Journal of Bioinformatics and Computational Biology

FORECASTING HOST CELLS FOR RECOMBINANT PROTEIN EXPRESSION

HUNG V. LE

ProSci LLC
Rockaway, NJ 07866, USA
hung.v.le@prosciconsulting.com

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Selection of an appropriate host cell is most critical when designing a recombinant protein expression system. The process is mainly based on subjective expertise in consultation with relevant literature, and intuitive decision making. The desired results are seldom instantly guaranteed and expensive laboratory trials are rather the norm. The current study leverages the collective experience in the field since inception in the late 1970's to build evidence-based predictive models for host cell selection. The Protein Database was used as key source of information and logistic regression methods were applied to relate structure-function parameters, stability index, subcellular localization, and post translational modifications, to expression preference for certain host cell types. The resulting models were validated providing a rational approach for forecasting the preference of specific protein sequences for expression in four host cell types: Escherichia coli, insect, mammalian, and yeast cells. Together with instructions they were made available to the public through a web server to facilitate applications and increase success in the laboratory.

Keywords: Recombinant protein, expression, host cells, logistic regression, modeling, prediction.

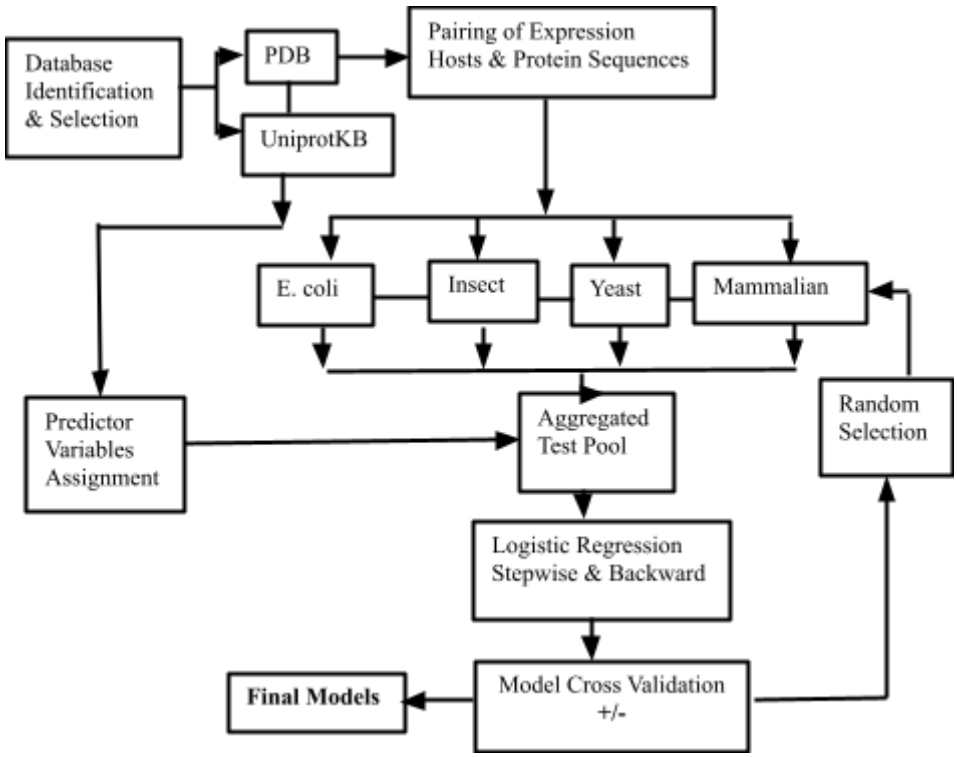
1. Introduction

Since the introduction of recombinant DNA technology[1-3] host cell selection for protein expression was essentially based on personal expertise and trials and errors. With the proliferation of available host cell types, vectors, and commercial reagent kits, some degree of success could always be assured in recombinant protein expression laboratories. Nevertheless, optimized expression to obtain large amounts of soluble active form with desirable post translational features for downstream processing, could be a daunting and costly iterative procedure. Computational tools have been used to address these issues and they were mainly directed toward predicting protein solubility[4-6] or correctly pairing vectors to protein sequences[7] for expression in E. coli. Alternatively, machine learning was applied to predict expression potential of specific sequences in B. subtilis[8]. Others chose translation optimization instead[9,10]. These tools are useful once the expression host has been selected. However, none of them addressed the initial task of pairing expression host to protein sequences, which when done correctly could negate all further needs for optimization. The vast amount of historical data available to date warrants an effort towards devising such a

method to supplement or replace the use of literature search engines and intuitive decision making. Here we seek to leverage the Protein Data Bank[11] (PDB, rcsb.org) with information related to expression in various host cell types. Logistics regression[12] was applied to identify structure function and subcellular localization attributes associated with preferred expression in certain hosts. Logit models were derived and validated for calculation of probability of expression and rank ordering of preference. We discussed the use of this evidence-based statistical approach as an alternative or supplement to traditional ones.

2. Methods

Building the models for predicting optimum pairing between expression hosts and protein sequences was accomplished according to the workflow shown below:



2.1 Building training and test sets.

Sets of protein sequences paired with specific host cells were mainly built from the PDB using its advanced search module. Random selections were accomplished in Excel using

either the Random Sampling or Random Number Generation modules of XLMiner Analysis Toolpak (Microsoft Corporation, Redmond, WA).

2.2 Predictor variable assignments.

Calculation of total number of amino acid residues (r), isoelectric point (pI), Kyte & Doolittle Grand Average Hydropathicity (GRAVY) index[13] (hp) and the instability index[14] (ii) was accomplished by submitting the amino acid sequences of respective PDB entries to the ProtParam tool[15] available from the ExPASy server operated by SBI Swiss Institute of Bioinformatics. Determination of tertiary structural classes was based on annotations of PDB entries when available, or alternatively by direct visualization using the 3DView module in Mol*[16]. When an experimentally determined structure is not available the tertiary structural class was obtained from a predicted structure for the exact protein sequence, or by default the closest homolog or ortholog, obtained from the AlphaFold Protein Structure Database[17] (https://alphafold.ebi.ac.uk/). SCOP classification of tertiary structural classes was applied with minor modifications. Annotations in the PDB and UniprotKB[18] were mainly used for quaternary structural classification and determination of membrane association status and molecular functions. In the absence of adequate annotations, membrane association was determined by prediction of transmembrane helices using TMHMM maintained by DTU Health Tech[19, 20] and Phobius[21, 22] maintained by SBC, Stockholm Bioinformatics Center, by prediction of lipid anchorage sites using GPS-Lipid[23] (S-Palmitoylation, N-Myristoylation, S-Farnesylation and S-Geranylgeranylation) and big-PI Predictor[24] (for GPI anchor), and by prediction of comprehensive subcellular localization using Deeploc[25] maintained by DTU Health Tech. The predicted number of N-glycosylation sites were obtained by submitting sequences, preferably with signal peptides when known, to NetNGlyc-1.0[26].

2.3 Logistic regression and model validation.

All logistic regressions were performed in Excel using the corresponding tool in the XLMiner Analysis Toolpak (Microsoft Corporation, Redmond, WA). Model validation studies were based on cross validation. The average probabilities of expression were calculated for random samplings of 25 sequences each from the training set versus a reserved test set, for each expression host. Analysis of variance was accomplished using the ANOVA module of the XLMiner Analysis Toolpak.

3. Results

3.1 Construction of training and validation test sets

The PDB serves as the primary source of information for identifying protein sequences expressed in various host types. As of this writing it presented 156,140 entries (redundancy included) for proteins expressed in both prokaryotes (87%) and eukaryotes (13%) (Table 1).

Table 1. Distribution of expression host/protein sequence pairs in the PDB

Expression host distribution in PDB*		156,140
Prokaryote expression		135,653
E. coli	134,513	
Others	1,140	
Eukaryote expression		20,487
Insect	9,637	
Mammalian	7,186	
Yeast	3,165	
Fungus	418	
Plant	56	
Protozoa	25	

* The information captured on 09/20/2021 was divided into two large groups of prokaryote and eukaryote hosts, and further subdivided into 2 subgroups of prokaryotes and 6 subgroups of eukaryotes.

Dominance by the former is understandable considering facile growth conditions, ample availability of vectors and cost effectiveness. Among the prokaryotes, variant strains of Escherichia coli (E. coli) were overwhelmingly represented (99%). Insect (47%), mammalian (35%) and yeast (15%) cells dominate eukaryotic expression systems while fungal, plant and protozoan cells together account for less than 2.5%. Considering this distribution, we focused our efforts towards building a database comprised of the larger group of hosts: E. coli, insect, mammalian, and yeast cells. Random samplings in batches of 50 protein sequences for each host cell type were initially performed, then combined for logistic regression in backward mode to retain the most relevant predictor variables. This preliminary training set of ca. 200 protein sequences paired to the four hosts was subsequently enriched iteratively in four additional sampling cycles resulting in stable sets of predictor variables and associated β's. In addition to the training set, small sets of 25 non overlapping sequences were also randomly selected from the four hosts and saved for validation studies.

3.2 Regression parameters consideration.

Protein sequences (the dependent variables) were coded dichotomously (1,0) for positive expression outcome or null in a particular host. The sequences are characterized by 9 independent variables (the predictor variables) associated with structure, stability, subcellular localization, and function. Three continuous variables: the total number of amino acid residues in the sequence (r), the isoelectric point (pI), and the Kyte & Doolittle's Grand Average Hydropathicity (GRAVY) Index[13] (hp), were chosen to provide primary structure information. Five polychotomous variables were included to represent the tertiary structural class (t), quaternary structural class (q), cell membrane association status (m), degree of N-glycosylation (ng), and molecular function (mf). Finally, *in vivo* stability was represented by the instability index (ii) of Guruprasad, Bhasker Reddy & Pandit[14].

The tertiary structural classification was essentially based on SCOP 2[27] (http://scop.mrc-lmb.cam.ac.uk/) and recognized 5 distinct classes: small protein (sp), all

alpha (α), all beta (β), alternating alpha/beta (α/β), and segregated alpha and beta domains (α+β). They were numerically coded 1-5, respectively. Large complex proteins exhibiting multiple domains of any permutations of α, β, and α/β were not assigned a weighted average but assigned code 5 as for (α+β). Likewise, the quaternary structural classes were divided into 5 categories, coded 1-5, respectively: monomer, homopolymer of nth order, homopolymer aggregate of undefined size, heteropolymer of nth order, and heteropolymer aggregate of undefined size. Three states of membrane association were represented: soluble protein with no known membrane association, embedded into cell membrane through nth transmembrane helices, and cell surface associated without transmembrane helices. They were numerically coded 0,1,2, respectively (Table 2).

Table 2. Numerical coding scheme for structural classes and state of membrane association.

Structural and membrane class definition*

Tertiary structural classes (t)	Numerical coding
Small protein (sp)	1
All alpha (α)	2
All beta (β)	3
Alternating alpha/beta (α/β)	4
Segregated alpha and beta (α+β)	5

Quaternary structural classes (q)	Numerical coding
Monomer	1
Homopolymer, nth order	2
Homopolymer, undefined size	3
Heteropolymer. nth order	4
Heteropolymer, undefined size	5

Table 2. (continued)

State of membrane association (m)	Numerical coding
None	0
Transmembrane helices, nth order	1
Membrane associated	2

*Tertiary structural classes were defined according to SCOP 2 with minor modifications. Quaternary structural classes and state of membrane association were derived from annotations in the PDB, Alpha Fold DB and UniprotKB.

Considering that the N-glycosylation status of the proteins selected in this study were not all determined experimentally, we chose instead an index based on the total number of potential sites predicted by the presence of the N-X-S/T sequence motifs[28]. Assignment of molecular function was accomplished according to a classification by the Gene Ontology Project[29, 30], version 2017-09-30. It recognized 15 categories coded as shown in Table 3.

Table 3. Numerical coding scheme for molecular functions.

Molecular function*	Numerical coding
Antioxidant	1
Binding	2
Catalytic	3
Hijacked molecular function	4
Molecular carrier activity	5
Molecular function regulator	6
Molecular transducer activity	7
Nutrient reservoir activity	8
Protein tag	9
Signal transducer activity	10
Structural molecule activity	11
Toxin activity	12
Transcription regulator activity	13
Translation regulator activity	14
Transporter activity	15

*Gene product molecular functions were derived from a classification by the Gene Ontology project, version 2017-09-30.

3.3 Logistic regression and modeling.

The proteins in the database were randomly selected from the PDB with respect to positive expression outcome in the four host cell types chosen for this study. Ultimately, a representative training set of 953 proteins comprising 260, 246, 215 and 229 known to be expressed in E. coli, insect, yeast, and mammalian cells (Appendix A) was used in this analysis. Stepwise logistic regression was carried in a backward manner testing for the significance of inclusion or elimination of predictor variables. The final models include six β coefficients at p values ≤ 0.05 for positive expression outcome in E. coli and yeast cells respectively, whereas seven could be obtained for insect and eight for mammalian cell expressions (Tables 4-7). Each set of predictor variables are distinct yet overlapping.

Table 4. Coefficients and associated statistics* for logistic regression on expression in E. coli.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	0.822	0.443	0.064	2.275	0.954	5.423
r	-0.003	0.001	2.1E-08	0.997	0.996	0.998
pI	-0.128	0.049	0.009	0.880	0.780	0.969
t	0.238	0.064	4.0E-04	1.268	1.118	1.438
q	-0.352	0.063	2.8E-08	0.703	0.621	0.796
ng	-0.209	0.054	1.1E-04	0.811	0.730	0.902
mf	0.048	0.023	0.036	1.049	1.003	1.097

*The model fitted from the data in the table is logit(p) = 0.822 - 0.003r - 0.128pI + 0.238t - 0.352q - 0.209ng + 0.048mf, where p is the probability of expression, r the number of amino acid residues, pI the isoelectric point, t the tertiary structural class,

q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from $\chi 2$ statistics of 149 for sample size N=953.

Table 5. Coefficients and associated statistics* for logistic regression on expression in insect cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-3.204	0.480	2.5E-11	0.041	0.016	0.104
r	0.001	2.0E-04	0.002	1.001	1.000	1.001
hp	1.128	0.323	5.0E-04	3.089	1.641	5.812
ii	0.038	0.008	3.9E-06	1.039	1.022	1.056
t	0.323	0.070	3.9E-06	1.381	1.204	1.583

Table 5. (continued)

m	0.381	0.167	0.023	1.463	1.054	2.031
q	-0.158	0.062	0.011	0.854	0.756	0.965
mf	-0.101	0.030	0.001	0.904	0.852	0.960

*The model fitted from the data in the table **is** logit(p) = -3.204 + 0.001r + 1.128hp + 0.038ii + 0.323t + 0.381m - 0.158q - 0.101mf, where p is the probability of expression, r the number of amino acid residues, hp the Grand Average Hydropathicity (GRAVY) index, ii the instability index, t the tertiary structural class, m the state of membrane association, q the quaternary structural class and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from $\chi 2$ statistics of 84 for sample size N=953.

Table 6. Coefficients and associated statistics* for logistic regression on expression in mammalian cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-1.977	0.567	0.001	0.138	0.046	0.421
r	-0.002	4.0E-04	2.0E-04	0.999	0.998	0.999
pI	0.125	0.054	0.020	1.133	1.020	1.259
ii	0.025	0.008	0.002	1.025	1.009	1.041
t	-0.334	0.072	3.6E-06	0.716	0.621	0.825
m	-0.834	0.282	0.003	0.434	0.250	0.755
q	0.343	0.064	8.0E-08	1.409	1.243	1.598
ng	0.243	0.039	2.8E-10	1.275	1.182	1.375
mf	-0.152	0.039	9.1E-05	0.859	0.797	0.927

*The model fitted from the data in the table is logit(p) = -1.977 - 0.002r + 0.125pI + 0.025ii - 0.334t- 0.834m + 0.343q + 0.243ng - 0.152mf, where p is the probability of expression, r the number of amino acid residues, pI the isoelectric point, ii the instability index, t the tertiary structural class, m the state of membrane association, q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from $\chi 2$ statistics of 162 for sample size N=953.

Table 7. Coefficients and associated statistics* for logistic regression on expression in yeast cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-0.798	0.352	0.023	0.450	0.226	0.898
r	0.002	3.0E-04	5.1E-06	1.002	1.001	1.002
ii	-0.046	0.009	1.1E-07	0.955	0.939	0.972
m	0.392	0.167	0.019	1.480	1.068	2.051
q	0.171	0.065	0.009	1.186	1.045	1.347
ng	-0.083	0.041	0.042	0.920	0.849	0.997
mf	0.096	0.023	1.8E-05	1.101	1.054	1.151

*The model fitted from the data in the table is logit(p) = -0.798 + 0.002r - 0.046ii + 0.392m + 0.171q - 0.083ng + 0.096mf, where p is the probability of expression, r the number of amino acid residues, ii the instability index, m the state of membrane association, q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio. Goodness of fit was ascertained from $\chi 2$ statistics of 108 for sample size N=953.

3.3.1 Common predictor variables.

The common denominators for all 4 expression hosts are the total number of amino acid residues (r), the quaternary structural class (q) and the molecular function (mf). The odd ratios for r were close to 1 (95% CI, 0.997<OR<1.002) for all 4 expression hosts, indicating they do not significantly affect the odds of expression outcome. Interestingly, the odd ratios for q were less than 1 for E. coli (OR=0.703) and insect cells (OR=0.854) but greater than 1 for mammalian (OR=1.409) and yeast (OR=1.186) cell hosts. These results indicate that multiple subunit proteins (homo or hetero) have better odds of expression in these two hosts and lesser when E. coli or insect cells are employed. The fourth common predictor variable mf had a small but measurable effect on expression outcome. Odd ratios for molecular function slightly exceeded 1 for E. coli (OR=1.049) and yeast cells (OR=1.101) and were less than 1 for insect (OR=0.904) and mammalian (OR=0.859) cells. Expression of proteins that are transporters, transcription regulators, or translation regulators, for example, would be marginally favored in yeast and E. coli but penalized in mammalian and insect cells.

3.3.2 Distinctive predictor variables for E. coli expression

The set of predictor variables that was distinct for E. coli expression consists of the isoelectric point (pI, OR=0.780), tertiary structural class (t, OR=1.268), and the predicted number of N-glycosylation sites (ng, OR=0.730). Two out of three predictor variables had odd ratio of less than 1 and thus adversely affect expression when their value trend higher. The odd ratio for tertiary structural class supports a slight increase in preference for the more complex types, e.g. (α/β) and (α+β), for expression in E. coli.

3.3.3 Distinctive predictor variables for insect cell expression.

The distinct set for insect cell expression included the GRAVY index (hp, OR=3.089), instability index (ii, OR=1.039), tertiary structural class (t, OR=1.381), and membrane association status (m, OR=1.463). Three of four indexes had odd ratios greater than 1 and therefore contribute towards higher odds of expression outcome when their values trend higher. Moreover, the results show that the effect is most pronounced with the GRAVY index, indicating that proteins that exhibit higher overall hydrophobicity would have higher odds of positive expression outcome in insect cells.

3.3.4 Distinctive predictor variables for mammalian and yeast cell expression.

Five predictor variables were distinct for mammalian cell expression including the isoelectric point (pI, OR=1.133), the instability index (ii, OR=1.025), the tertiary structural class (t, OR=0.716), membrane association status (m, OR=0.434) and the predicted number of N-glycosylation sites (ng, OR=1.275). Although counterintuitive, the membrane association status negatively impacts expression in mammalian cells. This is likely a reflection of the difficulty growing these cells to the high density required for obtaining practical amounts. Unsurprisingly, a higher degree of N-glycosylation increases the odd of successful expression in mammalian cells. Three predictor variables were distinct for yeast cells, the instability index (ii, OR=0.955), the membrane association status (m, OR=1.480) and the predicted number of N-glycosylation sites (ng, OR=0.920). As in insect cells higher index of membrane association increases the odd of expression in yeast cells. In contrast, a higher degree of predicted N-glycosylation is associated with a small reduction in the odd of expression.

3.4 Model validation.

The four models for recombinant protein expression in E. coli, insect, mammalian, and yeast cells are summarized in Table 8.

Table 8. Summary of logit models for recombinant protein expression in four principal hosts

Expression host	Logit model
E. coli	logit(p) = 0.822 - 0.003r - 0.128pI + 0.238t - 0.352q - 0.209ng + 0.048mf
Insect	logit(p) = -3.204 + 0.001r + 1.128hp + 0.038ii + 0.323t + 0.381m - 0.158q - 0.101mf
Mammalian	logit(p) = -1.977 - 0.002r + 0.125pI + 0.025ii - 0.334t- 0.834m + 0.343q + 0.243ng - 0.152mf
Yeast	logit(p) = -0.798 + 0.002r - 0.046ii + 0.392m + 0.171q - 0.083ng + 0.096mf

All 4 models were subjected to cross validation by comparing the average probability of expression predicted for a subset of 25 proteins in the training set versus the corresponding non-overlapping set selected from the PDB (Appendix B). As shown in Table 9 no difference could be discerned between the two averages for all 4 models, based on the analysis of variance using F-statistics at the common significance level of 0.05.

Table 9. Cross validation of prediction models for recombinant protein expression in four principal hosts.

Expression host	Probability average (Variance)		F-Statistics*	
	Training set*	Test set*	F	P value
E. coli	0.3548 (0.0260)	0.4116 (0.0243)	1.6001	0.2120
Insect	0.3316 (0.0177)	0.2956 (0.0206)	0.8460	0.3623
Mammalian	0.3568 (0.0302)	0.3108 (0.0259)	0.9422	0.3366
Yeast	0.3468 (0.0412)	0.4284 (0.0542)	1.7423	0.1931

*The protein components of the training and test set are described in Appendix B. The probability of expression was calculated using the respective logit model shown in Table 8. Variance analyses were performed in Excel using the single factor ANOVA module of the XLMiner Toolpak.

3.5 Practical applications.

The logit models summarized in Table 8 were used to predict the preference of specific protein sequences for host cell types. Table 9 shows the results of the applications of these equations for 5 sequences selected from recent entries (10/04/2021) in the UniprotKB database: the mating pheromone ER-23 of Euplotes raikovi (P58547), an E. coli Met repressor (B7UNQ8), a rat sodium-dependent glucose transporter (Q80T22), a homoserine O-methyltransferase (A8H6B7) from the marine bacteria Shewanella pealeana and the mitogen-activated protein kinase 1 (P26696) from Xenopus laevis.

Table 10. Probability of expression and preference rank order* for five candidate proteins.

	MER23_ EUPRA	METJ_ ECO27	Mfsd4 Naglt1	metAS Spea	mapk1 mpk1
Expression host					
E. coli	0.50 (1)	0.55 (1)	0.18 (1)	0.35 (1)	0.21 (1)
Insect	0.13 (0.26)	0.03 (0.05)	0.24 (1.39)	0.28(0.80)	0.26 (1.28)
Mammalian	0.31 (0.62)	0.07 (0.12)	0.03 (0.17)	0.17(0.48)	0.19 (0.90)
Yeast	0.11 (0.21)	0.36 (0.65)	0.48 (2.72)	0.14 (0.40)	0.28 (1.37)

Table 10 (continued)

*The probability of expression in a particular host cell was calculated using the respective logit model obtained from Table 8. Preference rank order shown in parentheses was based on the probability of expression normalized to that of E. coli. MER23_EUPRA is the mating pheromone Er-23 of Euplotes raikovi (P58547). METJ_ECO27 is an E. coli Met repressor (B7UNQ8). Mfsd4b Naglt1 is a rat sodium-dependent glucose transporter 1 (Q80T22). metAS Spea_2784 is homoserine O-succinyltransferase from the marine bacteria Shewanelle pealeana (A8H6B7). Mapk1 mpk1 is the mitogen-activated protein kinase 1 from Xenopus laevus (P26696).

They exhibited a range of size, molecular function diversity and species of origin. From the probability of expression and rank ordering each protein sequence shows distinct pattern of preference for expression hosts. Not surprisingly, the Met repressor (METJ_ECO27) and homoserine O-succinyl transferase (metAS Spea_2784) originating from bacterial species had

a clear preference for E. coli hosts. The mating pheromone ER-23 (MER23_EUPRA) although derived from Euplotes, a unicellular eukaryote, was more partial to E. coli than the eukaryote hosts. The rat sodium-dependent glucose transporter (Mfsd4b Naglt1) had clear preference for yeast and to a lesser extent insect cell as hosts. More importantly, the rank ordering also shows that certain proteins could be tolerant for expression in all four hosts. The case in point is Xenopus mapk1 (Mapk1 mpk1), which had low probability of expression in all four hosts and only marginally prefers yeast or insect cells over the other 2 expression hosts. Other protein sequences were more discriminating. For example, the order of preference for the Met repressor (METJ_ECO27) and rat sodium-dependent glucose transporter 1 (Mfsd4b Naglt1) shows 18- and 16-fold difference, respectively, between most and least preferred expression hosts. Such a large difference in preference allows for well informed decision over choice of hosts while taking other factors into consideration.

4. Discussion

In current practice the selection of a host cell for recombinant protein expression is largely driven by laboratory experience, available literature, and practical project needs. Literature search for information related to expression of the sequence of interest (exact or nearest homolog/ortholog identified by BLASTp[31]) remains the principal tool of the selection process. Even with the correct information, there is no guarantee the described expression host and attending vectors are the best possible choices. Practical experience has taught that a few amino acid differences could result in totally different expression outcome[32]. Current computational approaches[4-10] in recombinant protein expression have only addressed tasks that follow the initial choice of an expression host. Here we provide a tool that will rank the probability of expression in four common host types according to experience accumulated over ca. 40 years of practice. The ranking assists in making a rational decision for a host while factoring other considerations, and before proceeding with downstream optimizations. This is a critical step since the correct initial choice of a host could render many optimization steps unnecessary.

The recombinant protein expression literature is vast and scattered among papers covering different subjects. Finding all relevant publications and exploring the common denominators that could match a particular protein sequence to an optimum host cell is a daunting process, if possible, at all. On the other hand, the PDB with more than 150,000 entries to date offers a wealth of information besides the obvious structural aspects. Of relevance is the information on the expression system used to derive the pure proteins for structure determination. A system that can produce adequate amounts for structure determination provides assurance of functionality and practicality. The expression system reported in the PDB although heavily skewed towards E. coli hosts (Table 1) also provided enough entries for the other 3 host types, namely insect, mammalian, and yeast cells. Ultimately, a representative set of 953 entries was successfully constructed by random selection and cumulative addition, resulting in near equal distribution of all 4 host types (Appendix A).

The predictor variables were selected from common structural parameters and those related to subcellular localization, post-translational modifications, *in vivo* stability, and molecular functions. Multiple parameters could in fact be selected from each of these categories, but we prioritized parameters that were suspected by practitioners as potentially impacting expression, and those that are well established and readily obtained from online servers. They included the total number of residues (r), the isoelectric point (pI), the GRAVY index (hp) while ternary and quaternary information were coded by the polychotomous indexes t and q described in Table 2. Subcellular localization was limited to just membrane association and indexed to 3 categories: membrane associated or not, with the membrane associated state subdivided into 2 subcategories: associated via transmembrane helices or not. Membrane association is well known to impact the level of expression and judicious choice of hosts always a prerequisite.

Although multiple post-translational modifications are known[33], only N-glycosylation was selected as representative test case for this category. It is a modification that has been characterized in all 3 domains of life, eukaryotes, prokaryotes, and archaea. It is also known to impact both solubility and stability of the proteins. Since not all N-glycosylation motifs (N-X-S/T) in a sequence are necessarily glycosylated, and these sites have not been established experimentally for all proteins, the predicted number of potential sites[26, 28] was chosen as surrogate, and for uniformity. Assigning and codifying molecular functions were challenging for two reasons. First, the classification offered by the Gene Ontology Project[30] undergoes periodic updates and readjustment. The 2017 version was used as a starting point for this study. Secondly, molecular functions for a protein were assigned from the corresponding entry in UniprotKB, which quite often annotated the entry with multiple functions. Only the core function is recognized in this work. For example, an S-adenosylmethionine transferase would be annotated with 3 molecular functions: ATP binding, Mg binding and methionine adenosyltransferase activity. Its molecular function will simply be reported as "catalytic activity" and coded 3 as shown in Table 3.

Considering the predictor variables selected with the caveats discussed above, stepwise logistic regressions were conducted in backward mode to retain or eliminate variables at each stage based on p values of 0.05 or less. The data set of randomly selected protein candidates among the 4 host cell groups was enriched by batch of 200 at a time, with elimination of redundant identical sequences at each cycle. Further iteration was unnecessary when the data set grew to include 953 protein candidates. A stable set of predictor variables could be obtained with β coefficients at p value ≤ 0.05 as shown in Table 4-7. The unique logit model derived for each expression host (Table 8) was validated by showing that the average predicted probability of expression of a small random subset of sequences in the training set was not significantly different from that of a test set taken randomly from the PDB (Table 9). As a result, these logit models could be effectively used in predicting the probability of

expression, and rank ordering the preference of each protein sequence for a particular host type (Table 10). The preference patterns were clearly distinct and unique for each sequence. The logistic models are in that respect useful tools to deploy as a first step in designing expression systems for novel protein sequences before deploying currently available computational tools[4-10] and proceeding with bench experiments. They supplement the traditional approach and could ease the burden of extensive trials and errors, resulting in time and cost savings.

Future directions would include monitoring the evolution of the recombinant protein expression sample space and resampling as needed for improved data set and logistic models. Exploring the utility of other predictor variables, machine learning techniques, and extending the analysis to other host cell types are also options for development.

5. Conclusion

The present work offers a facile method for predicting the most suitable host cell for expression of novel protein sequences. It was based on regression analyses using a set of 953 protein sequences with known expression outcome in four host cells, E. coli, insect, mammalian, and yeast cells. The expression outcome was linked to predictor variables encompassing structural parameters, instability index, membrane association status, post-translational modifications, and molecular functions. The derived logit models were distinct for each expression host including unique sets of predictor variables. They were validated and shown to be effective in predicting probability of expression of protein sequences in particular hosts, and rank ordering their preferences. Deploying these tools in the initial phase of the design of expression systems could increase chance of success in laboratory trials and reduce associated cost. They are now accessible from the following public server:
https://www.prosciconsulting.com/expred-beta.html

Acknowledgments

Vincent Vinh-Hung, Catherine K. Smith and Andrew Prongay critically read the manuscript and provided valuable input.

References

1. Cohen SN, Chang ACY, Boyer HW, Helling RB, Construction of biologically functional bacterial plasmids in vitro, *Proc Natl Acad Sci USA* **70**:3240–3244, 1973.
2. Chang ACY, Cohen SN, Genome construction between bacterial species in vitro: Replication and expression of Staphylococcus plasmid genes in Escherichia coli, *Proc Natl Acad Sci USA* **71**:1030–1034, 1974.
3. Morrow JF, Cohen SN, Chang ACY, Boyer HW, Goodman HM, Helling RB, Replication and transcription of eukaryotic DNA in Escherichia coli, *Proc Natl Acad Sci USA* **71**:1743–1747, 1974.
4. Chang CCH, Song J, Tey BT, Ramanan RN, Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction, *Brief Bioinform* **15**: 953-962, 2013.

5. Habibi N, Hashim SZM, Norouzi A, Samian MR, A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli, *BMC Bioinform* **15**:134, 2014.
6. Bhandari BK, Gardner PP, Lim CS, Solubility-Weighted Index: fast and accurate prediction of protein solubility, *Bioinform* **36**: 4691–4698, 2020.
7. Chan W-C, Liang P-H, Shih Y-P, Yang U-C, Lin C-C, Hsu C-N, Learning to predict expression efficacy of vectors in recombinant protein production. 2010. *BMC Bioinform* **11(Suppl 1)**: S21, 2010.
8. Martiny H-M, Armenteros JJA, Johansen AR, Salomon J, Nielsen H, Deep protein representations enable recombinant protein expression prediction. *Comput Biol Chem* **95**: 107596, 2021.
9. Bhandari BK, Lim CS, Gardner PP, TISIGNER.com: web services for improving recombinant protein production, *Nucleic Acids Res* **49**: W654–W661, 2021.
10. Bhandari BK, Lim CS, Remus DM, Chen A, van Dolleweerd C, Gardner PP, Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites, *PLoS Comput Biol* **17**: e1009461, 2021.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank, *Nucleic Acids Res* **28**: 235-242, 2000.
12. Hosmer DW, Lemeshow S, Applied Logistic Regression, John Wiley & Sons, New York, 2000.
13. Kyte J, Doolittle RF, A simple method for displaying the hydropathic character of a protein, *J Mol Biol* **157**: 103-132, 1982.
14. Guruprasad K, Bhasker Reddy BV, Pandit MW, Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence, *Protein Eng* **4**: 155-161, 1990.
15. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A, Protein Identification and Analysis Tools on the ExPASy Server, in Walker JM (ed.), *The Proteomics Protocols Handbook*, Humana Press, Totowa, NJ, pp 571-607, 2005.
16. Sehnal D, Bittrich S, Deshpande M, Svobodová R, K. Berka K, V. Bazgier V, S. Velankar S, S.K. Burley SK, J. Koča J, A.S. Rose AJ, Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures, *Nucleic Acids Res* **49(W1)**: W431-W437, 2021.
17. Jumper J et al., Highly accurate protein structure prediction with AlphaFold, *Nature* **596**: 583–589, 2021.
18. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Res* **49**: D480-D489, 2021.
19. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol* **305**:567-580, 2001.
20. Sonnhammer ELL, von Heijne G, Krogh A, A hidden Markov model for predicting transmembrane helices in protein sequences, in Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C (eds.) *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp 175-182, 1998.
21. Käll L, Krogh A, Sonnhammer ELL, A Combined Transmembrane Topology and Signal Peptide Prediction Method, *J Mol Biol* **338**:1027-1036, 2004.
22. Käll L, Krogh A, Sonnhammer ELL, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server*, Nucleic Acids Res* **35**: W429-W432, 2007.
23. Xie Y, Zheng, Y, Li, H et al., GPS-Lipid: a robust tool for the prediction of multiple lipid modification sites, *Sci Rep* **6**: 28249, 2016.

24. Eisenhaber B, Bork P, Eisenhaber F, Prediction of potential GPI-modification sites in proprotein sequences, *J Mol Biol* **292**: 741-758, 1999.
25. Armenteros JJA, Sønderby CK, Sønderby SK, Nielsen H, Winther O, DeepLoc: prediction of protein subcellular localization using deep learning, *Bioinform* **33**: 3387-3395, 2017.
26. Gupta R, Brunak S, Prediction of glycosylation across the human proteome and the correlation to protein function, *Pac Symp Biocomput* **7**:310-322, 2001.
27. Murzin AG, Brenner SE, Hubbard T, Chothia C, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* **247**: 536-540, 1995.
28. Welply JK, Shenbagamurthil P, Lennarz WJ, Naid F, Substrate recognition by oligosaccharyltransferase: Studies on glycosylation of modified asn-x-thr/ser tripeptides, *J Biol Chem* **258**: 11856-11863, 1983.
29. Ashburner et al., Gene ontology: tool for the unification of biology, *Nat Genet* **25**: 25-29, 2000.
30. The Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine, *Nucleic Acids Res* **49**: D325-D334, 2021.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool, *J Mol Biol* **215**: 403-410, 1990.
32. Price II WN et al., Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli, *Microb Inform Exp* **1**: 6, 2011.
33. Huang KY, Su MG, Kao HJ, Hsieh YC, Jhong JH, Cheng KH, Huang HD, Lee TY, dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins, *Nucleic Acids Res* **44**: D435-D446, 2016.

Appendix A

Training sets

Randomly selected components of training sets identified by PDB numbers:

E. coli: 2MMW, 2MNQ, 1IU5, 3I8Z, 2ZW0, 1A8O, 1CTF, 5E11, 2UZG, 2LWB, 1M9Z, 2FYG,1VZI, 2D48, 1M48, 3BES, 2E7A, 1EXT, 1TNW, 3SE3, 1UEA, 2QPO, 1YV1, 1QLX, , 3P92, 1U5K,1OK3,1VK6, 1BR6, 1DCQ, 1L6J, 4ENE, 1QQT, 1SVL, 3FFN, 1MSW, 5TFR, 5B7I, 6YAK, 6YAK, 6YIJ, 6YKG, 6VWC, 7KFL, 6U4U, 6V1A-4, 6V1A-5, 5QOA, 6PLJ, 6PLD, 4QDB, 6Q64, 6Q61, 6U36-2, 6U36_3, 6MVT, 5A1V_1, 5A4J, 5DG3, 6DNV, 5PGI, 6H5V, 5NM8, 4UUS_1, 4UUS_2, 2MJ9, 2MPU, 2KS1_1, 2KS1_2, 2LU3, 3LNV, 4E6A, 4FRF, 4AJ5_1, 4AJ5_2, 4AJ5_3, 3LDN, 3THN, 2QUX, 2P7C, 2IFW, 3H7N, 3GLI, 3GLI_2, 3GLI_3, 3GLI_4, 1M01, 1DVY, 1SAY, 6C71, 6ST4, 6JXV, 6OL8, 6RNM, 5CS6, 5E67, 2DT3, 5DSD, 5E4R, 5DU2, 5E53, 5DTQ, 5DU9, 5DVA, 5DTP, 5E4Q,5DV8,5DTE, 5SA9, 4GQC, 3V5H_1, 3V5H_2, 4IFV_1, 4IFV_2, 4EYW, 4FFN, 4M2X, 5T0W, 3P30, 3ZDV, 3K4Y, 3JUU, 3KF9, 2B4V, 2IXF, 2I82, 3CXZ, 2POD, 2W4Q, 1ZXB, 3B8U, 2KHD, 1YEU_1, 1YEU_2, 1NMS, 1JGS, 1LX8, 1TF5, 1MSF, 1BD9, 1AXM, 1B56, 1J7H, 2R0H, 1G5I, 1JB3, 1X7X_1, 1X7X_2, 2Z8W_1, 2Z8W, 1RXT, 2AVS, 2C95, 2R2C, 2IPI, 3CRQ, 3D79, 2O93, 3CDV, 2LH8, 2XSR, 3HGV, 3KN2, 3NXS, 3NQA, 4DGZ, 2YEE, 4FI3, 4FI3_2, 4FI3_3, 4FOU_1, 4FOU_2, 4K0O, 3TSU, 6PG9, 6AIX, 6AIG, 6U2H_1, 6U2H_2, 6UGP, 5V6H_1, 5V6H_2, 6H5M, 6C1X, 5I5R, 5OO3, 6SOD, 6H2J, 6GV2, 5IAX, 5H11, 4XMD, 6SOR, 6JMD, 6KA2, 6KJ8_1, 6KJ8_2, 6KPH, 5LHM, 6HF4, 6H74_1, 6H74_2, 6EC7, 6EEJ, 5LBM, 6EHV, 5LFA, 5HT6, 5A3O, 5TNB, 6IYY, 5JJM, 6FUA_1, 6FUA_2,

4D4W, 6C37, 5LGW, 3UQA, 4FXV, 4LOB, 2L0I, 3RYI, 3RYI, 3NZ6, 3QS1, 2XBQ, 2VD1, 4GQL, 3U43, 3U43, 3TKP, 4IRP, 3GMV, 2BVZ, 2GB4, 1W6L, 2OMU_1, 2OMU_2, 3ETM, 2V2M, 2FL3, 2WBR, 3GSO_1, 1EI5, 1S2V, 1KAF, 1NIO, 1SNG, 1CKJ, 1KBC.

Insect: 1T50, 5MPO, 2VSD, 1Q8D, 1VYI, 3TJQ, 1Z92, 2HEW, 5MPO, 3SE3, 3N7O, 2ANW, 4H14, 3PE6, 3UG9, 5TIH, 2OU7, 1FO8, 3KJ6, 2V5M, 3SE3, 4GGA, 5UIG, 5X2M, 4O9R, 5X2M, 5T1A, 3NSJ, 4TNB, 4KRO, 1CK7, 1NUF, 3FVY, 2PGG, 5M05, 4KX7, 4XE0, 5UE8, 5ME3, 3S4Z, 6SA5, 6Q7E, 6TOU, 6S6Q, 6JOL, 6XY7, 6LKO, 6VHG,6V2W_2, 6TQ4, 6U8Z, 7CRC_1, 7CRC_2, 7D4F_1, 7D4F_2, 7D4F_3, 6PZD, 4RWS_1, 4RWS_2, 4ZG7, 5CHT, 5D9A_1, 5D9A_2, 5D9A_3, 6DBO_1, 6DBO_2, 5O1O, 5TCD, 6N52, 6C4D, 5AFM_1, 6C7I, 6AKY, 4G56_1, 4G56_2, 4MRO, 4BTJ, 4KC3_2, 4CI8, 3PIX, 4Q7X, 4PXW, 2ZVA, 3B2U_3, 1YAJ, 1ZCB, 1ZRZ, 1JK8_1, 1JK8_2, 1N52_1, 1N52_2, 1AYU, 1IEA_1, 6K42_1, 6K42_2, 6K42_3, 6K42_46K42_56THX, 7AFS, 6QZW, 6V6K, 7JHI, 5E1S, 5DXH_1, 5DXH_2, 6QTQ, 6U5O_1, 6U5O_2, 6GLA, 6CX9_1, 6CX9_2, 5FZF, 5Q5U, 5U7Z, 6ID5_1, 6ID5_2, 6HV0, 6FN6, 5I75, 5T89. 5U81, 4ZL4, 6SJ7_1, 6SJ7_4, 6SJ7_2, 3S8V_1, 4L45, 4EZL, 4FF8, 3TIA, 4Q9Z, 3LXN, 3N8Y, 3LFN, 3L9L_1, 3L9L_2, 3MQE, 2GCD, 2BGR, 2W9Z_1, 2W9Z_2, 3F82, 1W9L, 1N9A_1, 1N9A_2, 1LO6, 1VKG, 1A26, 1INP,1P93, 1FNG_1, 1FNG_2, 1U54, 2ATI, 2C3A, 2BR1, 3I5D, 2AUZ, 4RRV, 4MMX_1, 4MMX_2, 4DBR, 4FOD, 4RJ5, 4GRW_1, 4GRW_2, 4KBA, 4MLP, 4U45, 4WLJ, 4A55_1, 4A55_2, 4OJ2, 5E26. 6PT2, 6KUX, 6AC4, 6BAB, 6BM8, 6BV0, 6MJI_3, 5OSC, 3JBY_1, 3JBY_2, 5FDX, 5GQL, 5CGD, 5EML_1, 5EML_2, 5QCB, 5T1D_1, 5T1D_2, 5T1D_3, 5X93, 6OQL, 6XVK, 6QKK, 6OFY, 6LZ3, 6FIQ, 6GQ7, 4Z4F, 6PQU, 5EAW, 6S1L, 4RLP, 5cT7, 6CBX, 5QR2, 5BYZ, 4YVC, 6MHO, 5G02, 6MYN, 5BQG, 5MU7_1, 5MU7_2, 6ME5, 6ACJ_2, 6ACJ_1, 5QBZ, 4K1K, 5QJ5, 3OPM, 3G4G, 3VS2, 4FQJ_1, 4EHZ, 3J0A, 4LL0, 4ASD, 3OGM_1, 3OGM_2, 2YPT, 2V74, 1ZCA, 2O6S, 2V4B, 2ITV, 2P1O, 1UWY, 1PHZ, 6TT5, 6UP7_3, 6UP7_2.

Yeast: 1SS3, 1TCP, 5HPG, 3S64, 1BXM, 1SM7, 1GD6, 5T89, 2ZIB, 2KVA, 1XX9, 2FIN, 4WFE, 3FJU, 2O6X, 3SI1, 4N2Z, 3ZY2, 1H8L, 3N9K, 3BSG, 1WD3, 2XQR, 5GHK, 4AA1, 2VZ1, 6F7H. 5ED1, 6RW3, 5V5Z, 6SP2, 6Q81, 1E9T, 1EK6,1EQC, 6WHG, 6Z3Y, 7KAQ_1, 7KAQ_2, 7KAQ_3, 7KAQ_4, 7KAQ_5, 7KAQ_6, 7KY8_1, 7KY8_2, 6RB2, 6UCV_1,6UCV_2, 6UCV_3, 6UCV_4, 6UCV_5, 4QVM_1, 4QVM_2, 4QVM_3, 4QVM_4, 4QVM_5, 4QVM_6, 4QVM_7, 4QVM_8, 4QVM_9, 4QVM_10, 4QVM_11, 4QVM_12, 4QVM_13, 4QVM_14, 6FUF_1, 6FUF_2, 5VNY, 5XF8_1, 5XF8_2, 5XF8_3, 5XF8_4, 5XF8_5, 5XF8_6, 5XF8_7, 4ZR1, 6BM4_1, 6BM4_2, 6BM4_3, 6BM4_4, 6BM4_5, 6BM4_6, 6BM4_7, 6BM4_8, 6BM4_9, 6BM4_10, 6T8F, 6LCR_1, 6LCR_2, 6VFF, 4YB9, 5CO6, 5ESF, 6DZ7, 5VLJ_1, 5VLJ_2, 6GSA_1, 6GSA_2, 6GSA_3, 6GSA_4, 5KC2_1, 5KC2_2, 5FUF, 6P2R_1, 6P2R_2, 6DMU, 6I6J_1, 5M8C, 5V7V, 4ZRG, 6ROJ_1, 6ROJ_2, 3QVN, 4I5U, 3PLG, 2WFV, 3H2P, 2XA3, 2C32, 2VU8_1, 2FLP_3, 1X9Q, 1MM0, 1EX0, 1FTZ, 1GFT, 1GYR, 1J14, 1E78, 2DAA, 1BX7, 1CTE, 2IFF, 1KNT, 1DIX, 4WIS, 6PGM, 1QPG, 1QRK, 1CT5, 1A1S, 1LDT, 1A9W_1, 1A9W_2, 1FW8, 1R1J, 1GB5, 1H0K, 1J16, 1I8N, 1CC0_1, 1CC0_2, 1K9Z, 3CQQ, 1ZPU, 3UGX, 3W3X_1, 4K3G, 4AI6, 4A6K, 4EV5, 4C9Q, 6P25_1, 6SPEK, 6ROJ_1, 6E1K, 6FE8_3, 6U8A, 6H5Y, 6AFZ, 5AEZ, 5C1E,

5M5G_1, 5M5G_2, 5I6C, 5IFP, 6S4I, 4ZDY, 5SYT, 6PSY_1, 6PSY_2, 6CX0, 5WFD_2, 5OJS, 2WYT, 2LLI, 4AI6, 3OLE, 3OB8, 4AKJ, 2OKJ, 3FP7, 2IWH, 1GFK, 1OF6, 1N9G, 1OXM, 2WFJ, 3OEE_1, 3OEE_2, 3OEE_3, 3OEE_4, 3OEE_5, 1SJX, 1FMI, 4YBQ, 3O8O, 3O8O, 1DOR, 4A01, 2DKH, 3H8C, 1FBR, 5V6P, 4WJS, 1UL9.
Mammalian: 1ERG, 3T8X,1BQH, 5EN2, 1UEA, 1PEX, 1L0Y, 1SBB, 2ASU, 1KB5, 1FGX,3T8X,5LGJ, 3VI3, 3ZE2, 3S88, 3ZE2, 4XNU, 3V4V, 3HN3, 3VI3, 5NJ3, 4DB1, 4ZXB, 6V7M, 6PE8, 6HGA, 6USC, 1PV7, 6UMX, 1IJQ, 1IJY, 1IMV, 1P53, 1P8J, 6WGB_1, 6WGB_2, 6VWG_1, 7DPM_3, 7DPM_1, 7DPM_2, 7DUO_2, 7DUO_1, 7DUO_3, 7KMG_1, 7KMG_2, 7AH1, 5VSI_1, 5VSI_2, 6CH8_1, 6CH8_4, 6GL7_2, 6GL7_3, 5FMV, 4ZMA, 4YCI, 5HYS_1, 5HYS_2, 5A2E, 5A2F, 3ZHG_2, 4AKM, 3R08_3, 4EWS, 2JJW, 2DRU,2V5N, 3I9G_1, 3I9G_2,1DX5_2, 1T7W, 1N26, 1AU1, 3PP3_1, 3PP3_2, 3T3M_3, 3T3M_4, 4WUU, 5FN8, 5FCS_1, 5FCS_2, 4Z2A, 2UZX_2, 2VSC, 4F2M, 6U2F, 6TT1, 6VRT_1, 6VRT_2, 6XTA, 4R90_1, 4R90_2, 5KU6, 5WNA_1, 5WNA_2, 5KZW, 6IAS_1, 6IAS_2, 6IBT, 6OKP1_1, 6OKP_2, 6OKP_3, 6OKP_4, 6OKP_5, 6OKP_6, 6OML, 6UVO_1, 6UVO_2, 6EYJ, 5ANM_1, 5ANM_2, 5ANM_3, 5VH4_1, 5VH4_2, 5VH5, 4JDA, 2XFK, 4OGA_5, 3KWZ, 3REZ, 3RJR, 2NYY, 3E6P, 3HH2_1, 3HH2_2, 3EDY, 11N8Z, 1OYH, 1RF0_1, 1RF0_2, 1RF0_3, 1I8L_1, 1I8L_2,1J89, 1F42, 1PHM, 1CKL, 1CDU, 1EWF, 1OLZ, 1O86, 1M4U, 1JQF, 2F61, 2VXQ_2, 2VXQ_3, 2WNG, 1ZKZ, 2AM4, 1YIP, 1W0Y, 2V10, 2CF9, 6EHO, 4PO7, 3RM9, 2XMD, 2LDB_3, 2LDB_2, 4BDV_1, 4BDV_2, 4F5C_1, 4F5C_2, 4CSY, 3U7U, 5BO1_2, 5BO1_3, 5ALC_1, 5ALC_2, 5JNC, 5K33_1, 5K33_2, 6GSI_2, 5U6A_1, 5U6A_2, 5VR9_1, 5VR9_2, 5VQP, 6OMO, 6QB3_2, 6UVO_1, 6UVO_2, 6VRT_1, 6VRT_2, 6SRX_2, 6SRX_1, 6EDU_3, 3O9M, 2HWL, 1FRT_1, 1FRT_2, 6O9H, 6MF0, 3K71, 1SZH, 1MMP, 4LEO_3, 6LIQ_1, 1OQO_1, 1OQO_2, 5LY6, 2VNN, 1LTJ_3, 3NK3, 5VH4_1, 5VH4_2, 6MNF_2, 6MNF_1, 4BC0, 5XJE_2, 5XJE_1, 2V11, 6PYC_1, 6PYC_2, 2X8B, 5DTF_1, 5DTF_2, 3N2Z, 6DJP_1, 6DJP_2, 6SS2, 6U2F, 6U6U, 6UE7_1, 6UE7_2, 6UE7_3.
Randomly selected components of training sets identified by UniprotKB numbers:
E. coli: Q75NH7, P09603, P03069, P07951, P45379-6, P00797, P01009, P00970, P04535, P09922.
Mammalian: P60568, Q90495, P062133V4V.

Appendix B
Validation sets
Randomly selected components of training subsets spared for cross validation studies
E coli: Q75NH7, 1TNW, P03069, 5A4J, 2KS1_2, 3H7N, 3GLI_4, 3ZDV, 2POD, 1ZXB, 1B56, 1J7H, 1X7X_2, 2IPI, 3NQA, 4DGZ, 4FI3_1, 5V6H_1, 5IAX, 5H11, 6C37, 3QS1, 3U43, 4GQL, 1S2V.
Insect: 1Q8D, 4H14, 3UG9, 6Q7E, 5CHT, 4PXW, 1ZCB, 6THX, 6U5O_1, 6CX9_2, 5FZF, 3L9L_1, 1LO6, 1INP, 4RJ5, 4GRW_1, 4KBA, 3JBY_2, 5FDX, 6S1L, 4YVC, 6ACJ_2, 3OPM, 4LL0, 2O6S.

Yeast: 5HPG, 1XX9, 4WFE, 3BSG, 2XQR, 6F7H, 7KY8_2, 6UCV_2, 4QVM_6, 4QVM_13, 6BM4_1, 6BM4_7, 6BM4_8, 6BM4_9, 6VFF, 5VLJ_1, 6GSA_2, 5V7V, 3H2P, 1FTZ, 1GFT, 1E78, 1CT5, 1A1S, 1FW8.
Mammalian: 3T8X, Q90495, P06213, 7DPM_2, 3I9G_2, 3T3M_3, 3T3M_4, 6UVO_1, 5ANM_3, 1RF0_1, 1RF0_2, 1RF0_3, 4F5C_1, 4F5C_2, 5BO1_3, 6GSI_2, 5U6A_2, 5VR9_15VQP, 6VRT_2, 1OQO_1, 1OQO_2, 6MNF_1, 5DTF_2, 6U2F.
Randomly selected components of test sets selected from the PDB for cross validation studies
E. coli: 1P09, 1NCT, 1H7Q, 1ZUH, 4XZD, 1QSA, 1IJG, 2E7J, 2LX3, 4ZXS_1, 4ZXS_2, 1K0K, 2G0R, 5D45, 1K6O_3, 1K6O_4, 2HBT, 3LGW, 5ET0_1, 5ET0_2, 1L3L, 2PVG_1, 2PVG_2, 2PVG_3, 3MA9_1.
Insect: 1Q5K, 2YIY, 4MYW, 4Y5X_1, 4Y5X_2, 4Y5X_3, 6BKO_1, 6BKO_2, 6E7R_1, 6E7R_2, 6P7V_1, 6P7V_2, 6P7V_3, 6QWS, 7KN6_1, 7KNT_1, 7KNT_2, 2H2T, 3EU7, 3K7W, 4IB4, 4TYE, 4UXI, 6CZS, 5JFW.
Yeast: 1B7N, 1BX8, 1EAX, 1GAZ, 1H6M, 1I3D, 1UXL, 2E3F, 2J7O, 3L4K, 4BKE, 4FM9, 4QV3, 4W8F, 6ND1, 6QG3_1, 6QG3_2, 6QG3_3, 6QG3_4, 6QG3_5, 6QG3_6, 6QG3_7, 6QG3_8, 6T9J, 7OP1.
Mammalian: 1OAK_1, 1OAK-2, 1JMY, 3QRG_1, 3QRG_2, 4C8F, 2VXT_1, 2VXT_2, 2VDP_2, 4FTN, 2W0F_1, 2W0F_2, 3U9U_3, 4J1U_1, 4J1U_2, 4X96, 3QG7_1, 3QG7_2, 5EGH, 4X4M_2, 5L7I, 4LU5_2, 4LU5_3, 5UJZ_3, 6HYG_1.

Hung V. Le is currently Principal Consultant and Owner of ProSci LLC. Previously, he was Site Lead, Protein Science, Merck Research Laboratories, Cambridge, MA (2007-2010). In the Chemical Research Division of Schering-Plough Corporation from 1980-2006, he held positions of increasing responsibility from Senior Scientist in Protein Chemistry to Distinguished Fellow in Protein Engineering. From 1978-1980 he was Postdoctoral Fellow in Viral Oncology at the Roswell Park Memorial Institute. His PhD (1978) and BSc (1973) degrees were in Biochemical Pharmacology from the State University of New York at Buffalo.