Incomplete factorial and response surface methods in experimental design: yield optimization of $tRNA^{Trp}$ from *in vitro* T7 RNA polymerase transcription

Yuhui Yin and Charles W. Carter, Jr*

Department of Biochemistry and Biophysics, Campus Box 7260, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599-7260, USA

Received December 11, 1995; Revised and Accepted February 22, 1996

ABSTRACT

We have studied the yield of *Escherichia coli* $tRNA^{trp}$ obtained from *in vitro* T7 RNA polymerase transcription using incomplete factorial and response surface methods. Incomplete factorial experiments were first used to estimate the relative impact of six variables on the yield of $tRNA^{Trp}$. Fifteen trials were performed according to a balanced and randomized design. The correlation between observed yield and all experimental variables was identified by stepwise multiple linear regression analysis. The concentrations of T7 RNA polymerase, DNA template, NTP and $MgCl_2$ proved to be significantly correlated with the yield of $tRNA^{Trp}$. We then optimized the yield with respect to each of these four variables simultaneously with a designed, response surface experiment based on the Hardin–Sloane minimum prediction variance algorithm. Twenty experiments were performed, in duplicate, to sample the quadratic surface relating the yield to the four significant variables. Coefficients of the quadratic function with all two-factor interactions were evaluated by stepwise regression using least squares, and significant coefficients were retained. Partial differentiation of the resulting quadratic model showed it to possess an optimum. Transcription performed at the corresponding conditions yielded 6-fold more $tRNA^{Trp}$ than the initial conditions, confirming the predictive value of the experimentally determined response surface.

INTRODUCTION

Tryptophanyl-tRNA synthetase (TrpRS) catalyzes a multi-substrate reaction. The enzyme is believed to undergo conformational changes to accommodate the binding of different substrates (1) and an effort to understand these changes using X-ray crystallography is in progress (1–5). The structure of the complex between TrpRS and $tRNA^{Trp}$ is of particular interest because it is expected to reveal structural details of the catalytic center for acyl transfer and the mechanism of discrimination of cognate tRNA from other tRNAs.

Crystallographic study of the TrpRS–tRNA complex requires a large quantity of $tRNA^{Trp}$. *In vitro* tRNA transcription using T7 RNA polymerase provides a simple, rapid way to produce $tRNA^{Trp}$. One disadvantage of the *in vitro* transcription technique is that the transcribed tRNA does not have the usual modified bases. Fortunately, unmodified and native *E.coli* $tRNA^{Trp}$ have nearly the same acyl transfer activity (our unpublished results), as was found for other tRNAs (6). Our initial efforts to synthesize $tRNA^{Trp}$ using *in vitro* transcription were disappointing, however, producing only ~410 µg/ml of reaction mix, compared to ~1500 µg/ml (6).

Since the yield of *in vitro* transcribed tRNA can depend on a variety of factors whose relative importance are unknown, *a priori*, designed experiments aimed at finding optimal reaction conditions for a given synthesis can be useful. The components of the *in vitro* transcription reaction are well-defined and amenable to rational optimization. Differences in template sequence and length can result in substantial differences, for example in the optimal enzyme and template concentration required for the production of RNA. In addition, the effects of one or more of the components may depend on the choice of the others. Therefore, for initially screening the potential variables, the trial reactions should not only sample the full experimental space, but should also allow for estimation of interactions, or synergistic effects between sample variables.

Here we introduce a method combining incomplete factorial and response surface experimental designs to maximize the yield of tRNA from an *in vitro* transcription system. Our method of screening for important variables and then optimizing the yield of *in vitro* transcription with respect to those variables is quite generally useful in contexts where a process depends on many factors. It should therefore be appropriate, *mutatis mutandis*, for other multi-dimensional optimization problems in nucleic acids research, including both chemical and biochemical syntheses, expression and purification. Because both the experimental sampling designs and subsequent analysis procedures are unfamiliar, we present this example in detail.

Incomplete factorial designs were developed to efficiently and uniformly sample full-factorial designs involving large numbers of combinations of independent variables (7–9). In these designs two-way interactions are balanced, virtually without confounding between main effects and two-way interactions, so that multiple

[*] To whom correspondence should be addressed

linear regression models can be used to identify statistically significant main effects and potentially important synergistic effects. Hence, they provide effective and economical coarse screening of different possible factors to identify those most likely to be crucial for subsequent optimization.

Once these factors have been identified, optimization of reaction conditions can be accomplished using the response surface method. A response surface is an analytical model that tries to reproduce how the system actually responds to changes in the independent variables. We used a multivariate quadratic polynomial function fitted to a set of trial experiments. These experiments were designed to have maximal impact on the accuracy of the coefficients of the response-surface model (10). The method can be used to locate stationary points that may be optima, and hence to find the best conditions for a desired result, here the yield of tRNA^{Trp}. It is suitable whenever one knows something about where the best result (the optimum) might be obtained. Instead of sampling the experimental test space uniformly, the sampled points lie near the surface of a hypercube centered close to the suspected optimum. By sampling the surface in this way, one achieves the greatest contrast between results near the suspected optimum and those distant from it. A consequence is that the prediction variance of the resulting model is a minimum. Hence, these designs are called ‘minimum prediction variance’ designs, or ‘I-optimal’ (10).

Once coefficients of the model have been estimated, the real optimum can be estimated analytically by partial differentiation of the fitted quadratic model with respect to the experimental variables, setting the gradient equal to zero, and solving the resulting simultaneous equations. Normally, it is not exactly the same as the ‘suspected optimum’, but somewhere close-by. Therefore, the response surface method is used to ‘fine tune’ conditions. This fine-tuning can, nevertheless, result in a substantial improvement in the desired experimental result, because it takes into account the interaction terms in modeling the non-linear response of the system. A second important advantage of locating and using stationary points arises because the partial derivatives of the response surface with respect to the experimental conditions vanish at the stationary point, so the resulting optima are also points where the results are most reproducible.

We show below how we have increased the yield of tRNA^{Trp} 6-fold by combining incomplete factorial design to screen factors and response surface methods to optimize the four most important ones. This combination provides a powerful means for studying any optimization problem, using a relatively small number of experiments to quantitatively locate conditions for optimal results. We have previously described its use in screening for and optimizing protein crystal growth (3).

MATERIALS AND METHODS

Materials

The bacterial strain containing cloned *E.coli* tRNA^{Trp} gene was originally prepared for us by Dana Folkes (Pathology Department, UNC), then modified and kindly returned to us by Dr M. J. Rogers (Yale University), who also provided our starting conditions. The gene is under the control of the T7 promoter and is an *Eco*RI/*Bst*NI fragment in the plasmid pUC2119 (a derivative of pUC12 with an additional site at the polylinker sites). The plasmid carrying the tRNA^{Trp} gene (pUC2119) was purified from the DH5 strain transformed with the plasmid on a cesium chloride density gradient. T7 RNA polymerase was purified to the specific activity of 450 000 U/mg from *E.coli* strain BL21 with a cloned T7 RNA polymerase gene (pAR2119) following the procedure of Grodberg *et al* (11). Restriction enzyme *Bst*NI was purchased from New England Biolabs (Beverly, MA). Molecular biological grade nucleoside triphosphates ATP, GTP, CTP, UTP (NTPs) and yeast inorganic pyrophosphatase was purchased from Sigma Chemical Co. (St. Louis, MO). The RNase inhibitor RNasin was purchased from Promega Biotec. (Madison, WI). Toluidine Blue-o was purchased from Ernest F. Fullam Inc. (Schenectady, NY).

Transcription

To avoid ribonuclease-catalyzed degradation, all solutions were either treated with diethlypyrocarbonate (1% v/v) then autoclaved at 120°C for 30 min or supplemented with the RNase inhibitor RNasin (1000 U/ml).

Purified plasmid was linearized by digestion with *Bst*NI (1 U/µg of DNA) in a buffer containing Tris–HCl (10 mM, pH 8.0), NaCl (150 mM), $MgCl_2$ (10 mM) at 37°C for 2 h. An equal amount of *Bst*NI was added again, and the incubation continued for an additional 2 h. The reaction mixture was extracted with an equal volume of phenol/chloroform (1:1), and the digested DNA was precipitated with ethanol before being used in transcription reactions.

The tRNA^{Trp} was synthesized by *in vitro* runoff transcription of cloned DNA using T7 RNA polymerase (6). All reactions were performed at 37°C in the presence of spermidine (2 mM), dithiotheitol (10 mM), RNasin (1000 U), and yeast inorganic pyrophosphatase (5 U/ml). The concentrations of T7 RNA polymerase, DNA template and NTPs, together with four other factors (see below), were varied to survey their effects on the yield of transcribed tRNA^{Trp}. The reaction mixture was incubated overnight. DNase was added to give a final concentration of 10 U/ml, and the reaction mixture was incubated for 30 min. The mixture was extracted successively with equal volumes of phenol/chloroform/isoamyl alcohol (25:24:1) and chloroform/isoamyl alcohol (24:1), and then precipitated with 2.5 vol of ethanol by centrifuging at 14 000 *g* for 30 min. The recovered tRNA was dialyzed against 5 mM HEPES–KOH (pH 8.0) containing EDTA (1 mM) at 4°C overnight.

Quantitation of tRNA^{Trp}

It is crucial for structural studies that tRNA^{Trp} have the correct configuration in the active site, i.e., intact 3′ and 5′ termini. It has been known and proven in our system that T7 polymerase transcripts also contain, in addition to full length product, $N \pm 1$ and/or $N \pm 2$ products, i.e. one or two extra or fewer nucleotides at the 3′-terminus. These products are not suitable for our purpose. Therefore, the definition of yield in our reactions is tRNA which possesses tryptophan acceptor activity, and determined by measuring [^{14}C]tryptophanyl-tRNA^{Trp} formation, as described by Yarus *et al.* (12). To compare the relative amount of transcribed tRNA from a set of experiments, we also analyzed transcription products on a 50 × 30 cm, 10% acrylamide sequencing gel run at 400 constant voltage for 12 h. The distribution of tRNA was visualized by staining 10–20 min in a toluidine blue dye solution (4 g toluidine blue-o, 500 ml methanol, 10 ml glacial acetic acid and water to 1 l).

Table 1. Incomplete factorial experimental design matrix

Expt	GMP mM	NTP_{total} mM	Template µg	T7 pol U	$MgCl_2$ mM	GC:AU	$tRNA^{Trp}$ µg/ml	Yield / arbitrary densitometry units
1	20	5	25	1500	10	1:1	71	31
2	0	3	25	3000	0	1:1	357	284
3	20	3	50	3000	10	2:1	1392	1225
4	0	3	50	3000	0	2:1	1658	1623
5	0	3	25	1500	10	2:1	99	45
6	20	5	25	3000	0	1:1	31	35
7	0	5	50	1500	10	1:1	1504	1243
8	0	5	25	1500	0	2:1	141	136
9	20	3	25	3000	10	2:1	315	641
10	0	5	50	3000	10	1:1	2765	1561
11	20	3	50	1500	0	1:1	209	207
12	0	3	25	1500	10	1:1	92	57
13	0	5	50	1500	10	2:1	2535	1623
14	20	3	50	3000	0	2:1	1493	1089
15	0	5	25	3000	0	1:1	141	80

Identification of important variables by incomplete factorial design

Reaction components were first evaluated for their potential effect on $tRNA^{Trp}$ production. Six components were chosen as variables and tested at two levels using the incomplete factorial screening design in Table 1. These are listed below, with our rationale for including them and for choosing their values.

T7 RNA polymerase concentration. T7 RNA polymerase forms an initiation complex with DNA templates, and an elongation complex with DNA, template, and nascent transcribed RNA. The stability of these complexes depends on the sequence of the DNA template, the secondary structure of the transcribed RNA, and ionic strength of the reaction mixture. Interactions of T7 polymerase with DNA template, NTPs, and transcribed RNA distinguish it as an important potential factor in transcription optimization. The concentration of T7 polymerase should be carefully adjusted. Also, freshly purified polymerase had higher activity than did commercial enzyme. It has been suggested that 50–100 µg/ml of fresh purified T7 polymerase be used in the transcription reaction (13,14). It was used in screening experiments at 50 and 100 µg/ml.

DNA template concentration. As a component of initiation and elongation complexes, DNA affects the yield by its sequence and conformation, as well as its concentration. The stability of the elongation complex is also affected by the complementary structure of DNA and transcribed RNA. Since the T7 promoter has relatively low selectivity, the ratio of DNA to T7 polymerase concentrations is as important as, if not more important than the concentration of DNA template itself. The concentrations tested were 25 and 50 µg/ml.

Total NTP concentration. Although the overall K_ms for NTP incorporation are estimated to be in the range of 100 µM, a high level of incorporation of these components occurs only at a relatively high ratio of NTP concentration to the K_m (13). High concentrations of NTPs are also needed to overcome the consumption in abortive products, which can be formed in molar quantities as high as 10-fold that of the full-length product. The concentrations of NTP (each) tested were 2 and 4.5 mM, in accordance with recommendations from Dr M. J. Rogers (personal communication).

The ratio of GTP and CTP concentrations to ATP and UTP concentrations. The GC content in the *E.coli* $tRNA^{Trp}$ gene is almost 2-fold greater than the AU content. Providing a similar ratio to that in the $tRNA^{Trp}$ gene might improve the incorporation and thus increase the yield. The GC:AU ratios tested were 2:1 and 1:1.

GMP concentration. GMP has been suggested to be involved in initiation of the transcription reaction (13). The transcription reactions were performed at 0 or 50 mM.

$MgCl_2$ concentration. Mg^{2+} is an essential cofactor of RNA polymerase, forming a complex with the triphosphates and compensating the negative charges on the α-phosphates of the NTPs. Mg^{2+} may also induce a conformational change from right-handed B-form to left-handed Z-form in GC rich DNA. T7 RNA polymerase favors binding to the Z form (15). As the Mg^{2+} concentration affects the equilibrium between the B and Z conformations of the DNA template it therefore potentially affects the concentration of the structurally favorable substrate for the polymerase. Since NTP concentration varies in the experiments, additional $MgCl_2$ must be added in order to keep Mg^{2+} concentration greater than that of the NTP. Since the reaction buffer contains 20 mM $MgCl_2$, this variable was chosen to test an optimal excess level of $MgCl_2$. Supplemental $MgCl_2$ concentrations tested were 5 and 10 mM.

To identify the relative importance factors of these six factors for the yield of $tRNA^{Trp}$ transcription, they were tested in 15 experiments according to the design matrix given in Table 1. For six variables with two levels for each variable, there are $2^6 = 64$ possible combinations in the full factorial design. Incomplete factorial design allows one to select an effective sample from the total of 64 experiments by a procedure based on the advantages of randomized testing and balancing of two-factor interactions (7). The number of trial experiments necessary to identify the most important factors is thereby greatly reduced. In this case we selected 15 experiments using the computer program INFAC (8).

Our experience, conditioned by applications in screening crystal growth conditions and phase permutation experiments (16) suggests that the size of a screen required to preserve strong two-factor interactions is approximately $\sqrt{N}$ experiments, where N is the number in the full factorial. This 'rule of thumb' is suggested by analogy to sampling designs based on error correct codes (17), which also preserve higher-order interactions with $\sqrt{N}$ experiments. Therefore, with approximately twice this number, data from these sampled points should be sufficient to identify the most significant factors without performing the complete set of 64 experiments.

The yield of transcribed tRNA^{Trp} was measured as described above for 15 experiments. A regression model was established with the yield of tRNA as the dependent variable with a linear dependence on the six selected factors as independent variables. The significant coefficients of this model were identified by the stepwise regression method, as implemented in the MGLH module (multiple regression, general linear hypothesis) of the computer program SYSTAT 5.2 (18).

The yield of tRNA^{Trp} can be estimated by:

$$Yield_{\text{calc}} = a_0 + \sum_{i=1}^{N} a_i x_i \tag{1}$$

where $Yield_{\text{calc}}$ is yield calculated from the model, N is the number of the variables ($N = 4$), x_i ($i = 1, 2, ..N$) are the significant variables and a_i is the coefficient that describes the contribution of variable x_i to the yield of tRNA^{Trp}. The coefficients are estimated by minimizing the sum of the squared differences between $Yield_{\text{calc}}$ and experimental yield, $Yield_{\text{obs}}$. The absolute value of a coefficient, relative to its standard error, suggests the relative degree to which the corresponding variable contributes to the overall yield. A positive sign for a given coefficient suggests that higher concentrations of the variable increase the yield of tRNA^{Trp}, whereas a negative sign suggests that lower concentrations provide higher yields.

Stepwise regression facilitates the difficult task of selecting the right variables for the regression function by selecting an appropriate subset of variables from the full set. Variables are considered in order of their relative importance as indicated by Student t-tests based on the magnitude of the coefficient, relative to the standard error. Coefficients whose t-tests have a small probability (P<0.01) under the null hypothesis indicate significant contributions to the dependent variable. From the full list of independent variables the stepwise algorithm builds (or depletes) the regression model, one factor at a time. The independent variable selected at each step is that with the greatest (or least) potential to predict the remaining variation in the dependent variable, corrected for terms already in the model.

Optimization by response surface analysis

The linear model describing results of the incomplete factorial design indicates only trends arising from the different treatments of independent variables. Values for these variables are not necessarily at their best levels, because relationships between the dependent variable and independent variables are often nonlinear. The actual functional dependencies can be more accurately described by higher-order polynomial functions, the simplest of which is a quadratic polynomial. Having identified the significant variables and their effect on the yield of tRNA^{Trp} by the incomplete factorial method, we next aimed at maximizing the yield by fitting a quadratic model:

$$Yield_{\text{calc}} = \beta_0 + \sum_{i=0}^{N} \beta_i x_i + \sum_{j \to i, i=1}^{N} \beta_{ij} x_i x_j + \sum \beta_{ii} x_{ii} \tag{2}$$

where $Yield_{\text{calc}}$ is the calculated yield, N is the number of the variables in the model, β_o, β_i, and β_{ij} (i, j = 1, ...4) are coefficients determined by regression, and x_i and x_j are the variables.

Steps involved in optimization were as follows.

Establish the design matrix. Depending on the number of the variables, the number of the experiments to be performed to establish the model varies. The minimum number of coefficients in the general four-dimensional quadratic model is 15. Our model was fitted to data obtained using 20 sampling points. These sampling points were arrayed on a 4-dimensional hypercube, each dimension of the hypercube representing the range to be tested for one variable. The suspected optimum was chosen as the center of the hypercube. Two experiments were performed at the center of the hypercube with the rest of experimental conditions scattered essentially randomly near the borders of the hypercube so as to minimize the prediction variance of the resulting quadratic model. This aspect of the design matrix was carried out by the computer program, GOSSET (10). Specifications for 20 sampling points in our optimization experiments are encoded in the design matrix in Table 2. The point [0, 0, 0, 0] represents the mean value of the four variables, and +1 and –1 the highest and lowest levels tested for each variable.

Table 2. Hardin–Sloane minimum integrated variance design matrix for four factors, 20 experiments

Expt	NTP	DNA	T7 pol	$MgCl_2$
1	0.000	–0.056	0.000	–0.250
2	0.000	–0.056	0.000	–0.250
3	0.000	1.000	0.000	–0.250
4	0.000	–1.000	0.000	–1.000
5	1.000	–0.007	0.116	1.000
6	–1.000	–0.007	–0.116	1.000
7	0.210	0.108	–1.000	–1.000
8	–0.210	0.108	1.000	–1.000
9	–1.000	–1.000	1.000	–0.250
10	–1.000	1.000	–1.000	–0.250
11	1.000	–1.000	–1.000	–0.250
12	1.000	1.000	1.000	–0.250
13	0.492	–1.000	1.000	1.000
14	–0.492	–1.000	–1.000	1.000
15	–1.000	1.000	–0.577	–1.000
16	1.000	1.000	–0.577	–1.000
17	0.669	1.000	–1.000	1.000
18	–0.669	1.000	1.000	1.000
19	–1.000	–1.000	–1.000	–1.000
20	1.000	–1.000	1.000	–1.000

This design was prepared by N. J. A. Sloane using GOSSET (10). Matrix entries should be interpreted as: 0 = the centrer, –1 = the minimum end, and 1= maximum end of the variable range.

Determine physically reasonable mean, maximum, and minimum levels for each variable to be tested. The four significant variables selected from incomplete factorial design were concentrations of DNA template, NTP_{total}, T7 polymerase, and $MgCl_2$. The variables and their levels used are shown in Table 3.

Table 3. Variable range assignments for response-surface experiments

Variables	Center	Range
NTP (mM)	40	20–60
DNA (mg/ml)	100	50–150
T7 polymerase (µg/ml)	75	50–100
$MgCl_2$ (mM)	10	0–20

Perform the experiments. The conditions of the 20 experiments to be performed (Table 4) were determined from the mean and the range for each variable and the design matrix. The yield of tRNA from each experiment was measured as described above.

Table 4. Optimization experiments performed to establish the quadratic model

Expt	NTP_{total} mM	DNA µg/ml	T7 pol µg/ml	$MgCl_2$ mM	Yield µg/ml
1	40.0	95.0	75	7.5	1886
2	40.0	95.0	75	7.5	1755
3	40.0	150.0	75	7.5	2127
4	40.0	50.0	75	0.0	208
5	60.0	100.0	80	20.0	360
6	20.0	100.0	70	20.0	340
7	44.0	110.0	50	0.0	59
8	36.0	110.0	100	0.0	508
9	20.0	50.0	100	7.5	254
10	20.0	150.0	50	7.5	732
11	60.0	50.0	50	7.5	79
12	60.0	150.0	100	7.5	1710
13	50.0	50.0	100	20.0	716
14	30.0	50.0	50	20.0	213
15	20.0	150.0	90	0.0	1755
16	60.0	150.0	60	0.0	87
17	53.2	150.0	50	20.0	316
18	26.0	150.0	100	20.0	693
19	20.0	50.0	50	0.0	922
20	60.0	50.0	100	0.0	41

This matrix was generated by substituting Table 2 with the variable ranges in Table 3.

Develop a mathematical model. A quadratic model with the form of equation was obtained using SYSTAT 5.2 (MGLH module) starting with all the terms in equation **2** and using the same stepwise regression algorithm and criteria outlined above.

Locate and characterize the stationary points of the response surface. Having found an appropriate model, the optimum for the yield of tRNA can be determined analytically by partial differentiation against each variable and equating the gradient to zero. Coordinates of the resulting stationary point provide estimates for the variable concentrations giving the optimum results. The stationary point must then be verified by carrying out the experiment at those values.

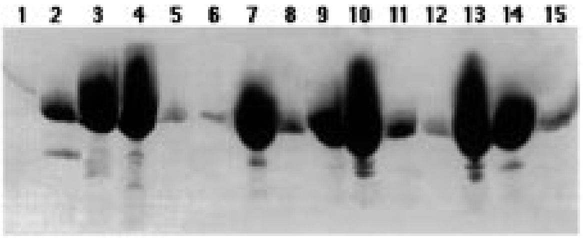

Figure 1. Transcripts from the 15 experiments in the incomplete factorial screening design. Equal volumes (5 ml of the 100 ml transcription mix) were applied to the 10% acrylamide gel. Band intensities therefore represent overall incorporation of nucleotides. *N*+1 products are visible in experiment 15 and incomplete transcripts are also seen in many of the lanes. For this reason, we used tryptophan acceptor activity, rather than the overall incorporation of radioactive bases, to quantitate the reaction yields in this and subsequent reactions.

RESULTS

Screening potential variables with an incomplete factorial design

There was considerable variation of the transcript yields among the 15 experiments from the incomplete factorial design shown in Table 1. Transcripts are illustrated in the electrophorogram in Figure 1 and results of densitometric tracing are included in Table 1. The intensity of bands on the gel corresponded to the tryptophan acceptor activity, with a squared multiple correlation coefficient of 0.95, confirming that the majority of the transcripts were $tRNA^{Trp}$. However, there is also variation in the degree of contamination by $N \pm i$ fragments. The variation in $tRNA^{Trp}$ tryptophan acceptor activity for the 15 experiments were analyzed by stepwise multiple regression (Table 5). Five of the six factors, the NTP, DNA, T7 polymerase and $MgCl_2$ concentrations, and the presence of GMP, all proved to have significant impact on the yield $tRNA^{Trp}$. Probabilities for their t-tests ranged from 0.003 (for the overall NTP concentration) to 3.4×10^{-12} (for the concentration of DNA template). The model was:

$$\text{Yield } (\mu g/ml) = -3115 - 577.8\,[GMP] + 416.8\,[NTP] + 1538.7[DNA] + 672.7[T7\ pol] + 603.7\,[MgCl_2] \qquad \mathbf{(3)}$$

The GC composition of the reaction was not a significant factor. The minus sign for the GMP coefficient indicates that the lower concentration of GMP produced, on average, a higher yield. Since the presence of GMP depressed the yield, it was not selected for further variation in the subsequent optimization phase. Coefficients for the remaining variables were all positive, indicating that higher concentrations of T7 polymerase, NTP, DNA template and $MgCl_2$ facilitate the production of $tRNA^{Trp}$. The yields estimated from the model agreed well with the experimental data (multiple R = 0.964 squared multiple R = 0.929), and the probability of the overall F-ratio test was *P* = 0.0001.

Table 5. Statistics for the incomplete factorial design model

Regression statistics

Variables	Coefficient	Std error	Std coefficient	Tolerance	Student's t	*P* (2-tail)
Constant	–3115	448	0.000	–	–7.0	0.34e–06
GMP	–578	127	–0.293	0.919	–4.6	0.13e–03
NTP	417	124	0.215	0.926	3.4	0.003
DNA	1539	121	0.796	0.980	12.8	0.34e–11
T7 pol	673	132	0.348	0.822	5.1	0.31e–04
$MgCl_2$	604	128	0.312	0.875	4.7	0.82e–04

Analysis of variance

Source	Sum-of-squares	df[a]	Mean-square	F-ratio	*P*[b]
Regression	0.253722e+8	5	5074430	47.7	0.108e–10
Residual	2551252	24	106302		

Dependent variable, yield
Number of experiment (including duplicates), 30
Multiple R, 0.953. Squared multiple R, 0.909
Adjusted squared multiple R, 0.890. Standard error of estimate: 326
[a]df, degrees of freedom.
[b]*P*, probability for F-ratio test.

Table 6. Statistics for the regression of the optimization model

Regression statistics

Variable	Coefficient	Std error	Std coefficient	Tolerance	Student-test	*P* (2-tail)
T7pol	23.115	8.300	1.780	0.022	2.785	0.015
NTP^2	–0.966	0.259	–2.107	0.028	–3.733	0.003
$NTP*MgCl_2$	2.137	0.633	1.069	0.088	3.377	0.005
NTP*T7pol	0.678	0.266	2.246	0.011	2.550	0.024
DNA*T7pol	0.099	0.029	0.831	0.148	3.397	0.005
$MgCl_2$	–4.750	1.270	–1.044	0.113	–3.741	0.002
$T7pol^2$	–0.346	0.108	–2.312	0.017	–3.191	0.007

Analysis of variance

Source	Sum-of-squares	df[a]	mean-square	F-ratio	*P*[b]
Regression	0.179925e+8	7	2570364	14.3	0.348e–04
Residual	2332981	13	179460		

Dependent variable, yield
Number of experiments, 20
Multiple R, 0.941. Squared multiple R, 0.885
[a]df, degrees of freedom
[b]*P*, probability for F-ratio test.

Optimization with response surfaces

Yields for the 20 experiments from the Hardin–Sloane matrix (Table 3) are shown in Table 4. The best quadratic model, summarized in Table 6, was:

$$\text{Yield} = 23.1[\text{T7 pol}] - 0.966[\text{NTP}]^2 + 0.678[\text{NTP}][\text{T7 pol}] + 2.14[\text{NTP}][\text{MgCl}_2] + 0.099[\text{DNA}][\text{T7 pol}] - 0.346[\text{T7 pol}]^2 - 4.75\,[\text{MgCl}_2]^2 \qquad \textbf{(4)}$$

There are no terms for either $[DNA_{template}]$ or $[DNA_{template}]^2$ in this equation, as neither term was significant. The quadratic term was positive, however, which suggests that it does not reach maxima within the experimental range. The upper-limit of the concentration range was used in calculations of the optimum.

The stationary point, i.e., the optimum transcription reaction conditions for the yield of $tRNA^{Trp}$, was calculated by taking partial derivatives of equation **4** with respect to [T7 polymerase], $[NTP_{total}]$, and $[MgCl_2]$, respectively. Equating these three derivatives to zero and solving for the coordinates of each variable ($T7pol_{opt}$, NTP_{opt}, $MgCl_{2\ opt}$) gave the following optimal concentrations: [T7 pol] = 98 µg/ml, $[NTP_{total}]$ = 46.4 mM (11.6 mM each NTP), $[MgCl_2]$ = 10.35 mM, while $[DNA_{template}]$ was given a value of 150 mg/ml. The yield predicted from the model under these conditions ($Yield_{calc)}$ was 2400 µg/ml. Transcription at these conditions furnished 2450 µg/ml of $tRNA^{Trp}$, which agreed well with the predicted value, and was 6-fold higher than the yield of 410 μg/ml from our initial conditions, and nearly 70% higher than that reported for $tRNA^{Phe}$ (6).

DISCUSSION

In vitro transcription is a convenient method to obtain tRNA for structural studies (20) and is representative of a variety of synthetic and other processes in widespread use. The efficiency of transcription and, ultimately, the overall yield depend on DNA template sequence, making each example unique with respect to its dependence on other reaction conditions. Each system is therefore likely to have different sets of optimum conditions. Efficient searches for optimal conditions are therefore desirable, especially in cases where preparative scale is important for subsequent studies.

The transcription reaction initiates by forming a complex of DNA template and T7 polymerase, and then an elongation complex of DNA, T7 polymerase and nascent RNA. The concentrations of template and polymerase affect the stability of these complexes, as does the superhelicity of the DNA template. Moreover, the subsequent elongation steps depend on the NTP concentrations and also affect the yield of tRNA. Moreover, any of these factors could affect the yield synergistically. Interactions between these factors raises the possibility that most ‘optimized’ conditions produce suboptimal yields because they are not at a global optimum. Moreover, the search for globally optimal conditions can be tedious because it must be carried out in many dimensions.

Our strategy uses a minimum number of sampled factorial experiments to capture the important multidimensional information about the system. This is accomplished by separating the task of identifying important variables from that of optimizing those variables with response surface methods. For each stage we use designed experiments that can be analyzed coherently to provide conclusions with adequate statistical significance. Incomplete factorial designs with uniform sampling properties are used in the first stage, while the second stage involves designs sampled in accordance with the property of minimum prediction variance for a quadratic response surface.

This study demonstrated that the maximum yield of tRNA depended most critically on four factors: the concentrations of T7 polymerase, NTP and $MgCl_2$ and DNA. Accurate relationships between these factors were described by a quadratic model, equation **4**, from which the yield can be estimated for any combination within the range on which it was determined. We successfully increased the yield of $tRNA^{Trp}$ by 6-fold at the stationary point of this model.

Screening variables

It is surprising that equation **3** suggests that higher yields occur in the absence of GMP. By the same token, the GC:AU ratio made

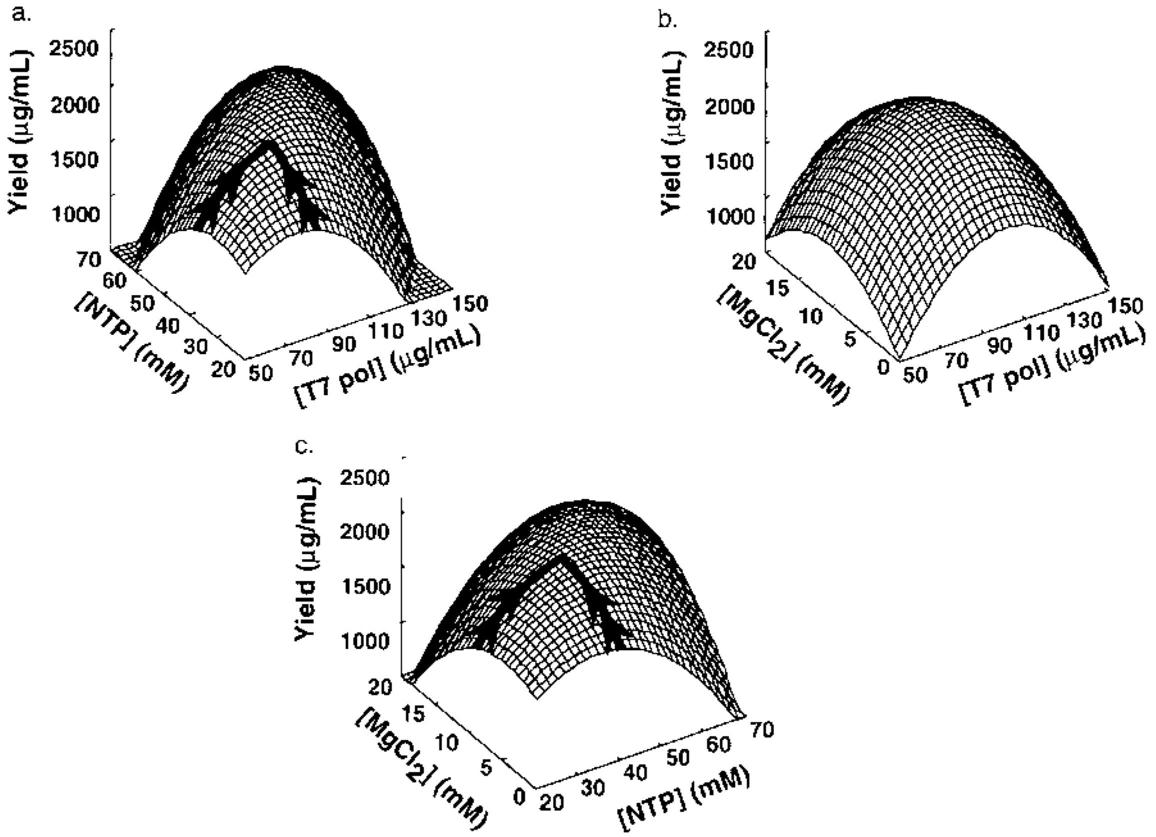

Figure 2 .Level surfaces for the response surface, equation **3**. (**a**) Yield of $tRNA^{Trp}$ versus the concentration of T7 polymerase and NTP at constant DNA and $MgCl_2$ concentrations ([DNA] = 150 mg/ml, $[MgCl_2]$ = 10 mM). The equation for this level surface is: $Y = 38[T7] – 0.35[T7]^2 + 24[NTP] – 0.97[NTP]^2 + 0.68[NTP][T7] – 509.0$. (**b**) Yield of $tRNA^{Trp}$ versus the concentration of T7 polymerase and $MgCl_2$ at constant level of DNA and NTP. The equation for this surface is: $Y = 69[T7] – 0.35[T7]^2 – 4.8[MgCl_2]^2 + 99[MgCl_2] – 2080.0$. (**c**) Yield of $tRNA^{Trp}$ versus the $MgCl_2$ and NTP concentrations at the optimum level of T7 polymerase and DNA. The equation for this surface is: $Y = 66[NTP] – 0.97[NTP]^2 + 2.14[NTP][MgCl_2] – 4.8[MgCl_2]^2 + 396.0$. The dark traces in (**a**) and (**c**) show that ‘one-at-a-time’ optimization of the variables will often miss the global optimum by a wide margin, resulting in substantial losses. These discrepancies are the result of the two-way interactions present in the corresponding level surface equations, and which are absent from the equation for (**b**).

an unimportant contribution to the yield. Moreover, the perceptive reader may notice that the yields for experiments 10 and 13 in Table 1 are somewhat higher than that obtained at the optimum. It is possible that GMP and the GC content do influence the yield via important interactions, so it is unlikely that we have found a global optimum for $tRNA^{Trp}$ production by *in vitro* transcription. No optimal template concentration was determined, nor have we made use of the two-way interaction between GMP and the GC:AU ratio, though examination of the data in Table 1 suggest that it may actually be important (data not shown).

Multidimensionality, optimization, and reproducibility

The relationship between the yield and the variables can be visualized by plotting values of equation **3** in graphs of the dependent variable versus two selected independent variables at a constant level of the rest of the variables. These graphs are called level surfaces. The curvature of the response surface reflects the degree of influence these two variables have on the dependent variable. We have plotted in Figure 2 the three level surfaces arising from each pair of independent variables whose quadratic terms had negative coefficients, and which therefore give rise to maxima. In our study the DNA template concentration did not reach an optimum, therefore the yield of tRNA continuously increased as the concentration of DNA template increased.

The T7 polymerase-NTP level surface (Fig. 2a) shows the relationship between the yield of $tRNA^{Trp}$ and the T7 polymerase and NTP concentrations at constant concentration of DNA template and at the optimal concentration of $MgCl_2$. The optimal concentrations, [T7 pol_{opt}] and [NTP_{opt}], are the concentrations of T7 polymerase and NTP corresponding to the maximum $tRNA^{Trp}$ yield. The sharpness of the peak indicates that the concentration of T7 polymerase is crucial to the yield, which will decrease rapidly as the concentration of T7 polymerase moves away from the optimal concentration. The synergistic effect of polymerase and template concentrations is important, however, as indicated by the strength of the corresponding coefficient in equation **4**, Table 6.

The T7 polymerase–$MgCl_2$ level surface (Fig. 2b) demonstrates the effects of T7 polymerase and $MgCl_2$ on the yield of $tRNA^{Trp}$ at constant concentrations of DNA template and NTP. The Mg^{2+} could play several roles. The yield of tRNA relies on the efficient formation of an initiation complex of T7 polymerase and DNA template. T7 polymerase prefers to bind to a left-handed Z-form DNA template (15). Mg^{2+} may therefore increase the yield of tRNA by favoring an appropriate DNA template conformation for T7 polymerase. The equation for this surface (see the figure legend) has no coefficient for the interaction term, suggesting that the polymerase and $MgCl_2$ concentrations act independently. There is no evidence in this level surface for strong synergistic effects between $MgCl_2$ and the polymerase.

The NTP–$MgCl_2$ response surface (Fig. 2c) again shows the effects of an important two-factor interaction. Both initiation and elongation of the transcription depend on efficiency of complex formation, which is affected by the conformation of DNA template and the incorporation of NTP. Mg^{2+} is an essential cofactor of T7 polymerase and generates its favorable DNA template conformation, facilitating elongation complex formation.

Two additional points should be made with respect to the use of stationary points for preparative syntheses of $tRNA^{Trp}$. First, there is an obvious relationship between the idea of a stationary point and the intrinsic reproducibility of conditions for maximum yield. The gradient is close to zero near the stationary point, and the yield at the stationary point will therefore be less vulnerable to experimental fluctuations represented by the partial derivatives that vanish.

Secondly, the multidimensional search inherent in our approach is inherently superior to optimizing one factor at a time. As shown by the dark traces in Figure 2a and c, the synergy between polymerase and NTP concentrations and between NTP and $MgCl_2$ concentrations means that optimization of one variable at a time is unlikely to lead to the global optimum. Unless one is fortunate enough to optimize the first variable at the optimum value of one or more of the others, the first optimization will predispose toward a false optimum because these two pairs of variables influence one another's optimum values. Moreover, the loss incurred by missing the optimum in any two-dimensional level surface is multiplicative, owing to the multidimensionality of the response surface. As a result, one can readily rationalize losses approaching an order of magnitude, even when each of several parameters has been 'optimized' by one-dimensional experiments. This phenomenon may also help explain why some published protocols can be hard to reproduce.

ACKNOWLEDGMENTS

We thank Drs Lee Kuyper and Howard Fried for critical reading of the manuscript. This work was supported by grants from the American Cancer Society and the NIH.

REFERENCES

1 Carter, C. W., Jr., Doublié, S. and Coleman, D. E. (1994) *J. Mol. Biol*, **238,** 346–365.
2 Coleman, D. E. and Carter, C. W., Jr. (1984) *Biochemistry*, **23,** 381–385.
3 Carter, C. W., Jr and Yin, Y. (1994) *Acta Cryst*, **D50,** 572–590.
4 Doublié, S., Bricogne, G., Gilmore, C. J. and Carter, C. W., Jr. (1995) *Structure*, **3,** 17–31.
5 Yin, Y. (1995) PhD thesis. University of North Carolina at Chapel Hill.
6 Sampson, J. R. and Uhlenbeck, O. C. (1988) *Proc. Natl. Acad. Sci. USA*, **85,** 1033–1037.
7 Carter, C. W., Jr. and Carter, C. W. (1979) *J. Biol. Chem*, **254,** 12219–12223.
8 Carter, C. W., Jr. (1990) *Methods: A Companion to Methods in Enzymology*, **1,** 12–24. This program is available from ftp://russell.med.unc.edu or from carter@med.unc.edu.
9 Carter, C. W., Jr. (1992) In Giegé, R., and Ducruix, A. (eds), *Crystallization of Proteins and Nucleic Acids, A Practical Approach*. IRL Press, Oxford University Press, Oxford, UK, pp. 47–71.
10 Hardin, R. H. and Sloane, N. J. A. (1993) *J. Stat. Plan. Inf.*, **37,** 339–369.
11 Grodberg, J. and Dunn, J. J. (1988) *J. Bacteriol.*, **170,** 1245–1253.
12 Yarus, M. and Berg, P. (1970) *Anal. Biochem.*, **35,** 450–465.
13 Milligan, J. F. and Uhlenbeck, O. C. (1989) *Methods Enzymol.*, **180,** 51–62.
14 Fuller, C. W. and Richardson, C. C. (1985) *J. Biol. Chem*, **260,** 3197–206.
15 Droge, P. and Pohl, F. M. (1991) *Nucleic Acids Res.*, **19,** 5301–5306.
16 Doublié, S., Xiang, S., Gilmore, C. J., Bricogne, G. and Carter, C. W., Jr. (1994) *Acta Cryst*, **A50,** 164–182.
17 Bricogne, G. (1993) *Acta Cryst*, **D49,** 37–60.
18 Wilkinson, L. (1987) 5.2.1 Ed. SYSTAT, Inc., Evanston, IL 60601.
19 Carter, C. W., Jr. (1993) *Annu. Rev. Biochem*, **62,** 715–745.
20 Perona, J. J. (1990) *Methods: A Companion to Methods in Enzymology*, **1,** 75–82.