Accounting for host cell protein behavior in anion-exchange chromatography

Ryan K. Swanson  
*Iowa State University*

Ruo Xu  
*Iowa State University*

Daniel S. Nettleton  
*Iowa State University, dnett@iastate.edu*

See next page for additional authors
Accounting for host cell protein behavior in anion-exchange chromatography

Abstract
Host cell proteins (HCP) are a problematic set of impurities in downstream processing (DSP) as they behave most similarly to the target protein during separation. Approaching DSP with the knowledge of HCP separation behavior would be beneficial for the production of high purity recombinant biologics. Therefore, this work was aimed at characterizing the separation behavior of complex mixtures of HCP during a commonly used method: anion-exchange chromatography (AEX). An additional goal was to evaluate the performance of a statistical methodology, based on the characterization data, as a tool for predicting protein separation behavior. Aqueous two-phase partitioning followed by two-dimensional electrophoresis provided data on the three physicochemical properties most commonly exploited during DSP for each HCP: pI (isoelectric point), molecular weight, and surface hydrophobicity. The protein separation behaviors of two alternative expression host extracts (corn germ and E. coli) were characterized. A multivariate random forest (MVRF) statistical methodology was then applied to the database of characterized proteins creating a tool for predicting the AEX behavior of a mixture of proteins. The accuracy of the MVRF method was determined by calculating a root mean squared error value for each database. This measure never exceeded a value of 0.045 (fraction of protein populating each of the multiple separation fractions) for AEX. © 2016 American Institute of Chemical Engineers Biotechnol. Prog., 32:1453–1463, 2016

Keywords
two-dimensional electrophoresis, multivariate random forest, anion-exchange chromatography, host cell proteins

Disciplines
Biological Engineering | Molecular, Cellular, and Tissue Engineering

Comments
This is the pre-peer reviewed version of the following article: Swanson, Ryan K., Ruo Xu, Daniel S. Nettleton, and Charles E. Glatz. "Accounting for host cell protein behavior in anion-exchange chromatography." Biotechnology Progress 32, no. 6 (2016): 1453-1463, which has been published in final form at DOI: 10.1002/btpr.2342. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving. Posted with permission.

Authors
Ryan K. Swanson, Ruo Xu, Daniel S. Nettleton, and Charles Glatz

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/cbe_pubs/307
Accounting for host cell protein behavior in

anion-exchange chromatography

Ryan K. Swanson\textsuperscript{a}, Ruo Xu\textsuperscript{b}, Daniel S. Nettleton\textsuperscript{b}, Charles E. Glatz\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a} Department of Chemical and Biological Engineering

\textsuperscript{b} Department of Statistics

Iowa State University, Ames, IA 50011

\textsuperscript{*} Corresponding author at: Department of Chemical and Biological Engineering, Iowa State University, 2114 Sweeney Hall, Ames, IA 50011-2230, USA. Tel.: +1 515 232 1017; fax: +1 515 294 2689. \textit{E-mail address: cglatz@iastate.edu} (C.E. Glatz)

A revised manuscript submitted to Biotechnology Progress

June 28, 2016
Abstract

Host cell proteins (HCP) are a problematic set of impurities in downstream processing (DSP) as they behave most similarly to the target protein during separation. Approaching DSP with the knowledge of HCP separation behavior would be beneficial for the production of high purity recombinant biologics. Therefore, this work was aimed at characterizing the separation behavior of complex mixtures of HCP during a commonly used method: anion-exchange chromatography (AEX). An additional goal was to evaluate the performance of a statistical methodology, based on the characterization data, as a tool for predicting protein separation behavior. Aqueous two-phase partitioning (ATPS) followed by two-dimensional electrophoresis (2DE) provided data on the three physicochemical properties most commonly exploited during DSP for each HCP: pI (isoelectric point), molecular weight, and surface hydrophobicity. The protein separation behaviors of two alternative expression host extracts (corn germ and E. coli) were characterized. A multivariate random forest (MVRF) statistical methodology was then applied to the database of characterized proteins creating a tool for predicting the AEX behavior of a mixture of proteins. The accuracy of the MVRF method was determined by calculating a root mean squared error (RMSE) value for each database. This measure never exceeded a value of 0.045 (fraction of protein populating each of the multiple separation fractions) for AEX.

Keywords: Two-Dimensional Electrophoresis (2DE), Multivariate Random Forest (MVRF), Anion-Exchange Chromatography (AEX), Host Cell Proteins (HCP)
1. Introduction

The production process for biologics runs into a manufacturing bottleneck during downstream purification as a result of a number of factors with two of the more critical ones being upstream titer improvements and increased economic pressure to be first to market [1]. This has led to the majority of production costs being devoted to the downstream purification.

We present a way to evaluate downstream purification alternatives using a multivariate statistical method to predict the separation behavior of host proteins. The method pairs observation of the anion exchange behavior of a large number of proteins with measurement of three molecular characteristics of the same set of proteins. We use measures of size, charge, and surface hydrophobicity for the molecular characteristics. By drawing from two different hosts, we expand the range of characteristics while offering the opportunity to contrast the two sets of host cell proteins (HCP).

Because of immune response concerns, regulatory agencies stipulate the maximum level of HCP allowable in the final product (for example < 100ppm). Hence, methods to quantify and characterize HCP throughout the manufacturing process are important [2-7]. Here we use one of those methods, two-dimensional (2D) electrophoresis, to obtain two protein characteristics – MW (size) and isoelectric point (pI, as an indication of charge) [8-9] as well as to track the elution behavior of individual proteins.

By partitioning the host extracts in an aqueous two-phase system that is sensitive to surface hydrophobicity (SH) and obtaining 2D gels of both phases, we obtain values for the partition coefficients (K) of the individual HCP. Log K has been correlated with other measures
of SH that are not applicable to complex mixtures [10]. A polymer-salt ATPS system was employed where the top phase is rich in polymer and is more hydrophobic than the bottom, salt-rich phase [11-14]. The combination of pl, MW and Log K is our 3D characterization.

This 3D characterization method has been described in greater detail elsewhere [11-14] as well as previously applied to a corn germ HCP mixture (i.e. corn germ extract) [12-13], soybean extract [15], alfalfa extract [16] and a mixture of model proteins [11]. This work expands the database of downstream purification methods and HCP that have been thus characterized.

Anion-exchange chromatography (AEX) offers mild and non-denaturing operating conditions and can achieve an attractive degree of resolution [11-12]. The mechanism by which proteins are retained during ion-exchange chromatography is primarily electrostatic interactions with the positively charged functional group (for AEX applications) immobilized on the surface of the resin. Other properties contribute to the retention behavior of a protein including charge asymmetry [17], molecular shape [18] and surface charge [19] in addition to some nonprotein factors such as the type of resin, sample load [20] and properties of the displacing salt [21]. Of those factors, our measured quantities provide pl (or its difference from pH) as an indication of charge, while MW could indirectly account for charge asymmetry (more possible with larger proteins) and resin type (e.g. pore size) and Log K could address support and spacer interactions.

1.1 Previous Methods
Various studies have sought to predict protein ion-exchange retention behavior [22-25], model elution curves [26-28] or other process characteristics [29-31]. All used model proteins and some a large number of protein descriptors (i.e. physicochemical properties) that come from having a known crystal structure. Prominent are quantitative structure-property (QSPR) or retention (QSRR) relationships for which excellent reviews are available [32-33]. These employ large numbers of descriptors based on protein crystal structure and statistically reduce the original set to a subset of descriptors required for correlation.

The goal of the current study was to develop a statistical method to predict separation behavior while using far fewer properties of a large number of proteins, thereby making it applicable to the large mix of HCP. When applied to IEX elution of a set of model proteins, an earlier statistical method based on only the three properties obtained by the 3D characterization method, provided accuracy comparable to the QSRR approach [11,34].

1.2 Proposed Approach for Correlation of Protein Properties with Separation Behavior

In this situation, with a large number of proteins and a small number of descriptors, we chose to evaluate the multivariate random forest (MVRF) statistical method. MVRF uses a decision tree format where the continuous inputs (i.e. predictor variables) are the three characterization properties (pI, MW, Log K). The outcomes (i.e. response variables) are the amount or fraction of a protein that elutes in each recovery fraction.

The MVRF methodology is an expansion of the classification and regression tree (CART) approach where the data is divided into subsets or nodes using a series of binary (yes or no) questions that involve a randomly selected predictor variable (e.g. Is pI > 6?). These questions
are phrased in order to maximize the degree of homogeneity with respect to the response variables amongst the resulting subnodes. The data is continuously split until a single data point remains, i.e. where a node has no descendants, which is known as the terminal node or leaf. The entire decision structure is a full grown, univariate response tree. A prediction is then obtained by routing a test case (i.e. a data point or protein that was withheld from the set used to grow the tree) through the tree based on the three predictor variables until a terminal node is reached [12,35-40].

The idea of a random forest (RF) involves multiple independent and randomly grown CART where each tree is grown using a different sampling set of data chosen at random using Bootstrap resampling [36]. The RF prediction results from averaging the test case predictions (i.e. terminal nodes) from each tree in the forest. De’ath [39] generalized the univariate CART ideology and created the multivariate response tree where a modified split function was implemented to allow for multiple response variables [37,40-41]. Lastly, Segal [37] developed the idea of the MVRF, which can be viewed as either a progression of the multivariate response tree or univariate-RF ideology. One of the advantages of the MVRF method is the ability to handle multiple predictor and response variables while computing predictions in a straightforward manner once the forest has been completed. There is no explicit prediction equation that can be reported, rather there is a computerized decision tree into which any number of proteins (each represented by its own set of molecular properties) can be entered to generate a predicted elution profile [12,42].
There are numerous other chemometric methods in the literature including partial least squares (PLS) [11,43], principal component analysis (PCA) [44-45], artificial neural network (ANN) [46-47], support vector machines (SVM) [22-23,48-50] and multiple linear regression (MLR) [47,50]. The MVRF methodology was chosen over these other options in part because of the ease in which it can handle data nonlinearities as well as complex response vectors.

1.3 Host cell protein considerations and previous MVRF separation method

*Escherichia coli* (*E. coli*) is a gram-negative bacterium that is commonly selected as an expression host because of the potential for rapid growth rates and high titers, which has led to a detailed understanding of the *E. coli* genome. As a result, *E. coli* has become the “workhorse” of molecular biology, genetics and biotechnology with many proteins now routinely being produced with high yields on the gram/liter scale, justifying inclusion in the current study [51-52].

Plant expression hosts offer unique advantages such as easy scale-up process by planting more acres (when using whole plant systems in the open field versus plant cell cultures in a bioreactor), the ability to perform PTM and the absence of human pathogens or endotoxins [51,53-57]. Corn was selected as the plant expression host for this work since it shares the advantages of other plant systems with the added benefit of having the capability for targeted expression (e.g. to the germ or endosperm fraction of the corn kernel) which can act as a pre-purification step due to the well-established kernel fractionation methods (e.g. dry- or wet-milling) that reduce the amount of residual biomass in the downstream process. There are, however, some disadvantages associated with corn such as low and inconsistent expression
levels, long incubation periods (i.e. long time to harvest) as well as the risk of cross contamination or the inadvertent gene transfer to non-transgenic organisms [55-56,58].

Previously, the separation behavior of a mixture of corn HCP extracted at neutral pH and low salt concentration from the germ-rich fraction (after dry-milling fractionation) was characterized using the 3D characterization method during cation-exchange chromatography (CEX) using a step elution protocol [12]. The results from that study indicated that the MVRF method was able to accurately predict the separation behavior of the individual corn germ HCP using only the three characterization properties (pI, MW, Log K) as inputs. As a result, AEX was selected for characterization while using the same expression host in order to expand the database of downstream methods. An expansion of proteins included was obtained by including *E. coli* extracts. A similar study by Nfor *et al.* applied multiple chromatographic techniques to a crude feedstock for the purposes of obtaining separation properties of a mixture of proteins [59]. An excellent review by Hanke & Ottens (2014) references not only this study but other knowledge-based approaches for chromatographic process development [60].

1.4 Summary

Empirical elution behavior for a large number of HCP was obtained by separately fractionating corn germ and *E. coli* extracts by anion exchange with 2DE of the step fraction eluates to track elution of individual proteins. For the same extracts, 3D characterizations of their individual proteins were obtained. An MVRF predictive tool was developed by using these results. This tool was then tested in a predictive mode by entering the 3D characteristics of a
set of model proteins (not part of the HCP extracts) for which AEX data using the same resin were available in the literature.

2. Materials and methods

2.1 Extract preparation

2.1.1 *Escherichia coli*

The strain of *E. coli* used was BL21(DE3), which was engineered for kanamycin resistance and generously provided by W. Kaar of the Center for Biomolecular Engineering Queensland University (Queensland, Australia). It is an *E. coli* strain representative of one used for protein expression. The *E. coli* cells were cultivated for 3 days while being adapted to minimal medium on minimal agar plates containing 50 mg L⁻¹ kanamycin sulfate by repeated streaking [61]. The resulting colonies were harvested and stored in 40% glycerol at -80°C. A seed culture was prepared by adding 200 mL of minimal medium to a 1 L flask along with 50 mg L⁻¹ kanamycin sulfate, then inoculating with a colony of cells before incubating on an orbital shaker for ca. 19 hours. Next, 400 mL of minimal medium containing 50 mg L⁻¹ kanamycin sulfate was added to each of 5, 2 L flasks before being inoculated with 30 mL of the seed culture broth and incubated again on an orbital shaker. The OD₆₀₀ of the culture broths took 24 hours to reach a value close to 1.0, at which point all 5 flasks were harvested. The harvested broth was centrifuged and after discarding the supernatant, the resulting cell pellet was resuspended in 0.9% (w/v) NaCl before centrifuging and discarding the supernatant as before. The final cell pellet was resuspended in 50 mM sodium phosphate and fed through a bead mill (BeadBeater, BioSpec Products, Bartlesville, OK) for homogenization with 0.1 mm diameter glass beads. The cell
suspension and glass bead mixture (60% glass beads, 40% cell suspension) was poured into the steel chamber of the BeadBeater which was then covered in ice and allowed to cool before being run for 5, 1 min intervals with 1 min cooling time in between. After completion, the homogenate was centrifuged before filtering the supernatant through 0.22 μm syringe tip filters (Corning Incorporated, Corning, NY). Duplicate *E. coli* cultures and extracts were prepared in this way before aliquoting and assaying each for an average total protein concentration of 2 mg mL⁻¹, representing an estimated 25% of the total host proteins.

### 2.1.2 Corn germ

The corn used was a transgenic B73 variety (with a modification, not a factor in this work, of targeted germ expression of Green Fluorescent Protein) provided by M. Paul Scott (USDA-ARS Iowa State University) [62] and fractionated by dry milling at the ISU Center for Crop Utilization Research [63]. The corn germ extract was prepared as before [12] and yielded a total protein concentration of 8.50 mg mL⁻¹ or 20 mg protein extracted/g corn.

The separation, processing and characterization methods described from this point were consistently applied to both *E. coli* and corn germ extracts. This allows for a direct comparison between host cell protein characteristics, prediction of results for the two expression hosts, and pooling of both host’s results for a broader-based database.

### 2.2 Anion-exchange chromatography (AEX)

The anion-exchange resin was Q Sepharose Fast Flow (GE Healthcare, Piscataway, NJ) which was packed into an 8 mL fixed-volume column (Amersham Biosciences, Piscataway, NJ).
The ÄKTA Explorer chromatography system (Amersham Biosciences) was used to perform the separation using the following buffers: equilibration buffer (A) composed of 50 mM sodium phosphate at pH 7.0, elution buffer (B) composed of 1 M NaCl with 50 mM sodium phosphate at pH 7.0, clean-in-place (CIP) buffer composed of 0.1 M NaOH with 1 M NaCl at pH 12.67. In order to avoid overloading the column while at the same time increasing the amount of protein captured in each step elution fraction, duplicate AEX separations were performed on each extract where the only difference was the amount of protein loaded; 10 mL of the *E. coli* extract (25.8 mg of total protein) was loaded onto the column during AEX separation 1 and 22 mL (56.7 mg) during AEX separation 2 where as for the corn germ extract, 5.5 mL (46.75 mg of total protein) and 15 mL (127.53 mg) were loaded for AEX separations 1 and 2, respectively. The AEX step elution protocol with buffer B consisted of: 34% B, 55% B and 100% B with each step run for five column volumes (CV). The flow rate and fraction size were kept constant throughout the separation at 1 mL min\(^{-1}\) and 5 mL, respectively. Five CV of CIP solution and flushing with storage solution (20% v/v ethanol) preceded column storage. The step elution protocol was decided upon after observing the chromatogram that resulted from performing a linear elution separation on both the *E. coli* and corn germ extracts separately while using the same buffers as described above. The amount of 1 M NaCl (i.e % B) was used as the determining factor when “grouping” the peaks together from the two linear elution chromatograms in order to ensure enough protein populated each step elution fraction for both hosts in order for the desired comparison of results to be possible. After both step elution separations, the protein-containing, matching fractions from the two AEX runs for each extract were pooled and assayed for total protein content. Figures 1 and 2 show the chromatograms resulting from both the AEX...
separations for the *E. coli* and corn germ extract, respectively, using the step elution protocol described above. In order to normalize the different loading volumes between the separations, “% AEX Protocol Completed” is shown along the lateral axis instead of volume or time.

2.3 Aqueous two phase system (ATPS) for characterization of surface hydrophobicity

The ATPS composition (15.7% (w/w) polyethylene glycol (PEG) MW = 3350 (Sigma-Aldrich, St. Louis, MO), 8.9% (w/w) sodium sulfate, 3.0% (w/w) NaCl all in 20 mM sodium phosphate at pH 7.0), mass (12 gram ATPS system), separation procedure and Log K calculations (for both HCP and individual model proteins) were the same as previously reported [12] with the only change being the *E. coli* loading ratio of 0.5 mg protein loaded/gram ATPS system. Some protein did precipitate at the interface during partitioning of both expression host extracts, behavior which has been previously observed for corn germ with consistent losses reported (22.63%) [12]. A portion of the proteins appear as interfacial precipitates in ATPS because neither phase is as good a solvent as the extraction buffer. A much larger percentage of *E. coli* extract proteins were observed to precipitate (ca. 80%), but no change was made in the ATPS system so that the log K determination would be consistent with that of the corn germ proteins [14]. The resulting low concentrations partitioning in the two phases did make it necessary to use a more sensitive fluorescent stain for spot quantitation after electrophoresis. The ATPS partitioning experiment was repeated five times for both extracts and average phase ratios of 0.90 ± 0.01 and 0.93 ± 0.01 were calculated for *E. coli* and corn germ, respectively, with 95% confidence intervals shown. The motivation for such a large number of replicates was to maximize what soluble protein mass was available in each phase
solution. Like phase solutions were pooled after separating and a small volume from each phase near the interface was not withdrawn in order to avoid possibly contaminating the top phase with bottom phase protein and vice versa. The fractional losses (precipitate and volume not withdrawn near interface) were assumed to be consistent across all proteins [12].

2.4 Preparation of protein fractions for characterization.

Five E. coli protein fractions result from AEX, and another four samples from the top and bottom ATPS phases along with the two E. coli extracts prepared from separate fermentations. An additional E. coli extract sample (“E. coli extract (ATPS”) was run in parallel with the ATPS separation fraction samples including being stained with the more sensitive fluorescent stain. Seven corn germ protein fractions (5- AEX, 2- ATPS,) result along with one corn germ extract sample.

Tables S1 – S2 (supplementary material) summarize all the sample processing steps undertaken to quantify the individual protein concentrations for each phase of the partitioned extract samples, including any steps, if needed to concentrate (by freeze-drying and reconstitution in 50 mM sodium phosphate at pH 7.0) and de-salt (by dialysis) the protein solutions.

2.5 Two-dimensional electrophoresis

The sample preparation, isoelectric focusing (IEF or 1st dimension separation), SDS-PAGE (2nd dimension separation), imaging and analysis methods were reported earlier [12] with notable deviations described here. Amounts of urea used to dissolve samples and amounts
loaded to the gels differed among samples and are shown in supplementary material section (Tables S1-S2). The alternative staining technique to detect low concentration proteins for E. coli fractions used Flamingo Fluorescent Stain (BioRad). The procedure for the gels stained with fluorescent dye was to soak first in fixative solution (10% acetic acid, 40% ethanol) then in the stain solution overnight in the cold room before washing with water and scanning on a Typhoon scanner (Amersham Biosciences). Four separate groupings (i.e. batches) of finished 2DE stained gels were uploaded into Progenesis SameSpots software (Nonlinear Dynamics, Durham, NC) with two batches from each expression host.

Errors in identifying protein (spot matching and undetectable levels) and assigning mass (intensity values and deviations from the mean recovery values) to those proteins were reduced by using the filters of our earlier work to select those satisfying mass balance and protein content criteria while still retaining a sufficient number of proteins [12]. Filtered subsets of proteins were calculated based on mass balance % (+ 30% of mass balance closure on amount loaded vs. amount eluted) and % of total corn germ extract (> 0.1%) criteria, where those proteins that meet both were put in the intersection subset. A final filter on binding required that a particular protein had ≥ 60% bind to the AEX resin at the loading step.

2.6 Development of multivariate random forest (MVRF)

MVRFs were developed for AEX by the process described in Swanson et al. [12] to relate the three protein properties to their distribution among the five separation fractions. To test the broader applicability of the resulting MVRF, a set of model proteins not present in building the MVRF method but for which literature sources provided properties and separation
outcomes was chosen. For AEX, reported linear gradient retention times for a set of model proteins [22], six for which we were able to obtain Log K values (see Table 1), were used.

The literature values for the AEX gradient elution times were converted to step fractions (also seen in Table 1) bracketing the salt strength of the elution time [12]. The limited availability of reported AEX retention times with existing experimentally determined Log K values resulted in a limited database of model proteins and, coincidently, all six of those eluted in the 34% B step elution fraction after converting from the published linear gradient. The MVRF-predicted elution distribution for these proteins was converted to an estimate of a single elution value calculated by multiplying the predicted % eluted by the average NaCl concentration for each step elution fraction (AEX 0.445 M for 55% B, 0.775 M for 100% B, etc.) resulting in a weighted average value of salt concentration for elution.

3. Results and discussion

3.1 Host Comparison

Collection of sufficient protein in each elution fraction for 3D analysis dictated using step elution for the modeling aspects of this work. Figures 1 and 2 show that material is available in each fraction and that the combination of data from the two hosts provides a set of proteins for MVRF development that cover a wide range of elution behaviors. In addition, potential advantages of correct host selection are evident in Figs. 1 and 2. Recovery from E. coli extracts would be favored for target proteins eluting in the 34% B fraction while target protein eluting in the higher salt fractions would have less contaminating corn germ proteins.
Methods of 3D analysis were adapted for each host in an attempt to carry the greatest number of proteins through the required processing of elution and ATPS phase samples for concentration determination, including changing the stain used for spot quantitation to provide greater sensitivity for the *E. coli* fractions. While this did result in resolving a greater number of proteins for *E. coli* (see Table 2), there was a lower yield of spots (compared to corn germ) that provided quality matches through sample processing because many of the spots were identified by their enhanced presence in elution fractions but had no match in the extract fraction. As a result, mass balance criteria could not be met. That being said, including the larger 1002 protein subset results using 100% of the protein mass could be as useful for the modeling aspect as those generated using the 30 protein subset, which only accounts for 7% of the total mass and covers a much smaller fraction of all dimensions of the protein property space. Elution fraction gels are provided in the supplementary materials (Figure S1), where one can see a general trend of the more basic proteins eluting in the wash or early step with acidic proteins eluting later.

For corn germ, Table 2 shows the higher yield of spots meeting both quality criteria. Of the total of 909 spots, 89, representing 29% of the total protein content of the extract met both criteria. Hence, the reduced subset is potentially more reliable for the modeling aspect (Elution fraction gels are provided as supplementary materials Figure S2). Pairing of subsets for the two hosts provides a still better sampling of proteins for building the MVRF as it provides not only more data but also a wider range of the hydrophobicity measure then either alone.
The scatter plots (Figure 3) indicate the limited range of characterization property values covered in the 24-protein subset. For pI, 19 of the 24 (79.2%) fall between 6 and 8, the same number of proteins have MW values less than 37 kDa and for Log K, 21 of the 24 (87.5%) have values less than -0.319 or the experimental value observed for the corn germ extract. The corn germ extract 2DE gel image from Figure 4 (a) shows that the region contained by MW <37 kDa and pI between 6 and 8 represents a small sampling of characterization property values, i.e. small number of spots in this region.

3.2 MVRF characterization of AEX elution behavior

Table 3 and 4 present the MVRF method prediction results in terms of root mean squared error (RMSE) and variable importance value (VIV) statistics, respectively, for the three protein subsets taken separately for each host and with the combined set. All three subset RMSE values (shown in units of fraction of protein populating each separation fraction) presented in Table 3 were low indicating the ability of the MVRF method to accurately predict a protein’s separation behavior during AEX. Only when corn germ was included do the spot quality criteria improve performance, supporting the earlier conclusion that it is better to include all E. coli proteins if only those data are to be used.

Plots of fitting error as a function of the three protein characteristics are provided as supplementary material (Figures S3: E. coli; Figure S4: Corn Germ). Upon examination of whether there was any trend in the protein characteristics resulting in poorer fitting using the MVRF, only one trend appears. Proteins with higher log K values were less accurately matched, with errors exceeding 25% on a scale of %elution. The likely explanation is that there were
relatively few such proteins in the mix on which the MVRF was developed and they were thus
less influential.

There are alternative characterization methods that have progressed to the point where
quantifying multiple properties of a given HCP is possible (e.g. using 2D electrophoresis or
orthogonal chromatographies, such as reversed phase for SH and chromatofocusing or ion
exchange for a charge measure, coupled with mass spectrometry to replace image analysis).
Utilizing such techniques in the context of the current work might provide the MVRF method
with better data from which to develop predictions by eliminating the errors associated with
spot matching and protein losses during sample processing. This would inevitably lead to a
better overall assessment of the capabilities of the MVRF methodology as it applies to
predicting protein separation behavior.

The VIV statistics represent how important each of the three characterization properties
was in determining a protein’s separation behavior. The VIV value represents the portion of the
elution response that can be described by that protein characteristic. A VIV = 0 would imply the
factor need not be included, while a VIV = 1 would imply the other two factors would not be
needed. A more detailed background, including how the VIV statistics are calculated, can be
found in Swanson et al. (2012) [12]. The prominent role of pI and MW is evident for each MVRF,
but the log K influence only shows up consistently for the combined subset (the strong
influence for the small filtered subset for corn germ may be the result of two small a training
set). The importance of pI and MW is physically reasonable for AEX since departures of pH
from pI would increase net charge and MW could serve as an indirect way of reflecting the
likelihood of charge clusters. The increased importance of log K resulting from the broader set of proteins in the combined set results from the very different distribution of log K values for the two hosts, combined with similar distributions for pI and MW. Hence, only log K can effectively lead to the significantly different elution outcomes for the two hosts (Figures 1 and 2).

3.3 Testing the predictive value of the MVRF with proteins not in the data set

Table 3 represents performance of nine different MVRF, each built on a different set of proteins. On the basis that the combined set of doubly-filtered proteins (54 proteins) serves the dual purpose of incorporating log K and providing low RMSE value, it was this MVRF that was tested for predictive ability on proteins not part of the data set. A set of “model target” proteins (Table 1) for which elution behavior from the same ion exchange resin had been previously published [40] was sent through the MVRF using literature values of pI and MW with log K either measured in this work or taken from earlier work in our lab.

While the experimental elution results were obtained using the same buffers, those experiments were run as gradient elution. Hence, our step elution or their gradient elution had to be put on an equivalent basis. Gradient was converted to step by assuming the protein would elute in the first step fraction with higher salt concentration than in the gradient as the peak eluted. The MVRF predicted distribution was converted to a gradient elution salt concentration by calculating a mass averaged salt concentration over the steps where the protein was detected. Recasting this prediction in order of elution provides a measure less affected by this approximate averaging method.
Table 5 summarizes all of these comparisons. After converting from linear to stepwise elution behavior, all six model proteins were projected to elute in the 34% B separation fraction. The MVRF method predicted that the step fraction to have the highest % elution of each of the six model proteins was in fact the 34% B fraction. Comparison of the predicted order of elution (Table 5) with the literature result (Table 1) shows good correspondence except for the reversal of position for the two eluted last – lipase and β-lactoglobulin A. The prediction of salt concentration at elution was high. The prediction of identical elution patterns for ovalbumin and BSA is a result of the MVRF method when predicting for proteins whose properties are effectively outliers (in this case for log K) relative to the training set, leading to limited possible forests to pass through. The same elution steps were predicted using MVRFs based on the individual host proteins (data not shown) though the orders of elution differed somewhat.

The question remains if the same MVRF method would perform as well with proteins that eluted in a higher or lower fraction but due to a lack of data availability in the literature, this has not been attempted. It was observed that the proteins eluting in the 34% fraction was a higher percentage in the highly filtered corn extract subset (the same was true for the E. coli filtered set, data not shown) as can be seen by comparing Figure 3 with Figure 2. However the MVRF was required to fit a much broader distribution of elution behavior for the HCP. Thus, its usefulness should not be limited to proteins which would elute in this step. Regardless, the prediction accuracy is encouraging since the range of pI and MW values of the six proteins, for the most part, lie outside those of the 54 proteins used to generate the MVRF.
4. Conclusions

Corn germ and *E. coli* HCP recovery profiles differ significantly. This offers an opportunity for matching host selection to the desired protein to simplify separation. When analyzing these profiles, one of the tradeoffs in the value of filtering out lower quality protein matches prior to building the MVRF is the reduction in property range coverage.

For AEX, the MVRF fit the extract sets and prediction of test proteins reasonably well. In addition, the VIV gave understandable physical interpretations for binding modes dependent on charge and charge heterogeneity.

Up to this point, the statistical analysis of the database of characterized expression hosts and purification methods has only been applied to a presumed first stage in the downstream process. The next logical step would be to expand on this idea and statistically analyze multiple permutations of expression host-purification method combinations with the goal of being able to predict the purity and yield of a target protein at the end of a series of separation methods (i.e. downstream process) while at the same time determining which host should be selected based on separation windows. This novel approach would provide the framework for designing a successful downstream process with minimal resources or time spent in the lab.

Acknowledgements

The authors would like to thank Dr. M. Paul Scott from the U.S. Department of Agriculture- Agricultural Research Service as well as Dr. Lawrence Johnson and Steve Fox from
the ISU Center for Crops Utilization Research for producing and supplying the fractionated corn, respectively, and William J. Colonna from the ISU Center for Crops Utilization Research for preparing the glycerol stocks of *E. coli* cells used to inoculate the seed culture minimal medium.

In addition, the authors would like to acknowledge Dr. Steve Cramer, RPI, for his sharing of elution data and sources for model protein comparisons. This work was sponsored by USDA CREES Grants #2008-34496-19348 and #2009-34496-19899. The authors have no conflicts of interest.

5. References


Identification of native *Escherichia coli* BL21 (DE3) proteins that bind to immobilized metal affinity chromatography under high imidazole conditions and use of 2D-DIGE to evaluate contamination pools with respect to recombinant protein expression level, Protein Expres. Purificat. 78(2) (2011) 216-224.


Table 1. AEX model protein dataset along with elution results taken from the literature.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>pI</th>
<th>MW (kDa)</th>
<th>Log $K^b$</th>
<th>Elution [salt]$^f$ (M)</th>
<th>Order of elution</th>
<th>Converted step elution fraction (%B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectin</td>
<td>9</td>
<td>49$^d$</td>
<td>-0.902</td>
<td>0.065</td>
<td>1</td>
<td>34% B</td>
</tr>
<tr>
<td>α-Lactalbumin</td>
<td>4.5</td>
<td>14.18</td>
<td>-1.761</td>
<td>0.088</td>
<td>2</td>
<td>34% B</td>
</tr>
<tr>
<td>Ovalbumin</td>
<td>4.9</td>
<td>44.29</td>
<td>-1.74</td>
<td>0.106</td>
<td>3</td>
<td>34% B</td>
</tr>
<tr>
<td>Bovine serum albumin</td>
<td>4.9$^e$</td>
<td>66.5$^e$</td>
<td>-1.222</td>
<td>0.118</td>
<td>4</td>
<td>34% B</td>
</tr>
<tr>
<td>Lipase</td>
<td>4.81$^c$</td>
<td>29.55$^c$</td>
<td>0.153</td>
<td>0.119</td>
<td>5</td>
<td>34% B</td>
</tr>
<tr>
<td>β-Lactoglobulin A</td>
<td>4.93$^c$</td>
<td>18.36</td>
<td>-1.702</td>
<td>0.148</td>
<td>6</td>
<td>34% B</td>
</tr>
</tbody>
</table>

$^a$ Values taken from Sigma-Aldrich product information sheet unless otherwise noted.


$^c$ Information from ExPASy Proteomics Server of Swiss Institute of Bioinformatics.

$^d$ Howard (1971) [64].

$^e$ Gu and Glatz (2007) [13].

$^f$ Calculated using retention times and elution gradient reported in Tugcu et al. (2003) [22].
Table 2.

Number of protein spots for both expression hosts included in each of the defined subsets for anion exchange.

<table>
<thead>
<tr>
<th>Description of filter or subset</th>
<th>AEX</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # spots (proteins)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>1002 (100%)</td>
<td>909 (100%)</td>
</tr>
<tr>
<td>Corn germ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combinedb,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intersection subseta</td>
<td>73 (18%)</td>
<td>89 (29%)</td>
</tr>
<tr>
<td>Intersection subset &amp; ≥ 60% protein bound</td>
<td>30 (7%)</td>
<td>24 (10%)</td>
</tr>
</tbody>
</table>

a Intersection subset describes the set of proteins where each individual protein accounts for ≥ 0.1% of the total E. coli or corn germ extract protein while also having a mass balance within 30%.

b Combined set of proteins was created by adding together all of the proteins present in both expression host subsets.

c The following property ranges correspond to the three E. coli subsets. 1002 spots: 3.51 - 9.44 pI, 8.13 - 149.67 MW (kDa), -1.329 - +1.383 Log K; 73 spot subset: 3.61 - 8.85 pI, 8.13 - 46.14 MW (kDa), -0.895 - +1.102 Log K; 30 spot subset: 3.61 – 8.45 pI, 8.29 – 41.67 MW (kDa), -0.895 - +0.976 Log K.

d The following property ranges correspond to the three corn germ subsets. 909 spots: 3.53 – 9.54 pI, 8.00 – 91.04 MW (kDa), -1.506 - +0.964 Log K; 89 spot subset: 4.07 – 9.20 pl, 8.62 – 68.35 MW (kDa), -1.097 - +0.381 Log K; 24 spot subset: 4.07 – 9.20 pl, 11.39 – 48.18 MW (kDa), -0.993 - +0.075 Log K.
The following property ranges correspond to the three of the combined host subsets. 1911 spots: 3.51 – 9.54 pl, 8.00 – 149.67 MW (kDa), -1.506 - +1.383 Log K; 162 spot subset: 3.61 – 9.20 pl, 8.13 – 68.35 MW (kDa), -1.097 - +1.102 Log K; 54 spot subset: 3.61 – 9.20 pl, 8.29 – 48.18 MW (kDa), -0.993 - +0.976 Log K.
Table 3.

Root mean squared error (RMSE) values for AEX analyzed for each subset of proteins for both expression hosts.

<table>
<thead>
<tr>
<th>Description of Subset</th>
<th>E. coli</th>
<th>Corn germ</th>
<th>Combined$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All proteins</td>
<td>0.034</td>
<td>0.033</td>
<td>0.041</td>
</tr>
<tr>
<td>Intersection subset$^a$</td>
<td>0.045</td>
<td>0.035</td>
<td>0.044</td>
</tr>
<tr>
<td>Intersection subset &amp; &gt; 60% protein bound</td>
<td>0.030</td>
<td>0.020</td>
<td>0.029</td>
</tr>
</tbody>
</table>

$^a$ Intersection subset describes the set of proteins where each individual protein accounts for > 0.1% of the total E. coli or corn germ extract protein while also having a mass balance within 30%.

$^b$ Combined set of proteins was created by adding together all of the proteins present in both expression host subsets.
Table 4.

Variable importance values (VIV) for AEX analyzed for each subset of proteins for both expression hosts.

<table>
<thead>
<tr>
<th>Description of Subset</th>
<th>E. coli</th>
<th></th>
<th>Corn germ</th>
<th></th>
<th>Combined&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pI</td>
<td>MW</td>
<td>Log K</td>
<td>pI</td>
<td>MW</td>
</tr>
<tr>
<td>All proteins</td>
<td>0.67</td>
<td>0.31</td>
<td>0.01</td>
<td>0.40</td>
<td>0.51</td>
</tr>
<tr>
<td>Intersection subset&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.69</td>
<td>0.27</td>
<td>0.03</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td>Intersection subset &amp; &gt; 60% protein bound</td>
<td>0.63</td>
<td>0.31</td>
<td>0.06</td>
<td>0.00</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<sup>a</sup> Intersection subset describes the set of proteins where each individual protein accounts for > 0.1% of the total E. coli or corn germ extract protein while also having a mass balance within 30%.

<sup>b</sup> Combined set of proteins was created by adding together all of the proteins present in both expression host subsets.
Table 5. Combined database AEX MVRF prediction performance using model protein dataset.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>pI&lt;sup&gt;a&lt;/sup&gt;</th>
<th>MW&lt;sup&gt;a&lt;/sup&gt; (kDa)</th>
<th>Log K&lt;sup&gt;b&lt;/sup&gt;</th>
<th>MVRF predicted elution values&lt;sup&gt;c&lt;/sup&gt; (%)</th>
<th>Single elution value [NaCl] estimate (M)</th>
<th>Single elution value [NaCl] predicted order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Wash</td>
<td>34% B</td>
<td>55% B</td>
</tr>
<tr>
<td>Lectin</td>
<td>9</td>
<td>49&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.902</td>
<td>31.98</td>
<td>60.29&lt;sup&gt;g&lt;/sup&gt;</td>
<td>0.86</td>
</tr>
<tr>
<td>α-Lactalbumin</td>
<td>4.5</td>
<td>14.18</td>
<td>-1.761</td>
<td>29.78</td>
<td>51.57&lt;sup&gt;h&lt;/sup&gt;</td>
<td>4.74</td>
</tr>
<tr>
<td>Ovalbumin</td>
<td>4.9</td>
<td>44.29</td>
<td>-1.74</td>
<td>29.77</td>
<td>50.56&lt;sup&gt;i&lt;/sup&gt;</td>
<td>5.29</td>
</tr>
<tr>
<td>Bovine serum albumin</td>
<td>4.9&lt;sup&gt;e&lt;/sup&gt;</td>
<td>66.5&lt;sup&gt;e&lt;/sup&gt;</td>
<td>-1.222</td>
<td>29.77</td>
<td>50.56&lt;sup&gt;i&lt;/sup&gt;</td>
<td>5.29</td>
</tr>
<tr>
<td>β-Lactoglobulin A</td>
<td>4.93&lt;sup&gt;c&lt;/sup&gt;</td>
<td>18.36</td>
<td>-1.702</td>
<td>28.77</td>
<td>49.62&lt;sup&gt;i&lt;/sup&gt;</td>
<td>5.89</td>
</tr>
<tr>
<td>Lipase</td>
<td>4.81&lt;sup&gt;c&lt;/sup&gt;</td>
<td>29.55&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.153</td>
<td>29.06</td>
<td>30.00</td>
<td>10.37</td>
</tr>
</tbody>
</table>

| % total corn germ extract protein eluted in each fraction<sup>k</sup> | NA | NA | NA | 51.51<sup>i</sup> | 31.17<sup>i</sup> | 2.38<sup>i</sup> | 1.55<sup>i</sup> | 13.01<sup>i</sup> | NA | NA |
| % total E. coli extract protein eluted in each fraction<sup>k</sup> | NA | NA | NA | 54.0<sup>i</sup> | 3.3<sup>i</sup> | 8.4<sup>i</sup> | 12.5<sup>h</sup> | 21.3<sup>i</sup> | NA | NA |

<sup>a</sup> Values taken from Sigma-Aldrich product information sheet unless otherwise noted.


<sup>c</sup> Information from ExPASy Proteomics Server of Swiss Institute of Bioinformatics.

<sup>d</sup> Howard (1971) [64].

<sup>e</sup> Gu and Glatz (2007) [13].

<sup>f</sup> Predicted using MVRF method developed using 54 protein subset.
Using linear elution retention time data (from Tugcu et al. (2003) [22]), underlined value indicates step elution fraction where protein would elute (i.e. mirrors column labeled “Converted step elution fraction (%B)” in Table 3).

Bold values indicate the fraction with the highest predicted % eluted using the selected MVRF method.

Bold and underlined value indicates that the fraction with the highest predicted % eluted matches the fraction where the protein would elute using the linear elution retention time data (from Tugcu et al. (2003) [22]).

The MVRF method predicted ovalbumin and bovine serum albumin to elute at the same [NaCl] (0.272 M) and are therefore both given the same predicted elution order value.

% of total protein extract that was experimentally observed to populate each AEX separation fraction for both E. coli and corn germ.

Values shown are the result of averaging the % of protein that populated each separation fraction from both AEX runs using the chromatograms for each expression host.
List of Figures:

Figure 1. Chromatogram resulting from both AEX separations of *E. coli* extracted at pH 7.0. The y-axis is scaled to mA/10 (milli-absorbance units/10) and corresponds to %B concentration as well. The fraction volumes containing a significant amount of each of the peaks were pooled to represent each separation fraction and to capture each step’s peak. In order to scale between the two separations, “% AEX step elution protocol completed” is shown along the lateral axis instead of volume or retention time.

Figure 2. Chromatogram resulting from both AEX separations of corn germ rich fraction extracted at pH 7.0. The y-axis is scaled to mA/10 (milli-absorbance units/10) and corresponds to %B concentration as well. The fraction volumes containing a significant amount of each of the peaks were pooled to represent each separation fraction and to capture each step’s peak. In order to scale between the two separations, “% AEX step elution protocol completed” is shown along the lateral axis instead of volume or retention time.

Figure 3. Corn germ AEX separation fraction 3D scatter plots: proteins with ≥ 60% total bound to resin are shown (24 proteins); all are ≥ 30% of mass balance and ≥ 0.1% of the total corn germ protein extract (a) Corn germ extract (AEX) where sphere size represents % of total protein in corn germ extract (2/3rd scale); (b) AEX Fraction: wash where sphere size represents % of protein in the wash (1/100th scale); (c) AEX Fraction: 34% B where sphere size represents % of bound protein that eluted in the 34% B step (1/100th scale); (d) AEX Fraction: 55% B where sphere size represents % of bound protein that eluted in the 55% B step (1/10th scale);
(e) AEX Fraction: 100% B where sphere size represents % of bound protein that eluted in the 100% B step (1/10\textsuperscript{th} scale); (f) AEX Fraction: CIP where sphere size represents % of bound protein that eluted in the CIP step (1/10\textsuperscript{th} scale)

Figure 4. Corn germ ATPS separation fraction 2D electrophoresis gel images: (a) Corn germ extract (ASP) with 200 μg protein loaded onto IEF strip; (b) ATPS top phase with 100 μg protein loaded onto IEF strip; (c) ATPS bottom phase with 175 μg protein loaded onto IEF strip. (b) & (c) reproduced from Swanson et al. [12].
Figure 1.
Figure 2.
Figure 3.
Figure 4.
Supplemental Material Section - Tables:

**Table S1**: *E. coli* concentrating and de-salting processing step information along with amounts loaded to first dimension IEF strips of 2D electrophoresis.

<table>
<thead>
<tr>
<th>Separation fractions/samples</th>
<th>Processing step 1: concentration (freeze-dry)(^a)</th>
<th>Processing step 2: de-salt (dialysis)</th>
<th>Processing step 3: concentration (freeze-dry)(^a)</th>
<th>First Dimension IEF strip loading details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Vol. 8 M urea-protein solution (μL)</td>
</tr>
<tr>
<td><em>E. coli</em> extract (ASP)</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>250</td>
</tr>
<tr>
<td><em>E. coli</em> extract (AEX)</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>213.96</td>
</tr>
<tr>
<td>AEX Fraction: wash</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>88.38</td>
</tr>
<tr>
<td>AEX Fraction: 34% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>43.46</td>
</tr>
<tr>
<td>AEX Fraction: 55% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>140.38</td>
</tr>
<tr>
<td>AEX Fraction: 100% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>134.22</td>
</tr>
<tr>
<td>AEX Fraction: CIP (titrated)</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>109.22</td>
</tr>
<tr>
<td>ATPS top phase (Flamingo)(^c)</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>57.3</td>
</tr>
<tr>
<td>ATPS bottom phase (Flamingo)(^c)</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>21.89</td>
</tr>
<tr>
<td><em>E. coli</em> extract (ATPS)</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>15.34</td>
</tr>
</tbody>
</table>

\(^a\) All samples that underwent concentration by freeze-drying (either processing step 1 or 3) were reconstituted after product was dry with 50 mM sodium phosphate at pH 7.0.

\(^b\) Mass loaded per strip is proportional to amount of protein populating each individual sample with the heavily populated samples loading the recommended maximum mass (200 μg) and the less populated samples loading less protein (35 μg, 75 μg, etc.).

\(^c\) *E. coli* sample gels stained with Flamingo Fluorescence Stain (BioRad) and therefore required less protein mass to be loaded onto IEF strip.
**Table S2: Corn germ sample concentrating and de-salting processing step information along with amounts loaded to first dimension of 2D electrophoresis.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn germ extract (ASP)</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>96</td>
<td>404</td>
<td>200 μg (200 μL)</td>
</tr>
<tr>
<td>Corn germ extract (AEX)</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>43.38</td>
<td>456.62</td>
<td>200 μg (200 μL)</td>
</tr>
<tr>
<td>AEX Fraction: wash</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>87.18</td>
<td>412.82</td>
<td>175 μg (200 μL)</td>
</tr>
<tr>
<td>AEX Fraction: 34% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>22.46</td>
<td>477.56</td>
<td>150 μg (200 μL)</td>
</tr>
<tr>
<td>AEX Fraction: 55% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>93.38</td>
<td>406.62</td>
<td>75 μg (200 μL)</td>
</tr>
<tr>
<td>AEX Fraction: 100% B</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>200</td>
<td>300</td>
<td>42.6 μg (200 μL)</td>
</tr>
<tr>
<td>AEX Fraction: CIP (titrated)</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>39.86</td>
<td>460.16</td>
<td>100 μg (200 μL)</td>
</tr>
<tr>
<td>ATPS top phase</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>160.4</td>
<td>339.6</td>
<td>100 μg (200 μL)</td>
</tr>
<tr>
<td>ATPS bottom phase</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>97.2</td>
<td>402.8</td>
<td>175 μg (200 μL)</td>
</tr>
</tbody>
</table>

*a* All samples that underwent concentration by freeze-drying (either processing step 1 or 3) were reconstituted after product was dry with 50 mM sodium phosphate at pH 7.0.

*b* Mass loaded per strip is proportional to amount of protein populating each individual sample with the heavily populated samples loading the maximum mass allowable (200 μg) and the less populated samples loading less protein (42.6 μg, 75 μg, etc.).
Supplemental Material Section - Figures

**Figure S1:** *E. coli* AEX separation fraction 2D electrophoresis gel images: (a) *E. coli* extract (AEX) with 200 μg protein loaded onto IEF strip; (b) AEX Fraction: wash with 175 μg protein loaded onto IEF strip; (c) AEX Fraction: 34% B with 100 μg protein loaded onto IEF strip; (d) AEX Fraction: 55% B with 125 μg protein loaded onto IEF strip; (e) AEX Fraction: 100% B with 75 μg protein loaded onto IEF strip; (f) AEX Fraction: CIP with 125 μg protein loaded onto IEF strip.

**Figure S2:** Corn germ AEX separation fraction 2D electrophoresis gel images: (a) Corn germ extract (AEX) with 200 μg protein loaded onto IEF strip; (b) AEX Fraction: wash with 175 μg protein loaded onto IEF strip; (c) AEX Fraction: 34% B with 150 μg protein loaded onto IEF strip; (d) AEX Fraction: 55% B with 75 μg protein loaded onto IEF strip; (e) AEX Fraction: 100% B with 42.6 μg protein loaded onto IEF strip; (f) AEX Fraction: CIP with 100 μg protein loaded onto IEF strip.

**Figure S3:** *E. coli* AEX separation fraction 3D mesh plots showing the difference between the % of each protein that was observed to elute (Experimental) and the % predicted to elute (Predicted) in each step elution fraction with respect to (a) pI, (b) MW and (c) Log K using the MVRF method developed with the 30 protein subset data.

**Figure S4:** Corn germ AEX separation fraction 3D mesh plots showing the difference between the % of each protein that was observed to elute (Experimental) and the % predicted to elute (Predicted) in each step elution fraction with respect to (a) pI, (b) MW and (c) Log K using the MVRF method developed with the 24 protein subset data.
Figure S1:
Figure S2:
Figure S3:
Figure S4: