Chapter 2

Designing Experiments for High-Throughput Protein Expression

Stephen P. Chambers and Susanne E. Swalley

Summary

The advent of high-throughput protein production and the vast amount of data it is capable of generating has created both new opportunities and problems. Automation and miniaturization allow experimentation to be performed more efficiently, justifying the cost involved in establishing a high-throughput platform. These changes have also magnified the need for effective statistical methods to identify trends and relationships in the data. The application of quantitative management tools to this process provides the means of ensuring maximum efficiency and productivity.

Key words: Protein expression optimization; Quantitative analysis; Experimental design; Screening; Statistics

2.1. Introduction

The amount of protein, particularly when recombinant, is more often described in qualitative than quantitative terms. Frequently, proteins are visualized on a gel to characterize amount and purity. The once-mandatory protein purification tables, describing protein production efficiencies, are now rarely found in publications. Protein yields are frequently described subjectively as estimates or percentages. This over-reliance on qualitative measurement reflects the difficulties encountered in accurately determining amounts of protein. The problem is only aggravated when working in a high-throughput protein production environment. This bottleneck in generating quantitative data has now effectively been removed with the development of the LabChip®90 protein assay system capable of analyzing >288 samples (three

Sharon A. Doyle (ed.), *Methods in Molecular Biology: High Throughput Protein Expression and Purification, vol. 498*

Book doi: 10.1007/978-1-59745-196-3

96-well plates) per chip priming *(1)*. When integrated into a high-throughput protein production platform, a vast amount of data can be generated, thereby, requiring effective quantitative and statistical methods to identify trends and relationships. In this chapter, we attempt to illustrate some of the advantages in using experimental design and show how, when combined with high-throughput expression, it can be used to optimize protein production. We encourage readers who are interested in this area to consult further references for a more detailed introduction to statistical analysis in experimentation *(2)*.

2.1.1. Design of Experiments

Statistical design of experiments (DOE), or simply experimental design, is a proven technique used extensively today in many industrial-manufacturing processes. Considering that this method was originally conceived to identify genetic variation in crops, it has not, until recently, been widely taken up by life scientists. As more research disciplines are using automation and microfluidics to obtain faster results, an increasing number of scientists are now recognizing the assistance that experimental design can provide. Consequently, this technique is finding increasing acceptance in many areas beyond its origins in genetics.

Among the advantages that DOE can provide is the increased amount of information per experiment compared to an ad hoc approach. The second benefit occurs in providing an organized approach toward analysis and interpretation of results, thus facilitating communication. Another advantage is the ability to identify interactions among factors, leading to more reliable prediction of response in areas not directly covered by experimentation. The fourth benefit is in the assessment of information reliability in light of experimental and analytical variation. The uptake of this mathematical technique has been greatly aided by the availability of DOE software packages, like JMP (*see* **Note 1**), making it accessible to the nonstatistician.

2.1.2. Optimization of Protein Production

Optimization of protein production using a conventional one-factor-at-a-time approach is a very labor-intensive endeavor, due to the large number of potential factors and their interactions that can affect expression. Interactions make it difficult to optimize factors independently, increasing the number of experiments required to cover the variable space to identify the maximum response. Through DOE techniques the total number of experiments can be reduced, by evaluating the more relevant interactions among variables, and through the use of partial factorial experimental models. Even then, however, the throughput of traditional protein expression is insufficient to perform the required number of experiments in a reasonable period of time and at a viable cost. Only now through the recent development of high-throughput protein expression platforms is it possible to take full advantage of DOE optimization of protein production.

2.2. Methods

2.2.1. Experimental Design

Good experimentation requires the establishment of a precise goal and objective; an ill-defined experiment will often produce ambiguous results and fail to reach any conclusion. The simplest experimental design is one where screening is used to identify key factors affecting a measurable response (*see* **Note 2**). In our case the response to be maximized is soluble protein production. Utilizing the high-throughput platform described in **Chapter 10** enables the analysis of soluble protein produced in *E. coli* and insect cells. Analysis of this quantitative response allows the experimenter to identify and optimize conditions critical to production of soluble protein.

In order to express a protein, many factors need to be examined experimentally, as it is difficult to know a *priori* what will succeed. Performing one-factor-at-a-time experiments (**Fig. 2.1a**), especially when there are many potential important factors, raises the risk of locating a local maximum, thereby missing the actual best condition.

Also, experiments are best executed in an iterative manner so that information learned in one experiment can be applied to the next. Typically a sequence of experiments is used to meet a defined objective. The experiments include screening designs based on a fractional factorial (**Fig. 2.1b**) to identify signifi-

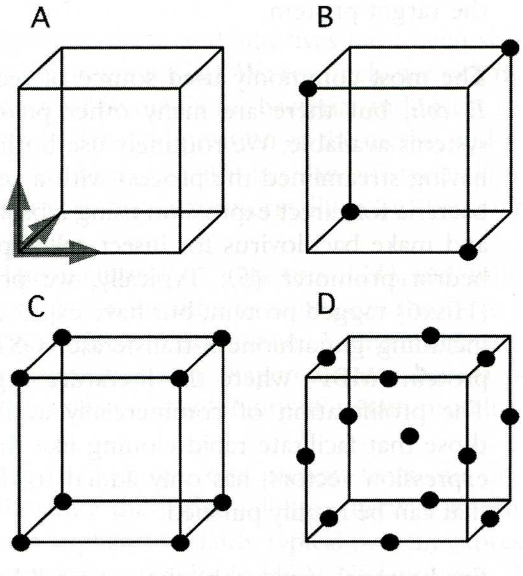

Fig. 2.1. Experimental design. (a) One factor at a time; (b) fractional factorial; (c) full factorial; (d) response surface model: Box-Behnken design for three factors.

cant factors, a full factorial (**Fig. 2.1c**) to identify interactions or response surface design (**Fig. 2.1d**) to fully characterize or model the effects, followed up with confirmation runs to verify the results (*see* **Note 3**).

2.2.2. Factors Affecting Expression

A factor is any variable associated with the product or process under experimental control. There are two different types of variables used in DOE: categorical and continuous. A categorical variable is a factor having only a discrete number of settings that have no intrinsic order, while a continuous variable can be assigned a numeric value. A number of factors affect recombinant protein expression including, but not limited to, construct length, vector, cell line, temperature, time, media, inducer concentration, and additives. The range of values used should be based on either literature precedent or previous experience expressing proteins (*see* **Note 4**). We will briefly discuss these factors and the approach we take to each expression system.

2.2.2.1. Construct

Once a target protein of interest is chosen, the first step is to design a number of constructs of varying length, as practical experience has shown that the exact construct limits can be critical to success. Alignments with homologous proteins that have been previously expressed can help limit the number of constructs, but it is unwise to choose only one. Limited proteolysis *(3)* or H/D exchange *(4)* of full-length protein can also be used to identify small, stable domains capable of being overexpressed and successfully used in structural studies downstream. Additional diversity can also be introduced into the experiment by exploring mutants and homologs of the target protein.

2.2.2.2. Expression System and Vector

The most commonly used source of recombinant expression is *E. coli*, but there are many other prokaryotic and eukaryotic systems available. We routinely use both *E. coli* and insect cells, having streamlined the process with a vector that can transform bacteria for direct expression using a bacteriophage T7 promoter and make baculovirus for insect cell expression using the polyhedrin promoter *(5)*. Typically, we produce a hexa-histidine (Hisx6)-tagged protein, but have explored other fusions options including glutathione-S-transferase (GST) and maltose-binding protein (MBP) where the literature suggests some advantage. The proliferation of commercially available vectors, especially those that facilitate rapid cloning like the Gateway™ destination expression vectors, has only added to the number of strategies that can be readily pursued.

2.2.2.3. Cell Line

For bacterial expression, choice of cell line can greatly affect the amounts of protein produced. There are a number of *E. coli* strains with genotypes engineered specifically to meet the needs of expressing recombinant proteins. While there is a wide variety of choice, the popularity of commercially available competent cells has reduced the number of cell lines more often used to just a few. Some of the most frequently used are derivatives of the BL21(DE3) cell line *(6)*, containing coresident plasmids to address specific protein expression issues, including toxicity, codon bias, and folding. For insect cell expression the choice of cell lines is smaller with Sf9, Sf21, and High-5 being the most commonly used. Despite the limited choice we have found, as in *E. coli*, proteins will have a distinct preference and it is worthwhile examining expression in as many cell lines as possible.

2.2.2.4. Temperature and Time

Temperature and time are frequently critical factors, especially since these two variables often interact. In bacteria, there are some proteins that benefit greatly from a slower, longer induction, which generally requires low temperature *(7)*. At high temperatures, bacterial cells will reach a maximum density and eventually run out of nutrients, at which point cell death will occur. If the protein of interest aggregates easily and cannot be overexpressed in a short time frame, then lowering the temperature is essential. We have expressed proteins anywhere from 15 to 37°C and 3–24 h. Insect cells are less tolerant of temperature variation, so we only examine expression at a single temperature (27°C). In both *E. coli* and insect cells the time of induction or infection, triggering the onset of expression, can also play a role in protein productivity. Induction of expression early or late in growth phase and its intensity directed either by IPTG *(8)* or multiplicity of infection *(9)* have been shown to influence protein levels.

2.2.2.5. Media

Specific media and additives have been shown to have an effect on expression ***(10)***. We routinely use rich media (*see* **Note 5**) and serum-free media for bacterial and insect cell expression, respectively. The composition of the media, and whether or not it contains serum, can also have an effect on expression, though we do not vary this factor normally in our basic screens.

2.2.2.6. Additives

The inclusion of cofactors *(11)* and inhibitors *(12)* into the expression medium has also been shown to affect levels of recombinant protein expression. Additionally, the coexpression of partner proteins and chaperones can have a positive effect on the expression and solubility of certain proteins *(13)*.

2.2.3. Full Factorial Design

We have chosen one protein from our production portfolio to illustrate the various designs used in optimization. This process was applied to a fairly typical protein expression experiment: the expression of soluble HCV NS3 protease domains (NS3-prt) in *E. coli*. The objective was to identify the significant factors and interactions involved in maximizing soluble expression. Having this goal clearly established, the experimental design can then be chosen. Our choice of design for this problem was a full factorial. This selection was based on our previous experience producing this poorly expressed protease and the identification of a small number of factors, including genotype, capable of influencing its soluble production. The soluble expression of six different NS3-prt genotypes in total was examined using four factors: three continuous (temperature, time, IPTG concentration) and one categorical (cell line). Both nominal and discrete variables were examined at two levels, high (+) and low (−), resulting in a 2^4 full factorial design. A total of 16 conditions per construct were examined with each condition being tested in triplicate. A full factorial experiment containing all possible combinations of factors represents not only the most conservative approach, but also the most costly in terms of experimental resources. As mentioned before, the availability of DOE software with custom design capability greatly facilitates this process for the nonstatistician. The JMP DOE software will determine how large a sample size is needed to identify a significant effect (*see* **Note 6**), guard against uncontrolled (or unknown) variables during execution of the experiment through randomization (*see* **Note 7**), and introduce blocking (*see* **Note 8**) when appropriate.

The amount of protein expressed (the response) quantified by the Caliper LabChip 90 system was transferred into DOE analysis software (*see* **Note 9**). The expression data when shown graphically (**Fig. 2.2**) readily illustrate that the categorical factor BL21(DE3) pLysS has a negative effect on the levels of soluble protein expression. Subsequent multiple regression analysis (*see* **Note 10**) of the data generated by the most constructs in BL21(DE3) identified the significant factors conducive to soluble expression, lower temperature (22°C), and shorter induction period (3 h), while IPTG concentration was not significant over the range examined (*see* **Note 11**). Relationships between factors are readily exposed in an interaction plot, with nonparallel lines produced by the interactive plot of NS3-prt (1b) L13K expression demonstrating that the effect of temperature is highly dependent on time (**Fig. 2.3a**). The level of expression over the time of induction, which had previously appeared to have little significant effect on the level of soluble expression of NS3-prt (1b) L13K, diverges widely at higher values of temperature. The interaction of time with temperature tended to mask the effect of time as a main effect.

2.2.4. Fractional Factorial Design

Unlike the example we have just used, experiments are often initiated knowing very little about what factors influence the expression of a particular protein. In such situations the preference would be to examine as many factors as possible. A large screen-

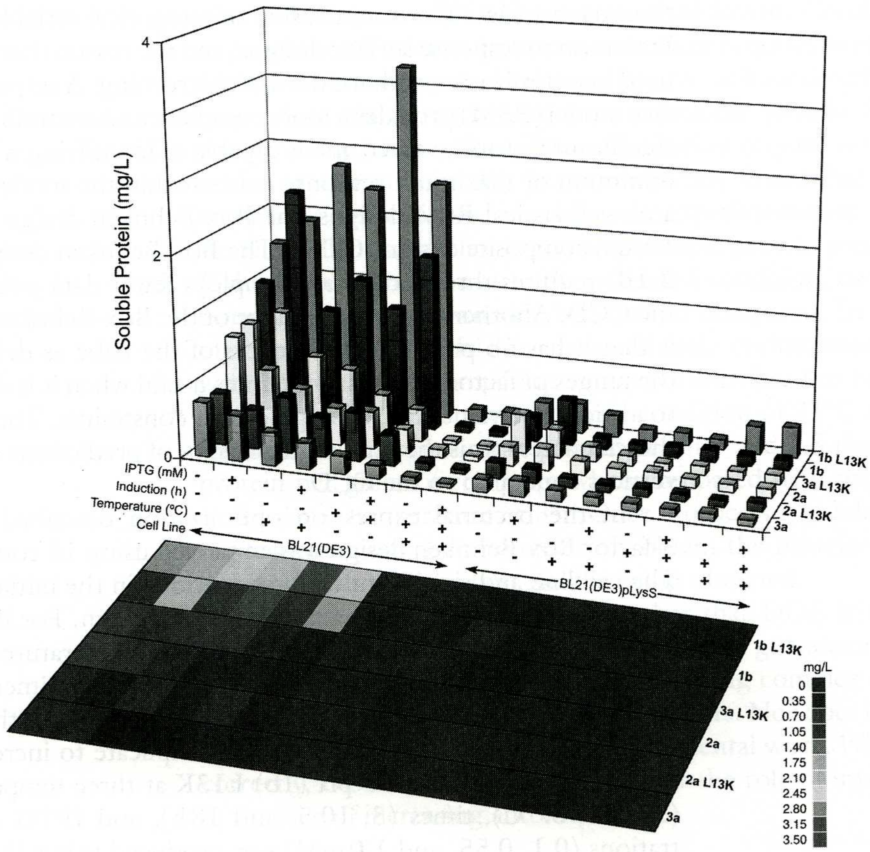

Fig. 2.2. Full factorial screening for effects on the soluble HCV NS3-prt expression in *E. coli*. Levels of expression are illustrated in a 3D Bar Chart and a Heat Map. Using a previously described high-throughput protein expression platform, each data point was obtained using the HT Protein Express 200 Chip run on the Caliper labchip 90 and measured in triplicate (*See Color Plate 1*).

ing approach is best accommodated using a fractional factorial experimental design, whereby the number of potential variables is reduced to a few effective ones. In this model, a partial combination of factors is capable of exploring the maximum number of variables, while requiring less experimentation, albeit at the cost of losing some information about possible interactions. Another consequence of using certain fractional factorial designs, particularly ones with low resolution, is effect aliasing (or confounding). This is where two or more variables have been changed at the same time in the same way resulting in their effects being aliased. This problem can be avoided using a 2-level full factorial or a higher resolution fractional.

2.2.5. Response Surface Designs

Once a process is close to optimum a response surface design can be used to fine-tune the conditions. Response surface designs are used to model the response of a curved surface to a range of continuous variables. The noninclusion of categorical variables is one limitation to response surface designs, and the reason that they are used in optimization and not the initial screening. A response surface model (RSM) provides a more complete understanding of the significant factors involved and is capable of identifying whether a minimum or maximum response exists within the model. There are two classical RSM designs, the Box-Behnken design and the central composite design (CCD). The Box-Behnken design (**Fig. 2.1d**) requires three factors and employs fewer data points than the CCD. Another important feature of the Box-Behnken design is that it has no points at the vertices of the cube as defined by the ranges of factors. This is sometimes useful when it is desirable to avoid these values due to engineering constraints. The cost of this characteristic is the higher uncertainty of predictions near the vertices compared to the CCD.

In the bacterial expression optimization described here a 3-factor Box-Behnken design was employed using 15 conditions. The cell line and significant factors identified in the initial screen were then applied to the customized RSM design. For instance, a protein with a strong preference for low temperature will be screened at lower temperatures in the RSM experiment. The design includes three center points, used to estimate the error of the process; each condition is run in triplicate to increase the accuracy. An RSM of NS3-prt (1b) L13K at three temperatures (21, 29, 37°C), times (3, 10.5, and 18 h), and IPTG concentrations (0.1, 0.55, and 1.0 mM) was produced using JMP software. The resultant RSM confirms the previous observation with expression peaking at high time and low temperature, and temperature being the most significant factor (**Fig. 2.3b**).

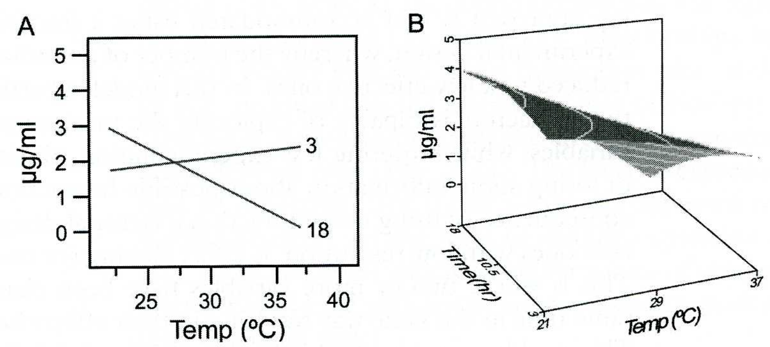

Fig. 2.3. (a) Interaction plot for temperature and time affects on the soluble expression of HCV NS3-prt (1b) L13K. The dependence of protein yield (y-axis) on temperature (x-axis) is plotted for two different times. The red line represents the 3-h data, while the purple line represents the 18-h data. (b) Response surface plot demonstrating expression as a function of time and temperature. The effect of time and temperature on protein yield at an optimal value of IPTG (0.78 mm) is depicted as a 3D surface using the statistical program JMP. The top surface is *colored purple* and the bottom is *colored gray*. Mapped to the surface in *blue* is the contour map of the same data (*See Color Plate 2*).

2.2.6. Validation

The final step in any DOE process is validation. Confirmation of conclusion(s) drawn by experimentation requires verification. Since the ultimate goal of the experiments we have described was to produce the maximum amount of soluble protein from liters of culture, the results of the optimization require verification at this greater volume. We have consistently found that optimal conditions determined by small-scale experimentation demonstrate excellent scalability in terms of soluble protein production. If there is any question as to the preferred condition, two conditions are chosen, grown side by side and compared. In the case of the NS3-prt (1b) L13K, our small-scale optimization results were confirmed by comparing two conditions head to head on a one-liter scale, where the optimized condition of 21°C and 18h resulted in ~3.5 times more protein per liter and 2.5 times more protein per gram of cell paste when compared with production at the same temperature at 3h. Once a process is validated, the experimenter can then reproducibly obtain the protein of interest, confident that the best yield is being obtained.

Finally, it is important to remember that DOE is merely a statistical tool, a means to an end. It does not guarantee success; it merely provides a framework for unraveling complex relationships between a response and multiple factors. Nor does it replace technical expertise or creativity in experimental work. When used correctly, DOE can be used to empower the role of investigator in the face of increasing automation.

2.3. Notes

1. Our preferred statistical program for use in experimental design is JMP 7.0 (SAS Institute Inc., Cary, NC, USA), but there are many other software packages are available. The following software contains DOE modules: Minitab (Minitab Inc, PA, USA), ECHIP (ECHIP Inc, DE, USA), and Stat-Ease (Stat-Ease Inc, MN, USA).
2. A single experiment can be defined as an experimental run. A design utilizes multiple runs directed toward meeting a single experimental objective.
3. One criticism of DOE is the potentially large number of runs at the onset of any investigation. Therefore many DOE texts recommend marshalling effort and not expending greater than 25% of resources on the initial screen, using the remainder for subsequent designs and the all-important validation step.

4. The change is response between two levels is termed the factor main effect. Since random error can easily obscure the main effect, if the levels are set too close together, the main effect can be estimated most precisely from extreme level settings of the factor.
5. We typically use brain heart infusion media (BHI) for *E. coli* growth, but have also explored other options, including autoinduction media formulated to support periods of cell growth leading to high densities at which point spontaneous induction of protein expression occurs from *lac* promoters, eliminating the need to monitor cell growth or induce with the addition of IPTG. For these reasons autoinduction medium, under its brand name Overnight Express™(Novagen, Madison, WI, USA), has been promoted as being ideally suited to high-throughput protein expression.
6. In order to design a meaningful experiment, an estimate of the response variable is required. Variability is expressed in terms of standard deviation, which is assumed to be constant over the range of response values encountered during experimentation. The spread of this variability will determine the size of the experiment and the number of runs required in the design.
7. Replication is used to dampen any uncontrolled variation (noise) that might occur, so that the variability associated with the phenomenon can be estimated. Replication requires more than simply resampling or taking additional measurements; the entire process must be repeated from start to finish. Where several samples are submitted from a given experiment, the response is generated as an average. The order in which the experiments are performed should also be randomized to avoid influences by uncontrolled variables such as material transfers, weighing error, and instrument readings. These changes, which often are time related, can significantly influence the response. If run order is not randomized, the analysis may indicate factor effects that are really due to uncontrolled variables that just so happen to change at the same time.
8. Blocking screens out noise caused by unknown sources of variation, such as raw materials, machine, or operator differences. By dividing experimental runs into homogeneous blocks and then arithmetically removing the differences, one increases the sensitivity of the DOE analysis. It is important not to block variables of potential interest.
9. Import raw expression data into JMP and convert to mg of protein per liter of culture. Multiply data by 0.033 and

0.024 for bacterial and insect cell expression, respectively (*see* **Chapter 10** for volumes of culture used).

10. In our example, we fitted the data using a multiple regression model since we were using only continuous predictors (time, temperature, and IPTG concentration) to explain a single continuous response (expression level). If one were using only categorical variables, one would fit an analysis of variance (ANOVA) model. In the case where both categorical and continuous predictors are used to fit a model for continuous response, it is called analysis of covariance (ANCOVA).

11. Look for significant factors and interactions by examining the *p*-values for each. The JMP program considers *p*-value <0.05 to be significant, but this is a matter of choice. Due to the variability in cell growth, we often use *p*-values of <0.01 as a cut-off, though one can choose to be more stringent.

References

1. Bousse L, Mouradian S, Minalla A, Yee H, Williams K, Dubrow R. (2001) Protein sizing on a microchip. Anal Chem 73(6):1207–12.
2. Box GEP, Hunter JS, Hunter WG. 2005 Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed. New York: Willey.
3. Chen GQ, Sun Y, Jin R, Gouaux E. (1998) Probing the d binding domain of the GluR2 receptor by proteolysis and deletion mutagenesis defines domain boundaries and yields a crystallizable construct. Protein Sci 7(12):2623–30.
4. Zhang Z, Smith DL. (1996) Thermal-induced unfolding domains in aldolase identified by amide hydrogen exchange and mass spectrometry. Protein Sci 5(7):1282–9.
5. Chambers SP, Austen DA, Fulghum JR, Kim WM. (2004) High-throughput screening for soluble recombinant expressed kinases in Escherichia coli and insect cells. Protein Expr Purif 36(1):40–7.
6. Studier FW, Moffatt BA. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189(1):113–30.
7. Schein CH, Noteborn MHM. (1988) Formation of soluble recombinant proteins in *Escherichia coli* is favored by lower growth temperature. Biotechnology 6:291–4.
8. Bocanegra JA, Bejarano LA, Valdivia MM. (1997) Expression of the highly toxic centromere binding protein CENP-B in E. coli using the pET system in the absence of the inducer IPTG. Biotechniques 22(5): 798–800, 802.
9. Licari PB, Bailey J. (1991) Factors influencing recombinant protein yields in insect cell-baculovirus expression systems: multiplicity of infection and intracellular protein degradation. Biotechnol Bioeng 37:238–46.
10. Moore JT, Uppal A, Maley F, Maley GF. (1993) Overcoming inclusion body formation in a high-level expression system. Protein Expr Purif 4(2):160–3.
11. De Francesco R, Urbani A, Nardi MC, Tomei L, Steinkuhler C, Tramontano A. (1996) A zinc binding site in viral serine proteinases. Biochemistry 35(41):13282–7.
12. Kumagai A, Dunphy WG. (1996) Purification and molecular cloning of Plx1, a Cdc25-regulatory kinase from *Xenopus* egg extracts. Science 273(5280):1377–80.
13. Ying BW, Taguchi H, Kondo M, Ueda T. (2005) Co-translational involvement of the chaperonin GroEL in the folding of newly translated polypeptides. J Biol Chem 280(12):12035–40.

Color Plates

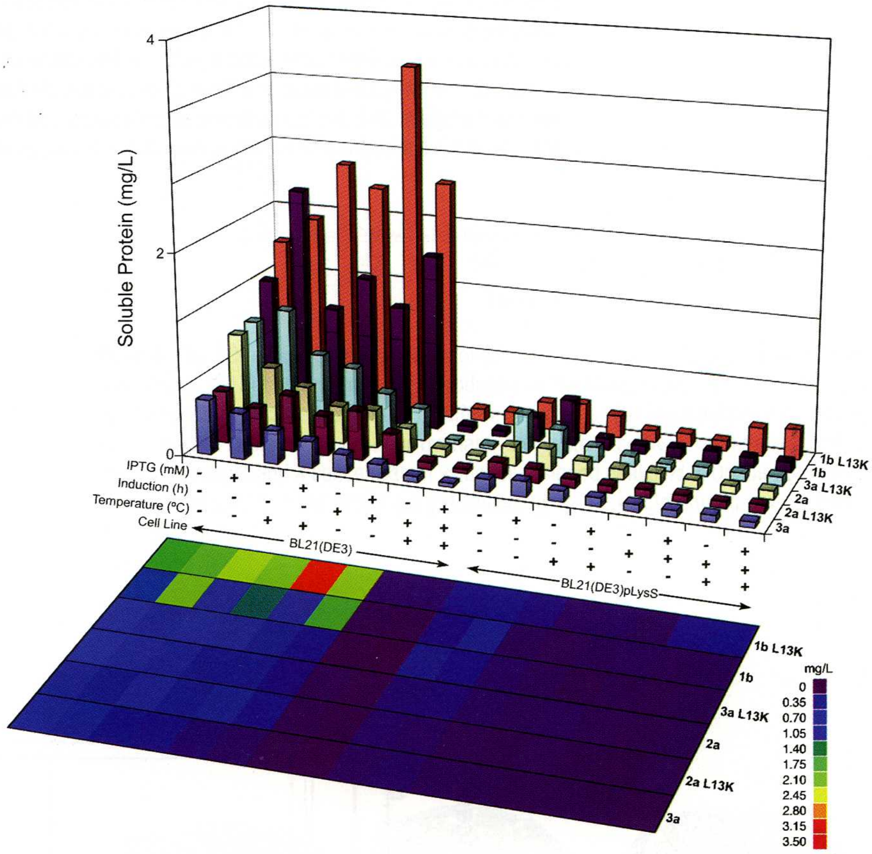

Plate 1, Fig. 2.2. Full factorial screening for effects on the soluble HCV NS3-prt expression in *E. coli*. Levels of expression are illustrated in a 3D Bar Chart and a Heat Map. Using a previously described high-throughput protein expression platform, each data point was obtained using the HT Protein Express 200 Chip run on the Caliper labchip 90 and measured in triplicate (see p. no. 25)

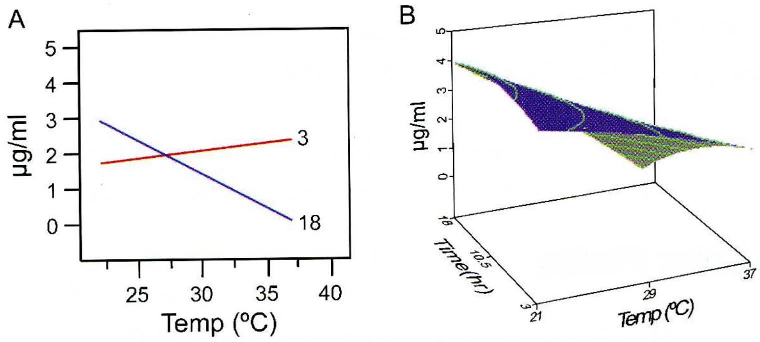

Plate 2, Fig. 2.3. (a) Interaction plot for temperature and time affects on the soluble expression of HCV NS3-prt (1b) L13K. The dependence of protein yield (y-axis) on temperature (x-axis) is plotted for two different times. The red line represents the 3-h data, while the purple line represents the 18-h data. (b) Response surface plot demonstrating expression as a function of time and temperature. The effect of time and temperature on protein yield at an optimal value of IPTG (0.78 mm) is depicted as a 3D surface using the statistical program JMP. The top surface is *colored purple* and the bottom is *colored gray*. Mapped to the surface in *blue* is the contour map of the same data (see p. no. 26)

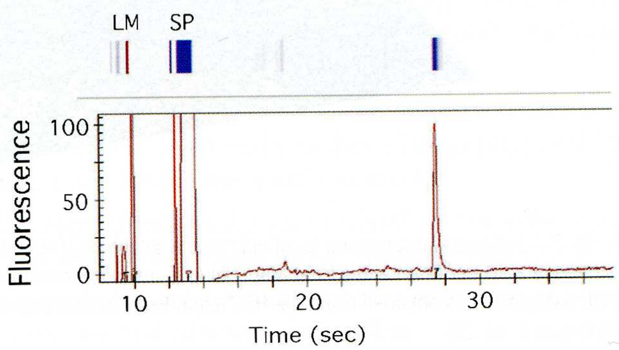

Plate 3, Fig. 10.3. Virtual gel image and electropherogram of protein generated by Caliper labchip 90. Protein was expressed in insect cells, grown in a deep well block, then purified and analyzed. The purified protein is readily identified on the gel and electropherogram, as is the lower marker (LM), an internal reference, and the system peak (SP) (see p. no. 152)