REVIEW

PREDICTING THE SOLUBILITY OF RECOMBINANT PROTEINS IN *ESCHERICHIA COLI*

David L. Wilkinson and Roger G. Harrison*

School of Chemical Engineering and Material Science, University of Oklahoma, Norman, OK 73019. *Corresponding author.

We have studied the cause of inclusion body formation in *Escherichia coli* grown at 37°C using statistical analysis of the composition of 81 proteins that do and do not form inclusion bodies. Six composition derived parameters were used. In declining order of their correlation with inclusion body formation, the parameters are charge average, turn forming residue fraction, cysteine fraction, proline fraction, hydrophilicity, and total number of residues. The correlation with inclusion body formation is strong for the first two parameters but weak for the last four. This correlation can be used to predict the probability that a protein will form inclusion bodies using only the protein's amino acid composition as the basis for the prediction.

A significant barrier to the full exploitation of recombinant DNA technology for protein production is the tendency for the targeted protein to form inclusion bodies. Although inclusion bodies have been observed mainly in *Escherichia coli*, evidence suggests they will also be a problem in other expression systems if yields start to approach those obtained in *E. coli*. Inclusion bodies are *in vivo* agglomerations of proteins, appearing as large dense bodies in electron microscope pictures.[1] They require strong denaturants for dissolution such as urea or guanidine hydrochloride, behaving like proteins which have been irreversibly precipitated, as noted by Schein[2]. Obtaining a usable product from a protein which has been expressed in inclusion bodies requires that the protein be denatured and then refolded to the native form, a slow, difficult procedure which greatly reduces the net yield.

Inclusion bodies are rare in nature, with sickle cell anemia[3] and other related blood diseases being some notable exceptions to the rule that proteins are always expressed in soluble form. The reason inclusion bodies frequently form in *E. coli* is basically their high level of expression, giving a concentration level rare in nature. Recent work[4] has linked inclusion body formation to temperature sensitive denaturation and indicates lowering the expression temperature from 37°C to 30°C reduces their frequency.

Inclusion body formation has been the subject of reviews by Krueger et al.[5], Marston[6], Mitraki and King[7], and most recently, Schein[8]. Schein, citing protein expression data for 22 proteins from the system of Nagai and Thogersen[9] in which all factors were kept constant except protein sequence, concluded there must be some "solubilizing characteristics" in the six proteins expressed in soluble form. From this small number of proteins, however, she was not able to draw any firm conclusions on the nature of these "solubilizing characteristics". By expanding upon the Nagai and Thogersen system data using other proteins expressed under similar conditions, we obtained enough data to allow a thorough statistical analysis of a number of protein composition parameters in relation to inclusion body formation. The proteins used in this analysis are given in Table 1. All were expressed in *E. coli* at 37°C at levels of 3% or more of total cell protein.

DATA MODELING AND VARIABLE SELECTION

Statistical analysis requires a large data base and an appropriate methodology. The methodology that can be used here is somewhat limited by the way *in vivo* solubility is reported in the literature. The only thing that can usually be ascertained from cloning and overexpression research reports regarding *in vivo* solubility besides the expression level is whether or not the protein forms inclusion bodies, that is, whether the protein fraction of interest after centrifugation appears chiefly in the supernatant or in the pellet fraction. This limits the analysis to what differentiates the group of soluble proteins from the group of insoluble proteins and if and how we can determine if a protein will form inclusion bodies by looking at its amino acid composition. Data in this form is suited to the technique of discriminant analysis. This technique statistically identifies and categorizes the difference between specified parameters for two sets of data, in this case *in vivo* soluble and insoluble proteins. The protein sequences for the 81 proteins are on data bases and the composition could be compared residue by residue, but the results of such an analysis might be difficult to interpret. A more tractable approach is to compare the two groups of proteins on the basis of physicochemical properties that are determined by protein composition and are hypothesized to be related to *in vivo* solubility.

From studies of phage P22 tailspike protein[10] the formation of inclusion bodies has been explained as arising from incorrect folding and the precipitation of folding intermediates. Analyses of the protein composition therefore focus on factors that relate to protein folding and protein solubility. Protein folding will be discussed first.

Folding. At least two aspects of protein composition can

be shown to be related to its ability to fold correctly. First, an important difference between the expression of mammalian proteins in *E. coli* and in their native environment is the inability of *E. coli* to form disulfide bonds, due to the reducing environment of its cytoplasm. The fraction of cysteine in a mammalian protein expressed in *E. coli* therefore is a measure of the difficulty a protein may have in assuming its correct conformation in *E. coli*. In addition, one study suggested that disulfide bonds are present in some inclusion bodies[10], although this has been disputed[7]. Another factor in the ability of proteins to fold correctly is the number of turns. Turns are the most difficult structures for proteins to form, so a high content of residues with a high Chou and Fasman index[11] for

TABLE 1 Parameters of proteins studied.

Proteins	Residues	Fractions			Hydrophilicity Index	Approximate Charge Ave.	Ref.
		Turns	Proline	Cysteine			
Insoluble Proteins							
(cII fusion proteins)							
1. human α-globin	179	.212	.039	.005	−.014	.023	9
2. human β-globin	184	.217	.038	.011	−.027	.006	9
3. pancreatic ribonuclease A	169	.244	.026	.049	.217	.039	9
4. human myoglobin	191	.204	.026	.005	.236	.005	9
5. human c-*myc*	483	.297	.073	.021	.317	−.023	9
6. chicken β-actin	419	.225	.046	.014	.085	−.024	9
7. chicken myosin light chain	204	.211	.092	.010	.385	−.030	9
8. *Xenopus* histone 2A	168	.244	.029	.000	.329	.101	9
9. *Xenopus* histone 2B	163	.226	.036	.000	.418	.108	9
10. *Xenopus* TFIIIA	345	.201	.039	.066	.290	.073	9
11. tobacco mosaic virus coat	187	.237	.041	.005	−.018	.016	9
12. yeast mating α1	213	.244	.028	.004	−.014	.083	9
13. N-end human gelosin	444	.245	.042	.009	.190	−.005	9
14. yeast SWI-5	717	.315	.058	.012	.230	.014	9
15. caltrin	85	.258	.023	.000	.494	.082	9
16. pre-prosubtilisin	193	.235	.031	.000	.055	.087	9
(Other proteins)							
17. human interferon α2	165	.175	.026	.036	.045	−.012	8
18. human interferon γ	146	.192	.012	.041	.350	.062	8
19. human basic fib. growth factor	146	.265	.055	.027	.017	.068	43
20. ricin A	267	.252	.055	.007	−.197	.003	44
21. insulin growth factor type 1	70	.257	.071	.085	.067	.014	41
22. diphtheria toxin	535	.271	.041	.007	.066	−.021	45
23. diphtheria toxin/melanocyte hormone	526	.271	.039	.007	.063	−.018	45
24. diphtheria toxin N-terminal	223	.272	.026	.008	.247	−.027	45
25. Mx protein	630	.184	.034	.006	.246	−.011	8
26. bovine prochymosin	381	.265	.041	.011	−.354	.048	46
27. bovine prochymosin N-terminal	84	.238	.047	.018	−.106	−.029	10
28. phage p22 tailspike protein	667	.284	.041	.011	−.090	−.019	47
29. sv40 small *t* antigen	174	.160	.034	.063	.152	.012	35
30. γδ resolvase	222	.208	.018	.004	.151	.038	40
31. β-galactosidase/HSV-1	1389	.257	.058	.007	−.095	−.025	31
32. β-gal/penicillin fragment	888	.242	.074	.015	−.085	−.013	39
33. β-gal/insulin A	1044	.234	.061	.019	−.085	−.040	48
34. β-gal/insulin B	1053	.233	.059	.017	−.085	−.039	48
35. *B. subtilus* φ29 Protein 13	44	.204	.136	.068	.466	−.045	38
36. *Leishmania major* DHFR-TS	521	.194	.057	.019	.004	−.004	24
37. CAT/ANF PTCAN-11	205	.168	.034	.024	−.155	−.019	49
38. CAT/ANF PTCAX-42	241	.192	.029	.024	−.332	−.008	49
39. phage λO protein	300	.244	.040	.016	.283	.027	26
40. phage T4 protein 23	522	.232	.042	.001	−.043	−.016	7
41. interleukin-2	133	.180	.037	.015	−.050	.000	50
42. interleukin-4	129	.139	.007	.046	.170	.054	25
43. renin	340	.268	.040	.017	−.210	−.029	28
44. prorenin	363	.267	.043	.016	−.160	−.024	28
45. prorenin/linker fusion	367	.266	.045	.016	−.160	−.024	28
46. human macrophage colony stimulating factor	218	.248	.073	.041	.151	−.040	29
47. H-*ras*, N-terminal	167	.178	.035	.017	.185	−.042	34
48. β-gal/somatostatin	1020	.236	.060	.016	−.085	−.039	51
49. β-gal/endorphin	1047	.239	.062	.014	−.085	−.032	52
50. T4 *reg A*	103	.146	.024	.009	.200	.022	53
51. salmon growth hormone	188	.239	.026	.021	.010	−.005	54
52. γ heavy chain	439	.275	.077	.029	−.122	.000	27
53. bovine growth hormone	191	.183	.031	.020	.040	.005	55
54. HGH/β-lactamase	373	.217	.040	.015	.040	−.025	56
Soluble Proteins							
(cII fusion proteins)							
55. human tropomyosin	322	.099	.000	.003	.919	−.090	9
56. human gelosin	794	.239	.043	.006	.190	−.016	9
57. human gelosin, C-terminal	389	.231	.044	.002	.190	−.034	9
58. troponin T	340	.155	.026	.000	1.22	−.007	4
59. chicken troponin C	200	.165	.004	.005	.637	−.144	9
60. CGN4 Ca+ binding domain	105	.209	.034	.000	.550	.100	9

forming turns, i.e. aspartic acid, asparagine, proline, glycine, and serine, may be indicative of a slow folding protein. The folding of proline in particular has been shown in some cases to be the rate limiting step for the folding of certain proteins[12]. Therefore, we have used the following three folding related parameters: the cysteine fraction, the proline fraction, and the combined fraction of asparagine, proline, glycine, and serine (aspartic acid is omitted for statistical reasons to be discussed; see Statistical Methods).

Solubility. The *in vivo* solubility, that is whether or not the protein forms inclusion bodies, is different in at least one aspect from the *in vitro* solubility, being related more to the solubility of folding intermediates than to the solubility of the mature protein. However, when *comparing* the solubility *in vivo* of one protein to another, we assume that the same parameters which affect *in vitro* solubility can be used. Three parameters that are important in this respect are net charge, hydrophilicity, and size.

Protein solubility is predicted for low ionic content solutions by the Debye-Hückel equation[13], with the log of protein solubility being proportional to the square of the net protein charge. Protein solubility thus increases with increasing net charge, positive or negative. At neutral pH's the net charge of a protein is very close to the difference between the number of basic and acidic residues, that is the number of arginines and lysines minus the number of aspartates and glutamates. This quantity divided by the total number of residues approximates the average charge per residue in a protein and is called the approximate charge average. *In vivo*, the approximate charge average should be slightly more electropositive than the actual charge average. The reason for this is the effects of the slightly basic pH of *E. coli* (7.5–7.9)[14] and of the ionic double layer[15], which each have a slightly electronegative effect. The exact amount of this effect is unknown. The effect of charge can therefore be estimated by the parameter |approximate charge average − a|, where "a" is the difference between the approximate charge average and the actual charge average, "a" being a constant obtained by modeling the data (see Modeling Results).

The Debye-Hückel equation also contains a size related term, the protein radius[13]. While the size of a protein cannot be directly determined from the protein composition, being structure dependent, the number of residues in a protein is roughly related to molecular size and can be easily measured.

Hydrophilicity or hydrophobicity affects proteins in a different way. The repulsion between hydrophobic aliphatic and aromatic side chains on amino acid residues and water has been shown by Kauzmann[16] to be a key factor stabilizing protein structure. Although many of the hydrophobic groups in proteins are buried within the molecule to minimize the contact area with water, some do contact water, and these hydrophobic interactions affect protein solubility (see Arakawa and Timasheff[17]). The Hopp and Woods formula[18] provides a convenient way of calculating average hydrophilicity, assigning negative values to aliphatic residues like valine or aromatic residues like tyrosine, and positive values to ionized residues like arginine, and is used in this model.

To summarize, the 81 proteins listed in Table 1 can be compared using six easily computed composition related parameters, cysteine fraction, proline fraction, turn forming residue fraction, |approximate charge average − a|, number of residues, and hydrophilicity, through the technique of discriminant analysis. This technique is used to maximize the statistical differentiation between the insoluble and soluble proteins, using a composite parameter of the individual parameters called the canonical variable, or CV. This composite parameter consists of the sum of the individual parameters multiplied by their respective adjustable coefficients, or λ's, as given by the equation: $CV = \sum_{i}^{n} \lambda_i x_i$ where λ_i = adjustable coefficient, x_i = composition related parameter, n = number of parameters.

The λ's were calculated from the two sets of data using standard statistical techniques described by Fisher[19]. The protein compositions were obtained from the University of Wisconsin Genetic Computer Group programs[20].

MODELING RESULTS

Various statistical parameters for the modeling using the 81 proteins are shown in Tables 2 and 3. The effectiveness of a parameter in determining membership in a group can be estimated by analyzing the parameter $(\mu_{insol} - \mu_{sol})/\sigma_{pooled}$, or the difference in means divided by the pooled group standard deviation. Discriminant analysis optimizes the λ's so that $(\mu_{insol} - \mu_{sol})/\sigma_{pooled}$ is as large as possible, and thus maximizes the possibility of correctly identifying a protein on the basis of these parameters alone as being soluble or insoluble. An initial evaluation can therefore be made on a parameter on the basis of

TABLE 1 (continued).

Proteins	Residues	Fractions			Hydrophilicity Index	Approximate Charge Ave.	Ref.
		Turns	Proline	Cysteine			
(Other proteins)							
61. thioredoxin	108	.194	.042	.018	−.01	−.046	32
62. maltose binding protein	370	.224	.051	.000	.002	−.041	42
63. T7 RNA polymerase	883	.200	.038	.014	.016	−.002	4
64. β-galactosidase	1029	.234	.055	.015	−.08	−.039	36
65. T4 DNA ligase	487	.190	.026	.016	.134	−.045	4
66. T7 gene 19 protein	550	.206	.043	.000	.100	−.027	57
67. CAT/ANF PTCAX-11	210	.161	.030	.024	−.24	−.016	49
68. CAT/ANF PTCAX-92	245	.187	.025	.024	−.24	−.002	49
69. β-gal/hirudin	1088	.239	.054	.020	−.08	−.045	30
70. CAT/ANF PTCAX-82	243	.197	.029	.024	−.28	.000	49
71. *B. subtilus* φ29 protein 13	103	.126	.044	.000	.410	−.097	38
72. *B. subtilus* φ29 protein 10	124	.209	.007	.000	.096	−.033	38
73. T3 protein 18	89	.146	.020	.000	.167	−.079	23
74. T4 RNA ligase	374	.197	.021	.013	.134	−.045	37
75. human growth hormone	192	.229	.037	.020	.01	−.026	48
76. human interferon α_1	166	.168	.032	.030	.171	−.048	58
77. H-*ras*	189	.185	.028	.031	.203	−.031	33
78. H-*ras*, C-terminal	167	.155	.032	.036	.172	−.048	33
79. α_1-antitrypsin	394	.200	.039	.002	.065	−.031	59
80. murine TNF	156	.256	.052	.012	−.13	−.039	60
81. human TNF	157	.280	.063	.012	−.17	−.007	60

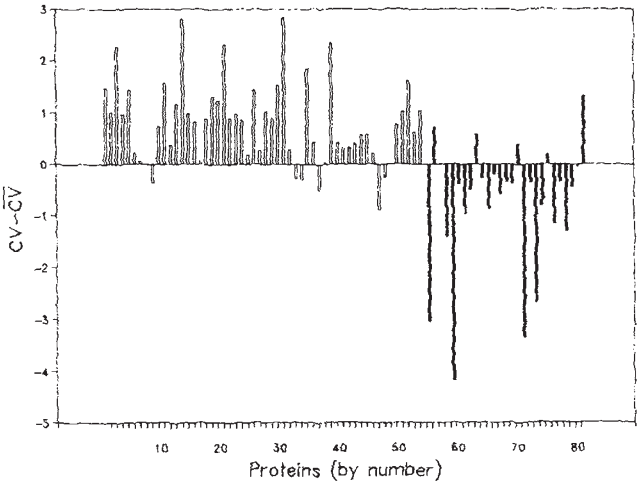

FIGURE 1 Canonical variable (CV) minus discriminant canonical value ($\overline{CV}$ = 1.15) for the proteins studied. Soluble proteins are shown with solid bars and insoluble proteins with open bars.

$(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$, before using it in the discriminant analysis calculations. The value for $(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$ for the size parameter was 0, so it was not used in the computer evaluation. The quantity 0.03 for "a" gives close to the highest value of $(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$ for |approximate charge average − a| and is within the expected range predicted by isoelectric point calculations and double layer theory[15]. The actual parameter used to measure charge is therefore |approximate charge average − 0.03|.

Values for other statistic parameters calculated for the individual parameters are also shown in Tables 2 and 3, with the first three rows representing values calculated on the individual composition parameters and the last two rows representing calculated values for the discriminant analysis equation. In Table 2 the turn fraction for the purposes of analysis is divided into two parts, the proline fraction and the fraction of the other three turn-forming residues, to determine if the proline fraction, as has been suggested[8], has any especially strong correlation with *in vivo* solubility as compared with the other three turn inducing residues used in the turn fraction compilation.

The F value calculated by the discriminant analysis computer program is used, as is $(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$, to estimate the statistical difference between groups. An F value of 1.0, as does a value for $(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$ of zero, means there is no statistical difference between groups. The F value may be used with the F distribution to statistically evaluate the significance of a given parameter's between group differences (see Statistical Methods).

The non-normalized λ's as stated were used directly in the CV equation. The normalized λ's give an estimation of the weighted contribution of each term in the equation, i.e., in terms of the magnitude of each λ adjusted for the size of its sequence parameter, by multiplying each λ by its parameter's pooled within-group variance. These results for normalized λ are consistent with the results for $(\mu_{insol}-\mu_{sol})/\sigma_{pooled}$, F, and the level of significance.

In Table 3 the proline fraction and the remaining turn fraction were combined into one parameter, as the proline fraction as seen in Table 2 contributes no more than its proportional share to the weighted λ relative to the other three residues in the turn fraction. As the separate consideration of proline does not add any accuracy and complicates the computer calculation, it is the Table 3 values which are used for the results that follow.

Values of (CV-$\overline{CV}$), where $\overline{CV}$ is the value of the discriminant (1.15), are shown in Figure 1. The value of CV is used to determine the statistical probability that a given protein will be soluble or insoluble, as shown in Figure 2. High values of |CV-$\overline{CV}$| mean the protein can be determined to be soluble or insoluble with a high degree of confidence. Low values of |CV-$\overline{CV}$| mean the protein is in the borderline region of solubility and has a roughly comparable chance of being either soluble or insoluble. Using the percentage probabilities to classify proteins as soluble or insoluble, discriminant analysis successfully classifies 22 of the 27 soluble proteins and 49 of the 54 insoluble proteins, for an overall accuracy of over 88%.

INTERPRETATION

The values for the λ's in the canonical variable equation as determined by discriminant analysis give the optimum formula for estimating the *in vivo* solubility for proteins. The probable solubility of other proteins which have not yet been expressed in *E. coli* can be evaluated, if the amino acid composition is known, by calculating the parameters as described above and inserting them into the discriminant analysis equation. The resulting value of CV may be used to obtain the probability of *in vivo* protein solubility or insolubility using Figure 2. CV values may also provide a good relative comparison of solubility for proteins expressed under different conditions than those specified for the Table 1 proteins; for instance if a protein expressed under a given set of conditions is insoluble, another protein with a higher CV expressed under the same conditions would likely also be insoluble.

These probability calculations concerning the chances of inclusion body formation should be helpful to researchers in microbial protein production. It should be possible after estimating the solubility of a cloned protein by calculations of the above type to plan the cloning procedure, especially promotor selection, fusion protein partner, and expression temperature accordingly to produce soluble or insoluble protein. For instance, a highly electronegative, low CV fusion protein partner for targeted proteins in lieu of the commonly used protein β-galactosidase (for which (CV-$\overline{CV}$) = −0.39) would likely be more efficient in producing soluble proteins.

In general, discriminant analysis adjusts the λ's so that the terms with the strongest correlation with membership in the different groups, in this case solubility or insolubility, have the strongest effect on the canonical variable or CV, while terms with a weaker correlation have a lesser effect on CV. The weighted λ's give a close estimate of the relative importance of each parameter in the CV equation. As can be seen from Table 3, the turn fraction and |approximate charge average − 0.03| make up most of the

TABLE 2 Statistical model: Results for five parameters.

Statistic Parameter	Fraction: Turn-Forming Residues (w/o Pro)	Fraction: Proline	Fraction: Cysteine	Hydrophilicity Index	|Approximate Charge Average −0.03|
$(\mu_{in}-\mu_{s})/\sigma_{pooled}$	0.88	0.35	0.54	−0.26	−1.37
"F" Value	13.86	2.16	3.53	2.03	32.78
Level of Significance	100%	85%	95%	85%	100%
Adjustable Coefficient (λ_i)	15.20	4.05	9.36	0.73	−33.27
Normalized or Weighted λ_i	0.50	0.08	0.15	0.19	−0.81

value of CV, the contribution of the other terms being small, indicating these two terms largely determine *in vivo* solubility. This is in line with their much stronger individual correlation with *in vivo* solubility as seen by their high F values and significance levels. The F values and significance levels for the cysteine fraction and hydrophilicity are low indicating a low level of correlation, and their contribution to the CV is correspondingly small.

It is interesting to relate these results to the model proposed by Mitraki and King[7] for inclusion body formation:

$$\text{translation} \rightarrow I^{pf} \leftrightarrow I^{pm} \rightarrow \text{native}$$
$$\updownarrow$$
$$I^{pf*}$$
$$\downarrow$$
$$\text{inclusion body}$$

where: I^{pf} is a soluble partially folded early intermediate
I^{pm} is the intermediate competent to form into the monomer
I^{pf*} is the species that generates the aggregate (i.e. inclusion body)

From this model it is clear how the *in vivo* solubility of proteins is affected both by folding rates and solubility factors. As an example of the effect of folding rates on solubility, a protein with many turn forming residues folds more slowly, hence has a higher concentration of folding intermediates, which increases the probability of these intermediates precipitating to form inclusion bodies. As an example of the effect of solubility factors, the folding intermediates of a highly electronegative protein should be more soluble than average and therefore less likely to precipitate to form inclusion bodies than less electronegative, less soluble proteins.

STATISTICAL METHODS

The statistical methods used in this analysis assume a standard Gaussian distribution, which however is not a critical assumption if the data is not too badly skewed. The discriminant analysis technique used assumes the individual parameters act independently, which for these parameters is substantially true. The value of the turn forming residue fraction does not include aspartatic acid to eliminate interference with the charge average, which also contains aspartic acid.

The pooled standard deviation combines the standard deviations from two different but statistically similar populations by adding their variances, i.e., for two populations 1 and 2 of size N_1 and N_2 and means of μ_1 and μ_2 respectively:

$$\sigma_{pooled} = \left(\frac{\sum(x_{i1} - \mu_1)^2 + \sum(x_{i2} - \mu_2)^2}{(N_1 + N_2 - 2)} \right)^{1/2}$$

The F value, like $(\mu_1 - \mu_2)/\sigma_{pooled}$, provides a test of the null hypothesis (that is, that there is no statistically significant between-group difference) but in a statistically quantifiable manner. It does so by comparing the between-group variance to the within-group variance. F values for n degrees of freedom are tabulated in standard statistical tables. If the calculated F value is larger than the tabulated value for a given confidence level, the between-group differences are statistically significant at that confidence level. If μ_T equals the combined group mean, the F value is given by the following equation[21]:

$$F = \frac{[N_2(\mu_1 - \mu_T)^2 + N_2(\mu_2 - \mu_T)^2]\,(N_1 + N_2 - 2)}{(N_1 - 1)\sigma^2_1 + (N_2 - 1)\sigma^2_2}$$

The discriminant analysis computer calculations were performed by the discriminant analysis program written by BMDP Statistical Software, Inc[22]. In addition to calculating the CV for a protein the computer also calculates the posterior probability that the protein will be soluble by using ratios called the Mahalonobis distances, which are derived from the covariance matrices used in the discriminant analysis calculations. Basically, if the CV for a protein is close to the $\overline{CV}$ value of 1.15, the posterior probability that the protein will be soluble or insoluble will be close to 0.5. If on the other hand the CV for a protein is far from the $\overline{CV}$ value of 1.15, the posterior probability that the protein is soluble or insoluble will be near unity. These computer calculated values were used to prepare Figure 2.

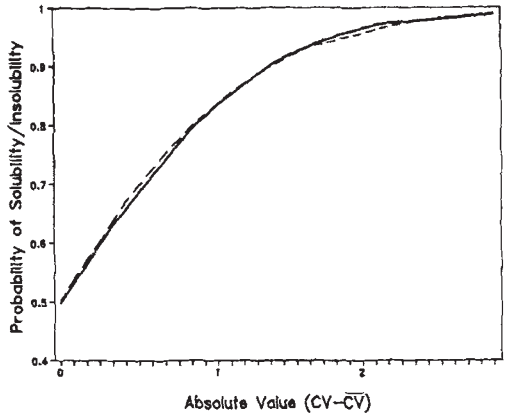

FIGURE 2 Predicted probability of protein solubility or insolubility based on canonical value of protein. $\overline{CV}$ = 1.15. If (CV−$\overline{CV}$) is positive, use dashed line for protein insolubility. If (CV−$\overline{CV}$) is negative, use solid line for protein solubility.

Acknowledgments

We would like to thank Jeff Harwell, Adele Hughes, Ed O'Rear, Bruce Roe and Dick van der Helm of the Univ. of Oklahoma for constructive criticism and advice. This work was partially supported by a grant-in-aid from the American Heart Association, Oklahoma Affiliate.

References

1. Williams, D. C., Van Frank, R. M., Muth, W. L. and Barnett, J. P. 1982. Cytoplasmic inclusion bodies in *Escherichia coli* producing biosynthetic human insulin protein. Science **215**:684–687.
2. Schein, C. H. 1990. Solubility as a function of protein structure and solvent components. Bio/Technology **8**:308–317.
3. Carrell, R. W., Lehmann, H. and Hutchisdon, H. F. 1966. Haemoglo-

TABLE 3 Statistical model: Results for four parameters.

Statistic Parameter	Fraction: Turn-Forming Residues (with Pro)	Fraction: Cysteine	Hydrophilicity Index	|Approximate Charge Average −0.03|
$(\mu_{in} - \mu_s)/\sigma_{pooled}$	0.98	0.54	−0.26	−1.37
"F" Value	17.19	3.53	2.03	32.78
Level of Significance	100%	95%	85%	100%
Adjustable Coefficient (λ_i)	13.40	5.83	0.80	−33.25
Normalized or Weighted λ_i	0.54	0.10	0.20	−0.81

bin Koln (beta 98 valine-methionine): an unstable protein causing inclusion body anemia. Nature **210**:915–916.
4. Schein, C. H. and Noteborn, M. H. M. 1988. Formation of soluble recombinant proteins in *Escherichia coli* is favored by lower growth temperature. Bio/Technology **6**:291–294.
5. Krueger, J. K., Kulke, M. N., Schutt, C. and Stock, J. 1989. Protein inclusion body formation and purification. BioPharm. **2**:40–45.
6. Marston, F. A. O. 1986. The purification of eukaryotic polypeptides synthesized in *Escherichia coli*. Biochem. J. **240**:1–12.
7. Mitraki, A. and King, J. 1989. Protein folding intermediates and inclusion body formation. Bio/Technology **7**:690–697.
8. Schein, C. H. 1989. Production of soluble proteins in bacteria. Bio/Technology **7**:1141–1149.
9. Nagai, K., Thogersen, H. C. and Luisi, B. F. 1988. Refolding and crystallographic studies of eukaryotic proteins produced in *Escherichia coli*. Biochem. Soc. Trans. **16**:108–110.
10. Brasnet, A. H., Schoemaker, J. M. and Marston, F. A. O. 1985. Examination of calf prochymosin accumulation in *Escherichia coli:* disulfide linkages are a structural component of prochymosin containing inclusion bodies. EMBO J. **4**:775–780.
11. Chou, P. Y. and Fasman, G. D. 1978. Empirical predictions of protein conformation. Ann. Rev. Biochem. **47**:251–276.
12. Evans, P. A., Dobson, C. M., Kautz, R. A., Hatfull, G. and Fox, R. O. 1987. Proline isomerism in staphylococcal nuclease characterized by NMR and site-directed mutagenesis. Nature **329**:266–268.
13. Tanford, C. 1958. Physical Chemistry of Macromolecules. John Wiley & Sons, New York.
14. Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., and Schaechter, M. (Eds.). 1987. *Escherichia coli* and *Salmonella typhimurium*. Cellular and Molecular Biology, Volume 2, p. 222–243. American Society for Microbiology. Washington, D.C.
15. Riddick, T. 1968. Control of Colloid Stability Through Zeta Potential. Livingston Publishing Company, Wynnewood, Pennsylvania.
16. Kauzmann, W. 1959. Structural factors involved in protein stability. Advances in Protein Chemistry. **14**:19–29.
17. Arakawa, T. and Timasheff, S. N. 1985. Theory of protein solubility. Meth. Enzym. **114**:49–77.
18. Hopp, T. P. and Woods, K. R. 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78**:3824–3828.
19. Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**:179–188.
20. Devereux, J., Haeberli, P. and Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acid Res. **12**:387–395.
21. King, R. S. and Bryant, J. 1982. Applied Statistics Using the Computer. Mayfield Publishing Company, Palo Alto, CA.
22. Dixon, W. J. 1988. BMDP Statistical Software Manual. University of California Press, Berkeley.
23. Hamuda, K., Furisawa, H. and Minagawa, T. 1986. Overproduction and purification of the products of bacteriophage T3 genes 18 and 19. Virology **151**:110–118.
24. Grumont, R., Worachart, S. and Sant, D. V. 1988. Heterologous expression of the bifunctional thymidylate synthase-dihydrofolate reductase from *Leishmania major*. Biochemistry **27**:3776–3780.
25. Bond, J. H., Van Kimnenade, M. W., Schumacher, C. and Kastelein, R. A. 1988. Expression, renaturation, and purification of recombinant human Il-4 in *E. coli*. Biochem. J. **174**:109–114.
26. Roberts, J. D. and McMacken, R. 1983. The bacteriophage λO replication protein: isolation and characterization of the amplified initiator. Nucleic Acids Res. **11**:7435–7452.
27. Cabilly, S., Riggs, A. D., Shivelly, J., Holmes, W., Rey, M., Perry, L. J., Heyneker, H. L. and Wetzel, R. 1984. Generation of antibody activity from immunoglobulin polypeptide chains produced in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **81**:3273–77.
28. Sharma, S. K., Evans, D. B., Tomich, C., Cornette, J. C. and Ulrich, R. G. 1987. Folding and reactivation of recombinant human prorenin. Biotechnology and Applied Biochemistry **9**:181–193.
29. Halenbeck, R., Kawasaki, E., Wrin, J. and Koths, K. Renaturation and purification of biologically active recombinant human macrophage colony-stimulating factor expressed in *E. coli*. Bio/Technology **7**:710–715.
30. Bergman, C., Dodt, J., Fink, E., Gassen, H. G. and Kohler, S. 1986. Chemical synthesis and expression of a gene coding for hirudin, the thrombin-specific inhibitor from the leech *Hirudo medicinalis*. Biol. Chem. Hoppe-Seyler. **367**:731–740.
31. Weis, J. H., Enquist, L. W., Salstrom, J. S. and Watson, R. J. 1983. An immunologically active chimaeric protein containing herpes simplex virus type 1 glycopotein D. Nature **302**:72–75.
32. Lunn, C. A., Kathju, S., Wallace, B. J., Kushner, S. R. and Pigiet, V. 1984. Amplification and purification of plasmid-encoded thioredoxin from *Escherichia coli* K12. J. of Biol. Chem. **259**:10469–10474.
33. Gross, M., Sweet, R. W., Sathe, G., Yokoyama, S., Fasano, O., Goldfarb, M., Wigler, M. and Rosenberg, M. 1985. Purification and characterization of human H-*ras* proteins expressed in *Escherichia coli*. Molecular and Cellular Biology **5**:1015–1024.
34. Lacal, J. C., Santos, E., Notario, V., Barbacid, M., Yamazaki, S., Kung, H., Seamans, C., McAndrew, S. and Crowl, R. 1984. Expression of normal and transforming H-*ras* genes in *Escherichia coli* and purification of their encoded p21 proteins. Proc. Natl. Acad. Sci. USA **81**:5305–5309.
35. Derom, C., Gheysen, D. and Fiers, W. 1982. High level synthesis in *Escherichia coli* of the SV-40 small *t* antigen under control of the bacteriophage λp_L promoter. Gene **17**:45–54.
36. Bastia, D., Germino, J., Mukherjee, S. and Vanaman, T. 1986. New approaches to the expression and isolation of a regulatory protein. Genetic Eng. **8**:333–351.
37. Rand, K. N., Singh, M., Thogersen, H. C. and Gait, M. J. 1986. T4 RNA ligase: New structural studies on an unusual but useful enzyme. Proc. Int. Symp. Biomol. Structural Interactions, Supp. J. Biosci. **8**:89–100.
38. Ibanez, C., Garcia, J. A., Carrascosa, J. L. and Salas, M. 1984. Overproduction and purification of the connector protein of *Bacillus subtilus* phage φ29. Nucleic Acids Res. **12**:2351–2366.
39. Ayala, J. A., Pla, J., Desviat, L. R. and De Pedro, M. A. 1988. A *lacZ-pbp*B gene fusion coding for an inducible hybrid protein that recognizes localized sites in the inner membrane of *Escherichia coli*. J. Bacteriol. **170**:3333–3341.
40. Reed, R. R. 1981. Transposon-mediated site-specific recombination: a defined *in vitro* system. Cell **25**:713–719.
41. Schulz, M., Buell, G., Schmid, E., Movva, R. and Selzer, G. 1987. Increased expression in *Escherichia coli* of a synthetic gene incoding human somatomedin C after gene duplication and fusion. J. Bacteriol. **169**:5385–5392.
42. Maina, C. V., Riggs, P. D., Grandea, A. G., Slatko, B. E., Moran, L. S., Tagliamonte, J. A., McReynolds, L. A. and di Guan, C. 1988. An *Escherichia coli* vector to express and purify foreign proteins by fusion to and separation from maltose-binding protein. Gene **74**:365–373.
43. Squires, C. H., Childs, J., Eisenberg, S. P., Polverini, P. J. and Sommer, A. 1988. Production and characterization of human basic fibroblast growth factor from *Escherichia coli*. J. Biol. Chem. **263**:16297–16302.
44. Piatek, M., Laird, W., Bjorn, M. J., Wang, A. and Williams, M. 1988. Expression of soluble and fully functional ricin A chain in *Escherichia coli* is temperature sensitive. J. Biol. Chem. **263**:4837–4843.
45. Bishai, W. R., Rappuoli, R. and Murphy, J. R. 1987. High-level expression of a proteolytically sensitive diphtheria toxin fragment in *Escherichia coli*. J. Bacteriol. **169**:5140–5151.
46. McCamen, M. T. 1989. Fragments of prochymosin produced in *Escherichia coli* form insoluble inclusion bodies. J. Bact. **171**:1225–1227.
47. Haase-Pettingell, C. A. and King, J. 1988. Formation of aggregates from a thermolabile *in vivo* folding intermediate in P22 tailspike maturation. A model for inclusion body formation. J. Biol. Chem. **263**:4977–4983.
48. Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Ross, M. J., Miozzari, G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K. and Riggs, A. D. 1979. Expression in *Escherichia coli* of chemically synthesized genes for human insulin. Proc. Natl. Acad. Sci. U.S.A. **76**:106–110.
49. Dykes, C. W., Brookless, A. B., Coomber, B. A., Noble, S. A., Humber, D. C. and Hobden, A. N. 1988. Expression of atrial natriuretic factor as a cleavable fusion protein with chloramphenical acetyltransferase in *Escherichia coli*. Eur. J. Biochem. **174**:398–410.
50. Weir, M. P. and Sparks, J. 1987. Purification and renaturation of recombinant human interleukin-2. Biochem. J. **245**:85–91.
51. Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F. and Boyer, H. W. 1977. Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. Science **198**:1056–1063.
52. Shine, J., Fettes, I., Lan, N. C. Y., Roberts, J. L. and Baxter, J. D. 1980. Expression of cloned β-endorphin sequences by *Escherichia coli*. Nature **285**:456–461.
53. Adari, H. Y., Rose, K., Williams, K. R., Konisberg, W. H., Lin, T.-C. and Spicer, E. K. 1985. Cloning, nucleotide sequence, and overexpression of the bacteriophage T4 *regA* gene. Proc. Natl. Acad. Sci. USA **82**:1901–1905.
54. Sekine, S., Mizukami, T., Nishi, T., Kuwana, Y., Saito, A., Sato, M., Itoh, S. and Kawauchi, H. 1985. Cloning and expression of cDNA for the salmon growth hormone in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. **82**:4306–4310.
55. Seeburg, P. H., Shine, J., Martial, J. A., Ivarie, R. D., Morris, J. E., Ultrich, A., Baxter, J. D. and Goodman, H. M. 1978. Cloning and expression of cDNA for bovine growth hormone in *Escherichia coli*. Nature **276**:795–798.
56. Szoka, P. R., Schreiber, A. B., Chan, H. and Murthy, J. 1986. A general method for retrieving the components of a genetically engineered fusion protein. DNA **5**:11–20.
57. White, J. H. and Richardson, C. C. 1988. Gene 19 of bacteriophage T7. Overexpression, purification, and characterization of its product. J. Biol. Chem. **263**:2469–2476.
58. Wetzel, R., Heyneker, H. L., Goeddel, D. V., Juhrani, P., Shapiro, J., Crea, R., Low, T. L., McClure, K., Thurman, J. E. and Goldstein, A. L. 1981. Production of biologically active Nα-desacetyl thymosin α *in Escherichia coli* through expression of a chemically synthesized gene, p. 251–266. *In*: Cellular Responses to Molecular Modulators. Mozes, L. W., Schultz, J., Scott, W. A., and Werner, R., (Eds.). Academic Press, New York.
59. Courtney, M., Buchwalder, A., Kohli, V., Lathe, R., Tolstoshev, P. and Lecocq, J.-P. 1984. High-level production of biologically active human α_1-antitrypsin in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. **81**:669–673.
60. Pennica, D., Hayflick, J. S., Bringman, T. S., Palladina, M. A. and Goeddel, D. V. 1985. Cloning and expression in *Escherichia coli* of the cDNA for murine tumor necrosis factor. Proc. Natl. Acad. Sci. U.S.A. **82**:6060–6064.