Pooled ORF Expression Technology (POET)

USING PROTEOMICS TO SCREEN POOLS OF OPEN READING FRAMES FOR PROTEIN EXPRESSION*[S]

William K. Gillette‡, Dominic Esposito‡, Peter H. Frank‡, Ming Zhou§, Li-Rong Yu§, Catherine Jozwik¶, Xiuying Zhang¶, Brighid McGowan¶, David M. Jacobowitz¶‖, Harvey B. Pollard¶, Tong Hao, David E. Hill**, Marc Vidal**, Thomas P. Conrads§, Timothy D. Veenstra§, and James L. Hartley‡ ‡‡**

We have developed a pooled ORF expression technology, POET, that uses recombinational cloning and proteomic methods (two-dimensional gel electrophoresis and mass spectrometry) to identify ORFs that when expressed are likely to yield high levels of soluble, purified protein. Because the method works on pools of ORFs, the procedures needed to subclone, express, purify, and assay protein expression for hundreds of clones are greatly simplified. Small scale expression and purification of 12 positive clones identified by POET from a pool of 688 *Caenorhabditis elegans* ORFs expressed in *Escherichia coli* yielded on average 6 times as much protein as 12 negative clones. Larger scale expression and purification of six of the positive clones yielded 47–374 mg of purified protein/liter. Using POET, pools of ORFs can be constructed, and the pools of the resulting proteins can be analyzed and manipulated to rapidly acquire information about the attributes of hundreds of proteins simultaneously. ***Molecular & Cellular Proteomics 4:1647–1652, 2005.***

Projects aiming to convert the thousands of genes made accessible by genomic sequences into their corresponding proteins have met with limited success (1) despite the expenditure of significant resources (2, 3). Expression of recombinant proteins in *Escherichia coli*, the primary host organism for high throughput applications, has been especially unsuccessful for metazoan proteins. For example, one effort directed at producing *Caenorhabditis elegans* proteins successfully purified only about 2% of those attempted (4).

In the standard approach to high throughput protein expression and purification used in these programs, genes are cloned individually into an expression vector, introduced into an expression host, and expressed in separate cultures. Each culture is subsequently tested for expression and solubility of the corresponding protein at which point new cultures of positive clones are grown and induced so that protein purification can be attempted. Even with intensive use of robotics, the logistics and costs of this strategy are considerable when thousands of genes are put into such a pipeline.

Here we describe a method called pooled ORF expression technology (POET)[1] that avoids many of the logistical issues associated with high throughput protein expression and purification. POET combines recombinational cloning and collections of sequenced ORFs with proteomic methods (two-dimensional gel electrophoreses (2DGE) and MS) to predict which ORFs in a pool will yield soluble, purified protein. We applied POET to a pool of 688 *C. elegans* ORFs. A high percentage of ORFs identified in this experiment yielded expressed, soluble, purified proteins in agreement with POET predictions.

EXPERIMENTAL PROCEDURES

ORFs—The *C. elegans* ORFeome version 1.1 has been described previously (5). The predicted initiating methionine of each ORF was changed to leucine (ATG → TTG), and the stop codon of each ORF was omitted. The DNA concentrations of the 752 Gateway entry clones used in this experiment were determined by PicoGreen (Molecular Probes) fluorescence and used to calculate the molar concentration of each plasmid (based on the size of each ORF and the size of the pDONR201 backbone), which ranged from 0 to 8.73 nM. All wells containing <0.15 nM plasmid concentration were omitted, leaving 688 ORFs. Plasmid DNAs were pooled in bins of 2-fold concentration range, starting with the most concentrated plasmid and going down 2-fold, 4-fold, etc. for a total of six subpools. These subpools were combined volumetrically, 1 volume of the most concentrated subpool, 2 volumes of the next most concentrated subpool, etc. The final pool was ethanol-precipitated and dissolved in Tris-EDTA to a final concentration of 2.5 ng/μl. The result of these manipulations was a single pool in which no plasmid varied in molar concentration from any other plasmid by more than 2-fold.

From the ‡Protein Expression Laboratory and §Laboratory of Proteomics and Analytical Technologies, Research Technology Program, SAIC-Frederick, Inc./NCI, National Institutes of Health, Frederick, Maryland 21702, ¶Department of Anatomy, Physiology, and Genetics, Uniformed Services University, Bethesda, Maryland 20814, **Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, and ‖Laboratory of Clinical Science, National Institute of Mental Health, Bethesda, Maryland 20892

Received May 4, 2005, and in revised form, August 17, 2005

Published, MCP Papers in Press, August 19, 2005, DOI 10.1074/mcp.M500128-MCP200

[1] The abbreviations used are: POET, pooled ORF expression technology; attL, site-specific recombination sites for Gateway cloning; LR Clonase, site-specific recombination enzyme mixture for Gateway cloning; 2D, two-dimensional; 2DGE, two-dimensional gel electrophoresis; LIT, linear ion trap.

Protein Expression—The 688 pooled ORFs (as Gateway attL entry clones (6)) were subcloned into pDest527 (T7 promoter, amino-terminal His_6 fusion) with LR Clonase (Invitrogen) according to the manufacturer's instructions except that the reaction was allowed to proceed for 5 h at 30 °C. Each ORF expressed in this vector contained the sequence MRSGSHHHHHHRSDITSLYKKAG added to its amino end and YPAFLYKVVISLAR added to its carboxyl end due to the lack of the native stop codon. Reaction products were transformed into DH5α cells (Invitrogen), and 1% of the SOC expression mixture was plated on ampicillin (145 colonies). The remaining 99% of the expression mixture was added to 50 ml of CircleGrow (QBiogene) containing ampicillin, and after overnight growth at 37 °C plasmid DNA was purified (Brinkmann Fast Plasmid). About 100 ng of pooled expression plasmids were electroporated into *E. coli* Rosetta (DE3) strain (Novagen), which compensates for eukaryotic codons that are rare in *E. coli*. Colonies from an aliquot of this reaction indicated that about 1.4 million transformants resulted. The 1 ml of SOC expression mixture was diluted into 50 ml of CircleGrow (containing 100 μg/ml ampicillin) and grown overnight at 37 °C. The overnight culture was diluted 1:100 into 1 liter of CircleGrow (containing 100 μg/ml ampicillin), grown at 37 °C to an A_{600} of 0.5, and cooled to 16 °C, and protein expression was induced by adding isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 0.5 mM. After 16 h at 16 °C cells were harvested and frozen at −80 °C.

Protein Purification—All steps were performed at 4 °C unless otherwise noted. *E. coli* cell pastes were resuspended with 2 volumes of extraction buffer/g of wet weight for a final concentration of 20 mM sodium phosphate buffer, pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 5% glycerol, 45 mM imidazole, and Complete protease inhibitor-EDTA (Roche Applied Science) at one tablet/50 ml of extract. Extracts were treated with lysozyme (0.5 mg/ml) for 30 min and with Benzonase (Novagen, 10 units/ml) for an additional 20 min. Samples were sonicated to lyse the cells (verified by microscopic examination), adjusted to 500 mM NaCl with solid NaCl, centrifuged at 111,000 × *g* for 30 min, filtered (0.45 μm, polyethersulfone membrane) and applied at 0.6 ml/min to 1-ml HisTrap columns (Amersham Biosciences) equilibrated with extraction buffer in 500 mM NaCl and 45 mM imidazole (binding buffer). The columns were washed with binding buffer until the levels of protein flowing through the column reached base line, and bound proteins were eluted with binding buffer + 500 mM imidazole, collected in 1-ml fractions, and analyzed by SDS-PAGE. The pools created from the IMAC fractions were precipitated by adding 25% (v/v) TCA to 6% (v/v) final concentration, vortexed, incubated on ice for 5 min, and centrifuged at 16,100 × *g* for 10 min. The supernatant was removed, and the pellet was incubated with ice-cold acetone for 5 min on ice and then centrifuged at 16,100 × *g* for 5 min. The supernatant was discarded, and the pellet was dried for 2 min at 70 °C, dissolved in room temperature solubilization buffer (8 M urea, 4% CHAPS, 50 mM Tris, pH 8.5) to a concentration of 20 mg/ml (Bio-Rad protein assay), and stored in 50-μl aliquots at −80 °C.

Two-dimensional Gel Electrophoresis—Two-dimensional PAGE of 200–1000 μg of pooled protein was performed according to the procedure of O'Farrell (7) with more recent modifications (8, 9). In the first dimension, isoelectric focusing was accomplished by IPG using the Amersham Biosciences IPGphor isoelectric focusing system. Affinity-purified samples were dissolved in 450 μl of rehydration buffer (8 M urea, 2% CHAPS, 7 mg DTT, and a trace of bromphenol blue). The rehydration buffer-protein mixture was placed in a 24-cm ceramic strip holder, and a 24-cm IPG strip was glided, gel side down, into the strip holder. Mineral oil was placed on top of the gel to minimize evaporation and covered with the strip holder plastic cover. The ceramic strip holder was placed in the IPGphor unit for isoelectric focusing and set at 30 V for 12 h, 500 V for 1 h, 1000 V for 1 h, and 8000 V for 8 h.

Prior to the second dimension step of separation by molecular weight, the IPG strip was equilibrated with an SDS buffer system. The equilibration solution contained 50 mM Tris-HCl, pH 8.8, 6 M urea, 30% glycerol, 2% SDS, and a trace of bromphenol blue. Prior to use, 100 mg of DTT was added in 10 ml of equilibration buffer. The IPG strips were placed in individual tubes containing the buffer. The tubes were then placed on a rocker and equilibrated for 12 min. A second equilibration was performed with 250 mg of iodoacetamide solution (instead of DTT) and incubated for another 12 min. The equilibrated IPG strip was then inserted into a cassette containing a precast Ettan DALT II 12.5% polyacrylamide gel, and contact was made with the gel. Enough melted agarose was added to cover the IPG strip. The 2DGE chamber was filled with anode buffer (0.5 M diethanolamine, 0.5 M acetic acid). Cathode buffer (0.1% SDS, 0.192 M glycine, 0.025 M Tris) was added to the top chamber. The running conditions were set in the power supply (phase 1, 5 watts/gel/15 min; phase 2, 150 watts/gel), and electrophoresis continued until the bromphenol blue dye front reached the bottom of the gel (~4–5 h). Once the dye front reached the end of the gel, the cassettes were removed, and the gels were placed in Coomassie Brilliant Blue staining solution (25% isopropanol, 10% acetic acid, 0.05% R250 Brilliant Blue) overnight. Gels were then placed in destain solution (30% methanol, 10% acetic acid). All staining/destaining procedures were carried out in glass trays placed on a slowly oscillating rocker table.

Spot Picking, Digestion, and Analysis—The spots on the 2D PAGE gel were numbered 1–170, and small pieces were retrieved from the center of each spot. Coomassie Blue-stained protein gel spots were digested with trypsin as described previously (8). Samples were desalted with C_{18} Zip Tips (Millipore, Bedford, MA) according to the manufacturer's protocols prior to MS analysis. Chromatographic separations of desalted tryptic peptides were conducted using a 75-μm-inner diameter × 360-μm-outer diameter × 10-cm-long fused silica capillary column (Polymicro Technologies Inc., Phoenix, AZ) with one end flame-pulled to a fine tip (~5–7-μm orifice). The column was slurry-packed in-house with 3-μm, 300-Å pore size C_{18} stationary phase (Vydac, Hercules, CA). Nanoflow reversed-phase LC was performed using an Agilent 1100 nanoflow LC system (Agilent Technologies, Palo Alto, CA) coupled on line to a linear ion trap (LIT) mass spectrometer (LTQ, ThermoElectron, San Jose, CA). Reversed-phase separations were conducted after injecting 5 μl of sample for each analysis. The columns were connected via a stainless steel union to an Agilent 1100 nanoflow LC system (Agilent Technologies), which was used to deliver solvents A (0.1% HCOOH in water) and B (0.1% HCOOH in CH_3CN). After sample injection, a 20-min wash with 98% mobile phase A was applied, and peptides were eluted using a linear gradient of 2% mobile phase B to 42% solvent B over 40 min with a constant flow rate of 200 nl/min. The column was washed for 15 min with 98% mobile phase B and re-equilibrated with 98% mobile phase A prior to subsequent sample loading.

The nanoflow reversed-phase LC column was coupled on line to a LIT mass spectrometer using the manufacturer's nanoelectrospray source with an applied electrospray potential of 1.5 kV and capillary temperature of 160° C. The LIT mass spectrometer was operated in a data-dependent mode where each full MS scan was followed by five MS/MS scans in which the five most abundant peptide molecular ions detected from the MS scan were dynamically selected for five subsequent MS/MS scans using a CID energy of 35%. The CID spectra were analyzed using SEQUEST operating on a Beowulf 18-node parallel virtual machine cluster computer (ThermoElectron) using a combined non-redundant *C. elegans*, *E. coli* proteome database (www.expasy.org). Only peptides with conventional tryptic termini (allowing for up to two internal missed cleavages) possessing Δ correlation scores (ΔC_n) >0.08 and charge state-dependent cross-correlation (X_{corr}) criteria as follows were considered as legitimate identifications:

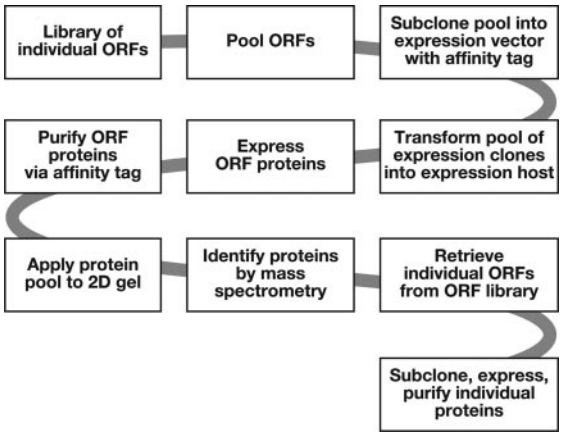

FIG. 1. **Schematic of the POET method.**

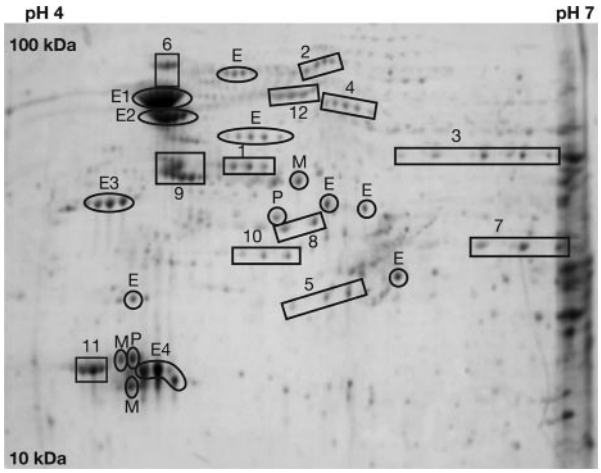

FIG. 2. **2DGE of proteins from 688 *C. elegans* ORFs expressed as His_6 fusions in *E. coli* and purified by affinity chromatography.** *Numbered rectangles* (*C. elegans* positives) correspond to *lane numbers* in Fig. 3. *Circles* and *ovals* identify *E. coli* proteins (*E*), mixtures of two or more proteins (*M*), and presumed proteolytic fragments (*P*). *E1*, DnaK; *E2*, GroEL; *E3*, OmpF; *E4*, SlyD.

>1.9 for +1 charged peptides, >2.2 for +2 charged peptides, and >3.1 for +3 charged peptides.

Testing of Predicted Positive ORFs—Based on visual inspection of the 2D gel and mass spectrometer identifications of the 165 total spots, 12 individual ORFs were retrieved from the ORF plates. These were subcloned into pDest527 and expressed in *E. coli* Rosetta in 700-μl cultures in a 24-well dish to an A_{600} of 0.5, then transferred to 17 × 100-mm polypropylene tubes (Falcon 2059), cooled to 16 °C, induced with isopropyl 1-thio-β-D-galactopyranoside (0.5 mM), and expressed overnight at 16 °C. To determine the fraction of soluble and insoluble protein, cells were lysed with detergent (ReadyPreps, Epicenter), and soluble and insoluble fractions were applied to SDS-PAGE. Recombinant His_6 fusion proteins were purified from the soluble fractions with Swell Gel beads (Pierce) and spin columns. Six ORFs chosen from this small scale experiment were grown at 1-liter scale and purified using the same procedure as the pool of 688 (above). Concentrations of the proteins in the small and large scale preparations were determined using the Bio-Rad protein assay.

RESULTS

Description of POET—The POET scheme is shown in Fig. 1. Hundreds of ORFs are pooled, and the pooled ORFs are subcloned en masse into a protein expression vector supplying an affinity purification tag. The resulting pool of expression plasmids is introduced into an appropriate host and expressed in a single culture of host cells, and the tagged expressed proteins are purified away from host proteins. (For a culture of volume *V* containing *n* ORFs, the protein from any particular ORF is derived from *V*/*n* cells, whereas host proteins are derived from *V* cells. Thus from mass considerations host proteins are much more abundant than proteins from individual ORFs.) The mixture of ORF proteins is then separated by 2DGE, and individual proteins are identified by MS. Proteins in intensely staining spots are predicted to be expressed as abundant, soluble proteins that can be easily purified. These predictions are confirmed by retrieving and expressing individual ORFs from the original ORF collection.

Subcloning of Pooled ORFs—A pool of 688 ORFs in the form of Gateway entry clones (6) was created from the *C. elegans* ORFeome (5). Wide variations in ORF DNA concentration were corrected by first creating subpools of clones having similar concentrations and then combining the subpools volumetrically (overall molar variation ≤2-fold). The pooled ORFs were subcloned via a Gateway LR reaction into an *E. coli* expression vector (pDest527) that added a hexahistidine (His_6) tag to the amino terminus of the protein expressed from each ORF. The LR reaction was transformed into a non-expression *E. coli* strain (DH5α), and purified DNA from this transformation was then transformed into *E. coli* strain Rosetta (DE3) for subsequent protein expression.

Protein Expression, Purification, and Identification—Proteins were expressed from the pooled ORFs at 16 °C (1-liter culture, equivalent to 1.45 ml for each ORF in the pool), the *E. coli* cells were lysed by sonication, and soluble proteins were purified by IMAC. The purified protein pool was precipitated with acid, dissolved in urea/CHAPS buffer, and resolved by 2DGE (Fig. 2). A large number of spots on the gel were identified by MS to understand more fully parameters important to the POET process. The most intense spots on the gel were identified as *E. coli* proteins DnaK, GroEL, SlyD, and OmpF (Fig. 2). Because the isoelectric focusing range on the gel was pH 4–7, about 200 of the 688 *C. elegans* proteins were predicted to appear on the 2D gel based on their calculated pI values. A total of 50 *C. elegans* proteins and 37 *E. coli* proteins were identified by MS of 165 spots selected from the gel for analysis (see the supplemental table).

Small Scale Verification of POET Results—Twelve *C. elegans* proteins were chosen from the 2D gel in a blinded fashion using only the number of spots and their intensity and purity (*i.e.* spots containing more than one protein were given less weight) as selection parameters (Table I). These 12 positive proteins were identified by an average of nine different tryptic peptides. Twelve negative proteins were also chosen from the set of 688 as having a predicted pI between 4 and 7

TABLE I

Expression and purification of 12 POET positive (1–12) and negative (13–24) ORFs

Lane numbers refer to Fig. 3. ND, not determined.

Lane no.	Protein annotation	Swiss-Prot accession no.	Protein molecular weight[a]	Yield	
				Small scale	Large scale
				mg/liter	
1	TAB1-like protein TAP-1	Q93375	47,735	37	ND
2	Protein with tau-like repeats, isoform a	O02592	53,682	5	ND
3	Machado-Joseph disease-like protein	O17850	40,130	10	ND
4	Hypothetical protein C17G10.2	Q09974	53,316	98	100
5	Skp1p homolog (SKR-12)	Q22871	23,192	34	ND
6	Hypothetical protein F09G2.9	O17406	48,553	77	47
7	Bag1 homolog protein 1	O44739	28,276	32	ND
8	Hypothetical protein F53F4.3	Q20728	29,707	144	364
9	Hypothetical protein D2096.8	Q19007	39,931	130	374
10	14-3-3-like protein 1	P41932	32,428	59	60
11	Troponin C, isoform 2	Q09665	22,493	163	ND
12	Troponin T	Q27371	51,307	80	125
13	Hypothetical protein F46A9.5	Q93647	24298	76	ND
14	Mitochondrial import receptor subunit TOM20 homolog	Q19766	24,692	20	ND
15	Probable mediator complex subunit soh-1	P91869	23,744	9	ND
16	Eukaryotic peptide chain release factor subunit 1	O16520	53,484	0	ND
17	CCR4-NOT transcription complex subunit 7	Q17345	38,104	0	ND
18	Probable coatomer ζ subunit	O17901	25,023	0	ND
19	Glutathione *S*-transferase 3	O16116	28,001	31	ND
20	Spermatocyte protein spe-27 (precursor)	P54218	19,449	0	ND
21	CCR4-NOT transcription complex subunit 7	Q17345	38,104	0	ND
22	Myoblast determination protein 1 homolog	P22980	40,715	0	ND
23	G_2/mitotic-specific cyclin A1	P34638	59,776	0	ND
24	Hypothetical protein B0035.11	Q17431	52,329	0	ND

[a] Predicted molecular weight of the protein expressed in pDest527.

and not being identified in the 2D gel spots that were examined. The 24 ORFs corresponding to these proteins were cloned individually into the same *E. coli* expression vector (amino His_6 fusions) described above and expressed in 700-μl cultures at 16 °C, and the resulting proteins were affinity-purified with spin columns. Total and soluble proteins from each culture (Fig. 3, *a* and *b*) and comparison of the purified proteins (Fig. 3*c* and Table I) validate the predictions of the POET method. Proteins from the positive clones were much more likely to be soluble (Fig. 3, *a versus b*) and give abundant purified protein (Fig. 3*c*) than the negative clones. An average of 6 times as much protein was recovered from the positive clones as from the negative clones (Table I). Most of the proteins migrated in accord with their predicted molecular weights. Proteins in *lanes 2* ("protein with tau-like repeats, isoform a") and *6* ("hypothetical protein F09G2.9") had extensive predicted hydrophobic regions that may account for their abnormally low electrophoretic mobilities in both the 2D (Fig. 2) and one-dimensional (Fig. 3) gels. We speculate that the "negative" protein in *lane 13* was not identified on the 2D gel because its solubility was low at its isoelectric point, and it failed to leave the isoelectric focusing strip.

Large Scale Verification—To verify that the small scale experiments could predict successful larger scale behavior, six of the positive ORFs were expressed in *E. coli* individually in 1-liter cultures. Soluble proteins were released from cells by sonication and ultracentrifugation and purified on preparative

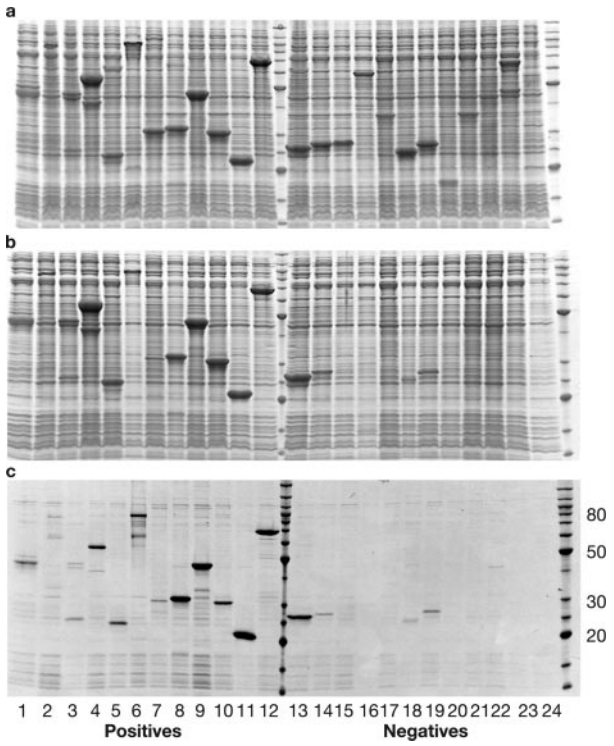

FIG. 3. **Expression analysis of 12 POET positives (intense spots on the 2D gel) and negatives (not identified on the gel) in small (700-μl) individual cultures.** *a*, total proteins. *b*, soluble proteins. *c*, purified proteins. See Table I for protein identities.

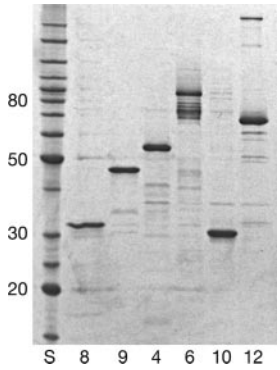

FIG. 4. **Six proteins chosen from Fig. 3 were purified from 1-liter *E. coli* cultures by IMAC.** 1 μg of each protein was applied to the gel. *Lane numbers* correspond to Fig. 3. Yields of the proteins are shown in Table I.

IMAC columns (Fig. 4 and Table I). All six ORFs yielded large amounts (47–374 mg/liter) of purified protein.

DISCUSSION

POET is a procedure for finding which ORFs in a collection of hundreds of ORFs can be most efficiently converted, by cloning, expression, and purification, into their corresponding recombinant proteins. By first combining *n* ORFs into a single pool, tasks that are difficult to accomplish hundreds of times (transformation, plating, colony picking, culture, induction, lysis, assays of solubility, and purification) are reduced in number *n*-fold. The problem that then arises, of course, is how to identify which proteins in the purified pool are the most abundant. MS can identify proteins with spectacular sensitivity, but it is also dramatically non-quantitative (due to the unpredictable ionization of peptides of different amino acid sequence). Thus MS alone, although it can identify all the proteins in the pool, cannot distinguish between the most abundant and least abundant proteins in that pool. (Isotopic labeling methods such as ICAT (10) are not useful, because they give relative quantitation of the same protein in two different samples, whereas POET requires relative quantitation of many different proteins in one sample.)

We chose 2DGE to determine abundance of the expressed, purified proteins in our POET pool. 2D gels have limitations of size and pI, and running them is not trivial. But they can resolve thousands of proteins, and often individual spots contain a single protein (see the supplemental table). For the purposes of the present study we assumed that the size and intensity of stained spots was a reasonable indicator of the abundance of each protein. Combinations of liquid chromatography columns could be used to generate dozens of fractions of the protein pool, but few if any fractions would contain single proteins, and quantitation would thus be unsatisfactory.

We wished to identify a large number of the spots on the 2D gel so that we could understand the parameters of the POET experiment more completely. Many of the highly abundant spots that were identified were *E. coli* stress response proteins that would not require reidentification in the analysis of subsequent ORF pools, allowing the focus to be only on new spots from *C. elegans* clones. Because many of the manual steps at the end of POET can be automated (spot identification, picking, digestion, MALDI plate preparation, and MALDI-TOF/TOF peptide identification), we estimate the net efficiency gain for POET to be 10–100-fold when compared with automated or manual one-by-one methods, respectively.

Although the POET scheme is straightforward, we recognize there are underlying assumptions that can affect its results. 1) All the ORFs in a pool should retain their representation during subcloning and subsequent transfer into expression hosts. Recombinational cloning appears to be essential to minimize size bias and maximize efficiency. Loss of some clones due to toxic effects of ORF expression will tend to increase the representation of remaining clones on the 2D gels. 2) POET assumes that the intensities of spots on 2D gels reflect the amounts of each protein in the purified pool prior to electrophoresis. However, not all proteins remain soluble during isoelectric focusing, and some proteins migrate as multiple spots. 3) Soluble proteins may interact as they are released from cells. The assumption is that as the pool is purified the effect of any one recombinant protein on the behavior of any other member of the pool is small. 4) More than one ORF may be expressed in a particular host cell. POET assumes that coexpression of any two ORFs in the same cell will be distributed more or less randomly among all the ORFs in the pool, and the influence of any one ORF on the behavior of any other ORF is small.

We foresee numerous applications of POET. 1) The large data sets that can be produced by POET experiments could provide researchers with *a priori* knowledge of what is the most appropriate context to express and purify any candidate protein of interest. 2) Because the solubility of overexpressed proteins is often low, one could take the insoluble fraction of proteins from a POET experiment, divide the insoluble proteins into aliquots, subject each aliquot to a different refolding regimen, and identify which procedure works best for any protein in the pool, thus obtaining hundreds of results for each refolding protocol. 3) Small differences in amino acid sequence can cause vastly different behaviors of proteins during overexpression and purification. Using POET, proteins from mouse, rat, and other model organisms can be attempted if purification of the homologous human proteins fail. 4) Because membrane proteins are difficult to extract and purify, ORF pools comprising the extracellular and/or intracellular domains of hundreds of membrane proteins could be constructed, expressed, purified, and analyzed in POET experiments. 5) The optimal number of ORFs in a pool can be adjusted in conjunction with protein expression, purification, and separation parameters. Clearly the larger the number of ORFs in each pool, the fewer the number of overall experiments are required. However, as the number of ORFs increases the average intensity of each ORF spot on the 2D gel

decreases, and the intensities of ORF spots decrease compared with spots from expression host proteins. Analysis of relatively large pools of ORFs should be improved by comparison of multiple pools that contain unrelated ORFs because host proteins in each pool are relatively constant and can be ignored.

Gene libraries such as the National Institutes of Health Mammalian Gene Collection (mgc.nci.nih.gov/) and collections of ORFs from *Homo sapiens* (11–16) and other model organisms (5, 17–21) already exist or are coming into being. In combination with these resources, POET may help make large numbers of important proteins accessible to the scientific community.

Acknowledgments—We thank Robert Stephens of the Advanced Biomedical Computing Center, SAIC-Frederick, Inc., for valuable bioinformatics support and Sukanya Chowdhury, Megan Bucheimer, and Kelly Esposito of the Protein Expression Laboratory for skilled technical assistance.

* This work was supported in part by the Intramural Research Program of the National Institute of Mental Health (to D. M. J.) and by United States funds from the NCI, National Institutes of Health, under Contract NO1-CO-12400 (to W. K. G., D. E., P. H. F., M. Z., L.-R. Y., T. P. C., T. D. V., and J. L. H.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

‡‡ To whom correspondence should be addressed: Protein Expression Laboratory, NCI, National Institutes of Health, SAIC-Frederick, Inc., P. O. Box B, Frederick, MD 21702. Tel.: 301-846-7375; Fax: 301-846-6631; E-mail: hartley@ncifcrf.gov.

REFERENCES

1. Service, R. F. (2002) Tapping DNA for structures produces a trickle. *Science* **298,** 948–950
2. Lattman, E. (2004) The state of the Protein Structure Initiative. *Proteins* **54,** 611–615
3. Frazier, M. E., Johnson, G. M., Thomassen, D. G., Oliver, C. E., and Patrinos, A. (2003) Realizing the potential of the genome revolution: the genomes to life program. *Science* **300,** 290–293
4. Luan, C. H., Qiu, S., Finley, J. B., Carson, M., Gray, R. J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D. E., Vidal, M., Delucas, L. J., and Luo, M. (2004) High-throughput expression of C. elegans proteins. *Genome Res.* **14,** 2102–2110
5. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004) A map of the interactome network of the metazoan C. elegans. *Science* **303,** 540–543
6. Hartley, J. L., Temple, G. F., and Brasch, M. A. (2000) DNA cloning using in vitro site-specific recombination. *Genome Res.* **10,** 1788–1795
7. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021
8. Gorg, A., Postel, W., and Gunther, S. (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **9,** 531–546
9. Gorg, A., Obermaier, C., Boguth, G., and Weiss, W. (1999) Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures. *Electrophoresis* **20,** 712–717
10. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
11. Rual, J. F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P. O., Clingingsmith, T. R., Hartley, J. L., Esposito, D., Cheo, D., Moore, T., Simmons, B., Sequerra, R., Bosak, S., Doucette-Stamm, L., Le Peuch, C., Vandenhaute, J., Cusick, M. E., Albala, J. S., Hill, D. E., and Vidal, M. (2004) Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.* **14,** 2128–2135
12. Pearlberg, J., and LaBaer, J. (2004) Protein expression clone repositories for functional proteomics. *Curr. Opin. Chem. Biol.* **8,** 98–102
13. Messina, D. N., Glasscock, J., Gish, W., and Lovett, M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* **14,** 2041–2047
14. Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T., and Sugano, S. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.* **14,** 2048–2052
15. Wiemann, S., Arlt, D., Huber, W., Wellenreuther, R., Schleeger, S., Mehrle, A., Bechtel, S., Sauermann, M., Korf, U., Pepperkok, R., Sultmann, H., and Poustka, A. (2004) From ORFeome to biology: a functional genomics pipeline. *Genome Res.* **14,** 2136–2144
16. Collins, J. E., Wright, C. L., Edwards, C. A., Davis, M. P., Grinham, J. A., Cole, C. G., Goward, M. E., Aguado, B., Mallya, M., Mokrab, Y., Huckle, E. J., Beare, D. M., and Dunham, I. (2004) A genome annotation-driven approach to cloning the human ORFeome. *Genome Biol.* **5,** R84
17. Labaer, J., Qiu, Q., Anumanthan, A., Mar, W., Zuo, D., Murthy, T. V., Taycher, H., Halleck, A., Hainsworth, E., Lory, S., and Brizuela, L. (2004) The Pseudomonas aeruginosa PA01 gene collection. *Genome Res.* **14,** 2190–2200
18. Dricot, A., Rual, J. F., Lamesch, P., Bertin, N., Dupuy, D., Hao, T., Lambert, C., Hallez, R., Delroisse, J. M., Vandenhaute, J., Lopez-Goni, I., Moriyon, I., Garcia-Lobo, J. M., Sangari, F. J., Macmillan, A. P., Cutler, S. J., Whatmore, A. M., Bozak, S., Sequerra, R., Doucette-Stamm, L., Vidal, M., Hill, D. E., Letesson, J. J., and De Bolle, X. (2004) Generation of the Brucella melitensis ORFeome version 1.1. *Genome Res.* **14,** 2201–2206
19. Bonaldo, M. F., Bair, T. B., Scheetz, T. E., Snir, E., Akabogu, I., Bair, J. L., Berger, B., Crouch, K., Davis, A., Eyestone, M. E., Keppel, C., Kucaba, T. A., Lebeck, M., Lin, J. L., de Melo, A. I., Rehmann, J., Reiter, R. S., Schaefer, K., Smith, C., Tack, D., Trout, K., Sheffield, V. C., Lin, J. J., Casavant, T. L., and Soares, M. B. (2004) 1274 full-open reading frames of transcripts expressed in the developing mouse nervous system. *Genome Res.* **14,** 2053–2063
20. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469
21. Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R. P., Bitton, F., Caboche, M., Cannoot, B., Chardakov, V., Cognet-Holliger, C., Colot, V., Crowe, M., Darimont, C., Durinck, S., Eickhoff, H., de Longevialle, A. F., Farmer, E. E., Grant, M., Kuiper, M. T., Lehrach, H., Leon, C., Leyva, A., Lundeberg, J., Lurin, C., Moreau, Y., Nietfeld, W., Paz-Ares, J., Reymond, P., Rouze, P., Sandberg, G., Segura, M. D., Serizet, C., Tabrett, A., Taconnat, L., Thareau, V., Van Hummelen, P., Vercruysse, S., Vuylsteke, M., Weingartner, M., Weisbeek, P. J., Wirta, V., Wittink, F. R., Zabeau, M., and Small, I. (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res.* **14,** 2176–2189