins

Our strategy for high-throughput protein expression and purification is outlined in Figure 1. The use of 96-well microtiter plates throughout provides a format for manipulation that is amenable to automation. Selected IMAGE cDNAs, from 384-well master plates, are rearrayed into the 96-well format. Recombinant baculoviruses are produced using a PCR-based method that eliminates the subcloning of cDNAs into transfer vectors and the routine plaque purification steps commonly employed in the generation of recombinant baculoviruses. A 2 ml, 96-well insect cell culture system has been designed for the expression of small amounts (2–10 μg) of epitope-tagged protein expressed from recombinant baculoviruses.

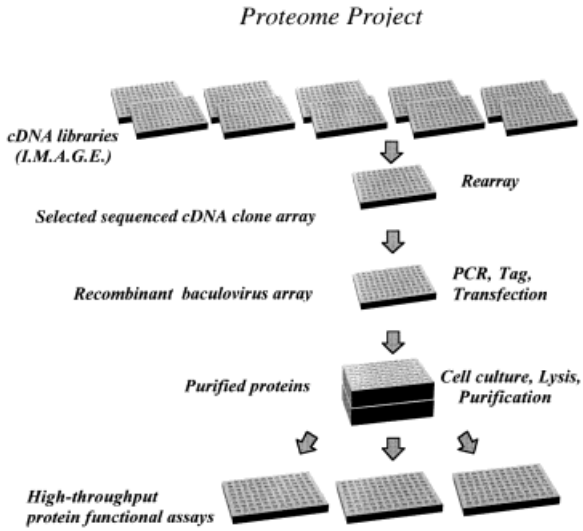

Fig. 1. Schematic of high-throughput protein expression and purification strategy.

Immunoaffinity chromatography of insect cell lysates results in a protein library, in the 96-well format, ready for downstream applications. Throughout, the cDNAs, viruses, and proteins remain addressable, within the original array. A bioinformatics system is under development to manage the data generated from cDNA selection through protein purification.

As a test of transfection efficiency and proof-of-principle, a cDNA encoding the β-glucuronidase gene was cloned into an IMAGE vector and transformed into *E. coli* as a prototype clone. The bacteria containing the *Gus* construct were then inoculated into every well of a 96-well microtiter plate. The cDNA was amplified by PCR and purified in a 96-well format. Recombinant *Gus* virus was produced and β-glucuronidase activity assessed using a colorimetric assay. An 88% expression efficiency was observed, demonstrating the potential for efficient production of recombinant virus starting from an "IMAGE" cDNA clone in a high-throughput format.

In order to test our strategy for automated protein production, a "protoplate" (a prototype, 96-well microtiter plate) was designed to contain a variety of cDNAs selected and rearrayed from the IMAGE collection. The protoplate comprised 88 cDNAs and eight controls. Each cDNA conformed to set criteria including length of open reading frame, brevity of 5′ untranslated region, and absence of upstream stop codons. The majority of cDNAs rearrayed

onto the protoplate were full-length clones representing genes of known function. In several instances partial sequences were selected if the full-length cDNA failed to conform to the designated criteria or it was the only clone available for a given gene. In addition, some partial cDNAs were chosen to represent protein functional domains.

To streamline cDNA selection, a web-based tool, IMAGEne [Cariaso et al., 1999; http://bbrp.llnl.gov/imagene/bin/search], was used to select the "best" IMAGE clone available for a given gene. This is a multimodular tool that clusters ESTs from IMAGE clones to known genes, ranks these clones within a cluster, and displays the alignments of ESTs to their associated genes via a HTML/Java-based interface. Clustering algorithms have recently been completed for application toward unknown genes in order to examine novel proteins.

The selected cDNAs were amplified by PCR using a high-fidelity polymerase. Forward oligonucleotide primers were designed to ensure that each cDNA insert was in-frame with respect to the epitope tag and the start codon within the PCR primer sequence. To determine the appropriate frame and primer for each cDNA to be amplified, we have developed a software tool, Framefinder™. This program was designed to identify the most probable open reading frame (ORF), determine the exact cDNA insertion site, and indicate the correct reading frame of that cDNA relative to the vector sequence. In some cases all three forward primers (designating all possible ORFs) were employed in a single PCR reaction. These were used to assess the applicability of this method of amplification toward cDNA clones of unknown or insufficient sequence.

PCR products were purified (seven clones failed to produce a detectable product) and incorporated into recombinant baculoviruses in the 96-well plate using a transfection protocol. Cells infected with amplified recombinant virus were harvested after 72 h for protein expression. Western blot analysis using the Glu-Glu antibody demonstrated that 34 of 81 wells yielded soluble, recombinant protein. The expressed proteins incorporate an amino-terminal epitope tag to facilitate purification and analysis. A 2 ml, 96-well insect cell culture system has been designed for immunoaffinity purification, which maintains the high-throughput format. Lysates were prepared from insect cells infected with recombinant virus and adsorbed onto an immunoaffinity matrix. Recombinant protein was eluted from the immunoaffinity matrix by competition with synthetic peptide. Purification yielded similar results and purity to routine protocols for larger volumes of insect cell culture [Rubinfeld and Polakis, 1995]. Purified recombinant proteins maintain an addressable format and can be stored or rearrayed for subsequent analysis.

A Proteome database has been designed to track the progression of each clone from IMAGE cDNA through the processes of PCR, protein expression, purification and functional characterization. Information regarding clone identification, sequence accession numbers, viral passages, and additional pertinent details are recorded for each plate. It has a World Wide Web user interface and was created utilizing the Sybase database with links to the IMAGE database. Ultimately, the data generated from the Framefinder™ and IMAGEne programs will be combined with the Proteome database resulting in a single, integrated bioinformatics platform.

In summary, the strategy described provides the basis for high-throughput protein expression and purification on a genomic scale. It has been designed for automation allowing the high-throughput protein production essential to generate the human proteome. This system provides the ability to select, array, express and re-express proteins from the largest public cDNA collection. Several future applications for functional and structural studies using this baculoviral expression paradigm can be envisioned, for example, high-throughput structural analysis by NMR and X-ray crystallography, small molecule binding screens and high-throughput enzymatic assays. The versatility of our protein production system coupled with the unique resources of IMAGE is a powerful combination for the advancement of functional genomics.

ACKNOWLEDGMENTS

The authors are grateful to Barbara Belisle and Chris Celeri of Onyx Pharmaceuticals, Inc., for their technical expertise. We acknowledge the invaluable assistance of Christa Prange and LeAnne Mila of the IMAGE Consortium at LLNL for providing the cDNA clones. The authors would also like to thank Emily Quinnan for her contributions to system

automation, and special thanks to Dr. James Sommercorn for his advice and helpful discussions.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403–410.

Bartel PL, Roecklein JA, SenGupta D, Fields S. 1996. A protein linkage map of *Escherichia coli* bacteriophage T7. Nat Genetics 12:72–77.

Bollag G, McCormick F, Clark R. 1993. Characterization of full-length neurofibromin: tubulin inhibits Ras GAP activity. EMBO J 12:1923–1927.

Cantor CR, Little DP. 1998. Massive attack on high-throughput biology. Nat Genetics 20:5–6.

Cariaso M, Folta P, Wagner M, Kuczmarski T, Lennon G. 1999. IMAGEne I: clustering and ranking of IMAGE cDNA clones corresponding to known genes. Bioinformatics, in press.

Davies AH. 1994. Current methods for manipulating baculoviruses. Biotechnology 12:47–50.

Dove A. 1999. Proteomics: translating genomics into products? Nat Biotechnol 17:233–236.

Grussenmeyer T, Scheidtmann KH, Hutchinson MA, Eckhart W, Walter G. 1985. Complexes of polyoma virus medium T-antigen and cellular proteins. Proc Natl Acad Sci 82:7952–7954.

Guengerich FP, Gillam EMJ, Ohmori S, Sandhu P, Brian WR, Sari M-A, Iwasaki, M. 1993. Expression of human cytochrome P450 enzymes in yeast and bacteria and relevance to studies on catalytic specificity. Toxicology 82:21–37.

Harlow E, Lane D. 1988. Antibodies: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Press.

Hua S, Luo Y, Qui M, Chan E, Zhou H, Zhu L. 1998. Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. Gene 215:143–152.

Humphery-Smith I, Blackstock W. 1997. Proteome analysis: genomics via the output rather than the input code. J Protein Chem 16:537–544.

Kidd M, Emery VC. 1993. The use of baculoviruses as expression vectors. Appl Biochem and Biotech 42:137–159.

Lennon G G, Auffray C, Polmeropoulos M, Soares M. 1996. The IMAGE Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151–152.

Porfiri E, Evans T, Bollag G, Clark R, Hancock JF. 1995. Purification of baculovirus-expressed recombinant Ras and Rap proteins. Methods Enzymol 255:13–21.

Roberts RW, Szostak JW. 1997. RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci USA 94:12297–12302.

Rubinfeld B, Munemitsu S, Clark R, Conroy L, Watt K, Crosier WJ, McCormick F, Polakis P. 1991. Molecular cloning of a GTPase activating protein specific for the Krev-1 protein p21 Rap1. Cell 65:1033–1042.

Rubinfeld B, Polakis P. 1995. Purification of baculovirus-produced Rap1 GTPase-activating protein. Methods Enzymol 255:31–38.

Smith DB, Johnson KS. 1988. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene 67:31–40.

Summers MD, Smith GE. 1987. A manual of methods for baculovirus vectors and insect cell culture procedures. Tex Agric Exp Stn Bull B1555:1–56.

Van Dyke MW, Sirito M, Sawadogo, M. 1992. Single-step purification of bacterially expressed polypeptides containing an oligo-histidine domain. Gene 111:99–104.

Wasinger VC, Cordwell SJ, Ceroa-Poljak A, Yan JX, Gooley AA, Wilkins KL, Humphery-Smith I. 1995. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis 16:1090–1094.

Williams KL. 1999. Genomes and proteomes: toward a multidimensional view of biology. Electrophoresis 20: 678–688.