COMPLEX SYSTEM MODELING WITH STATISTICAL METHODS

John J. Helly, Jr., Edward C. DeLand
Departments of Computer Science and Anesthesiology
University of California, Los Angeles

Modeling of complex systems can be greatly facilitated by the inclusion of empirical data directly into the solution of the model. Data can then be used to provide information about the fidelity of the model (goal) to the real system and/or act as a temporary model component for a subsystem not yet well-defined (probe).

This method utilizes existing, highly-developed statistical packages to reduce development effort as well as obtain valuable statistical information useful in model validation. An example of the method applied to a molecular model of hemoglobin is provided.

1. Introduction

One of the most difficult and central problems in mathematical and computer simulation of complex systems is the incorporation of empirical data by the computer model. Such data is necessary as a simulation goal and as a probe for system identification. As a goal, data represent the output of a real-world system which the modeler is attempting to represent mathematically. As a probe, the data represent criteria by which model components can be evaluated. these two viewpoints are not entirely distinct. The latter perspective views the model as simply composed of sub-models, one for each identifiable subsystem. However, models are often not simply decomposable into smaller components especially during development when the components may not yet be identified or adequately represented. The utility of data inclusion in the modeling process then has a dual function, as both a goal and probe. Conceptually, these functions are quite different and it is, therefore, convenient to retain the distinction between them.

Computer models of biological systems tend to be large, complex, and highly parameterized systems of equations which face the modeler with at least three major problems. First, an explicit representation of the system to be simulated must be derived. This requires the collation of unrelated results from different workers in different laboratories as a basis of system identification and model definition. Second, initial values must be provided as a starting point for the simulation. If precise values are not available then a method for their estimation must be provided. Third, the results of the simulation experiment must be analyzed and evaluated with respect to the behavior of the real system being modeled. Based on the analysis, either the simulation results are accepted or one or more of the steps above are repeated until acceptance occurs or the model is discarded.

This paper describes a powerful technique for the inclusion of empirical data in the modeling process in each of these stages of development. The method allows the investigator to examine the behavior of specific subsystems in the model for which empirical data are available (goal) and to measure the accuracy of the model with respect to the data (probe). The investigator may also substitute data for a model component when an explicit description of that component is not available (identification/representation).

This technique yields a method for combining standard statistical procedures with usual modeling techniques to improve

81CH1709-5/81/0000-0611$00.75 © 1981 IEEE

and clarify model design. As an example, the method is presented in the context of a steady-state, molecular model of hemoglobin. The hemoglobin example illustrates the use of this method in parameter estimation. In the sense of the previous discussion this is a goaling strategy. It is important to realize that the emphasis of this discussion focuses on an approach to modeling and not on any particular implementation. The hemoglobin example is presented simply to help elucidate this idea and its potential for facilitating the interaction between theory and data.

2. Statistical Methodology

Least-squares regression models employing standard statistical procedures contend with the same three problems described above for model development since a regression function is merely a model of the behavior of a dependent variable with respect to an independent variable. However, regression usually involves only a few closed form expressions with a relatively small number of parameters to be estimated. In this case the function is regressed against a set of observations. The success or failure of the model can be judged by the value of the residual sum of squares after the regression is performed.

Usually, regression is used to investigate a hypothesis about the system from which the data is obtained; linearity of the system, for example. In this sense, regression is trivially a goaling procedure in which the goal is the accurate simulation of experimental data by the regression function. However, when we incorporate the use of regression into the solution of a large, complex model, the goal in the regression procedure may represent only a small component of the overall solution sought by the simulation. Then the regression procedure acts as a constraint on the model by holding, or attempting to hold model parameters at values consistent with experimental data. In this sense the regression is a probe in that it represents a subsystem of the model for which there is not yet an explicit, deterministic representation.

Two difficulties with the use of regression in this way are the general non-linearity of large, biological models and their lack of a closed form expression. A few statistical software packages, notably BMDP (Dixon, 1979) and SAS (Barr, 1979), currently provide efficient procedures for handling non-linear problems. The absence of a closed form solution is not a difficult problem. If a numerical value can be computed by the model to be passed to the statistical program as if it were the output of the regression function, the regression can proceed as usual. The output of the model appears exactly as if it were computed by an explicit regression function. In the hemoglobin model, the oxygen saturation curve is the experimental data and the entire model is the regression function.

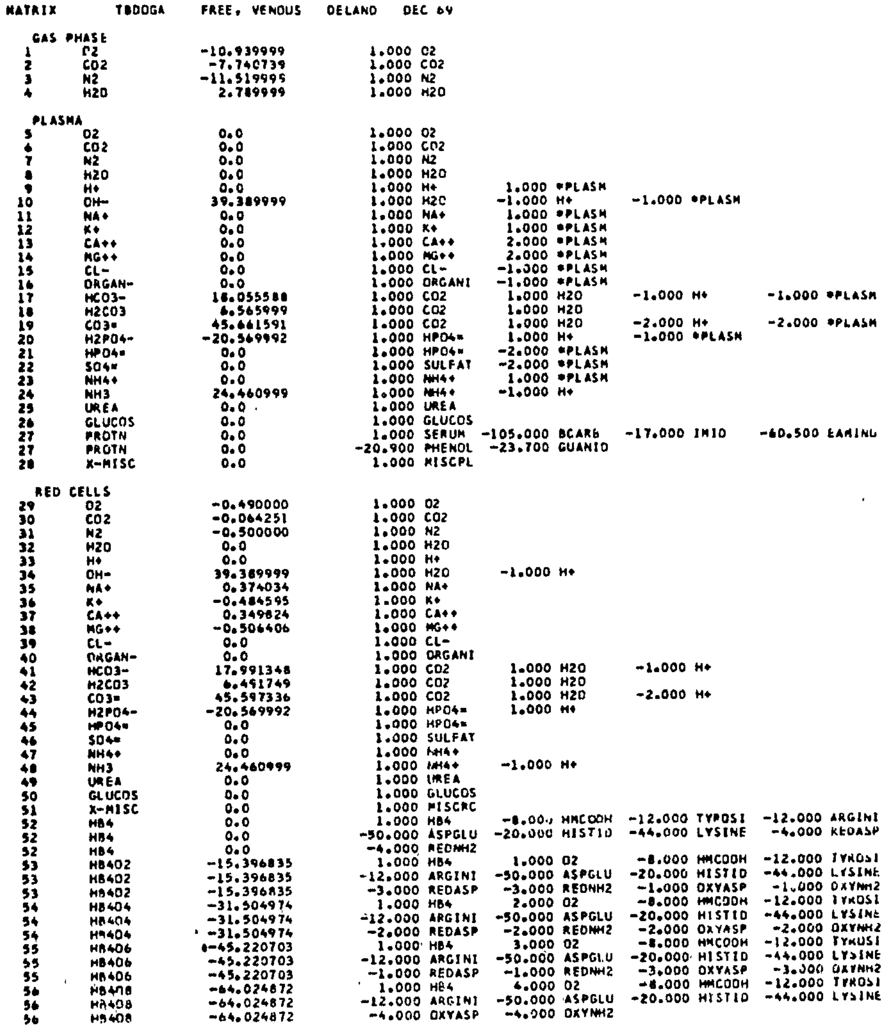
Regression analysis provides parameter estimates, the principle reason for its use, but can also provide confidence intervals for the parameters, a correlation matrix between parameters, and a residual sum of squares. The residual sum is an estimate of the accuracy of the model with respect to the data and condence intervals provide a measure of the precision of the model on the basis of the parameter estimates. These statistics permit the comparison of model results on the basis of hypothesis tests between parameter estimates obtained from different sets of data or from other models. Residual and variable plots are also available and these are useful in providing insight into the general behavior of the model.

4. Modeling Methodology and the Hemoglobin Model

The hemoglobin molecule is one of the most completely studied biological molecules known to science (Perutz, 1970; Monod et al, 1963). Although models exist which do quite well in predicting saturation curves they do not attempt to relate the structure of hemoglobin to its behavior except in a very general way (Fell, 1978; Seaton, 1974).

Using the steady-state modeling system CHEMIST (DeLand, 1967), Hemoglobin is described as a list of chemical equations (Fig. 1). Each reaction is described by its stoichiometry and an estimated equilibrium constant. These reactions are logically grouped according to their role in hemoglobin function. Consequently, some reactions are grouped as plasma, red cells, oxidized heme reactions, DPG binding, etc. Only part of the complete model is shown. There are four oxygen binding constants which require estimation in this model, one for each subunit of the hemoglobin molecule.

Empirical data used in the estimation of the binding constants is taken from Severinghaus (1966). The data are ordered pairs of values for the percentage oxygen saturation of hemoglobin at selected values of partial pressure of oxygen. A model run uses fifteen data points selected from the complete saturation curve. These are selected to facilitate the regression procedure and may be weighted. Only a small number of points are selected in order to reduce the time per iteration of the regression. There is no theoretical limit to the amount of data that may be used but there is a tradeoff between



```
MATRIX        TBDOGA    FREE, VENOUS   DELAND   DEC 69

  GAS PHASE
 1      O2        -10.939999       1.000 O2
 2      CO2        -7.740739       1.000 CO2
 3      N2        -11.519995       1.000 N2
 4      H2O         2.789999       1.000 H2O

  PLASMA
 5      O2          0.0            1.000 O2
 6      CO2         0.0            1.000 CO2
 7      N2          0.0            1.000 N2
 8      H2O         0.0            1.000 H2O
 9      H+          0.0            1.000 H+        1.000 *PLASM
10      OH-        39.389999       1.000 H2O      -1.000 H+        -1.000 *PLASM
11      NA+         0.0            1.000 NA+       1.000 *PLASM
12      K+          0.0            1.000 K+        1.000 *PLASM
13      CA++        0.0            1.000 CA++      2.000 *PLASM
14      MG++        0.0            1.000 MG++      2.000 *PLASM
15      CL-         0.0            1.000 CL-      -1.000 *PLASM
16      ORGAN-      0.0            1.000 ORGANI   -1.000 *PLASM
17      HCO3-      18.055588       1.000 CO2       1.000 H2O       -1.000 H+        -1.000 *PLASM
18      H2CO3       6.565999       1.000 CO2       1.000 H2O
19      CO3=       45.661591       1.000 CO2       1.000 H2O       -2.000 H+        -2.000 *PLASM
20      H2PO4-    -20.569992       1.000 HPO4=     1.000 H+        -1.000 *PLASM
21      HPO4=       0.0            1.000 HPO4=    -2.000 *PLASM
22      SO4=        0.0            1.000 SULFAT   -2.000 *PLASM
23      NH4+        0.0            1.000 NH4+      1.000 *PLASM
24      NH3        24.460999       1.000 NH4+     -1.000 H+
25      UREA        0.0            1.000 UREA
26      GLUCOS      0.0            1.000 GLUCOS
27      PROTN       0.0            1.000 SERUM  -105.000 BCARB     -17.000 IMID      -60.500 EAMINO
27      PROTN       0.0          -20.900 PHENOL  -23.700 GUANID
28      X-MISC      0.0            1.000 MISCPL

  RED CELLS
29      O2         -0.490000       1.000 O2
30      CO2        -0.064251       1.000 CO2
31      N2         -0.500000       1.000 N2
32      H2O         0.0            1.000 H2O
33      H+          0.0            1.000 H+
34      OH-        39.389999       1.000 H2O      -1.000 H+
35      NA+         0.374034       1.000 NA+
36      K+         -0.484595       1.000 K+
37      CA++        0.349824       1.000 CA++
38      MG++       -0.506406       1.000 MG++
39      CL-         0.0            1.000 CL-
40      ORGAN-      0.0            1.000 ORGANI
41      HCO3-      17.991348       1.000 CO2       1.000 H2O       -1.000 H+
42      H2CO3       6.451749       1.000 CO2       1.000 H2O
43      CO3=       45.597336       1.000 CO2       1.000 H2O       -2.000 H+
44      H2PO4-    -20.569992       1.000 HPO4=     1.000 H+
45      HPO4=       0.0            1.000 HPO4=
46      SO4=        0.0            1.000 SULFAT
47      NH4+        0.0            1.000 NH4+
48      NH3        24.460999       1.000 NH4+     -1.000 H+
49      UREA        0.0            1.000 UREA
50      GLUCOS      0.0            1.000 GLUCOS
51      X-MISC      0.0            1.000 MISCRC
52      HB4         0.0            1.000 HB4      -8.000 HMCOOH   -12.000 TYROSI   -12.000 ARGINI
52      HB4         0.0          -50.000 ASPGLU  -20.000 HISTID   -44.000 LYSINE    -4.000 REDASP
52      HB4         0.0           -4.000 REDNH2
53      HB4O2     -15.396835       1.000 HB4       1.000 O2        -8.000 HMCOOH   -12.000 TYROSI
53      HB4O2     -15.396835     -12.000 ARGINI  -50.000 ASPGLU   -20.000 HISTID   -44.000 LYSINE
53      HB4O2     -15.396835      -3.000 REDASP   -3.000 REDNH2    -1.000 OXYASP    -1.000 OXYNH2
54      HB4O4     -31.504974       1.000 HB4       2.000 O2        -8.000 HMCOOH   -12.000 TYROSI
54      HB4O4     -31.504974     -12.000 ARGINI  -50.000 ASPGLU   -20.000 HISTID   -44.000 LYSINE
54      HB4O4     -31.504974      -2.000 REDASP   -2.000 REDNH2    -2.000 OXYASP    -2.000 OXYNH2
55      HB4O6     -45.220703       1.000 HB4       3.000 O2        -8.000 HMCOOH   -12.000 TYROSI
55      HB4O6     -45.220703     -12.000 ARGINI  -50.000 ASPGLU   -20.000 HISTID   -44.000 LYSINE
55      HB4O6     -45.220703      -1.000 REDASP   -1.000 REDNH2    -3.000 OXYASP    -3.000 OXYNH2
56      HB4O8     -64.024872       1.000 HB4       4.000 O2        -8.000 HMCOOH   -12.000 TYROSI
56      HB4O8     -64.024872     -12.000 ARGINI  -50.000 ASPGLU   -20.000 HISTID   -44.000 LYSINE
56      HB4O8     -64.024872      -4.000 OXYASP   -4.000 OXYNH2
```

Figure 1. Partial Listing of CHEMIST Hemoglobin Model.

computing time and increased precision of the parameter estimates obtained from a larger sample size. Initial values for the binding parameters are obtained from DeLand (1970).

The incorporation of the hemoglobin model into the regression procedure BMDP3R (Dixon, 1979) is illustrated in Figure 2. The interface between BMDP3R and CHEMIST passes model parameters as well as control information to BMDP3R. Partial derivatives of oxygen saturation with respect to each of the oxygen binding constants are obtained from the Jacobian matrix computed by CHEMIST. These values are determined for each data point during each iteration of the regression procedure, hence the desire to minimize the number of data points if computing time is costly.

When the regression procedure has altered the model parameters, during its parameter search, the model is executed again to obtain new steady-state values in accord with the new parameter estimates. The procedure continues until the convergence criteria of the regression

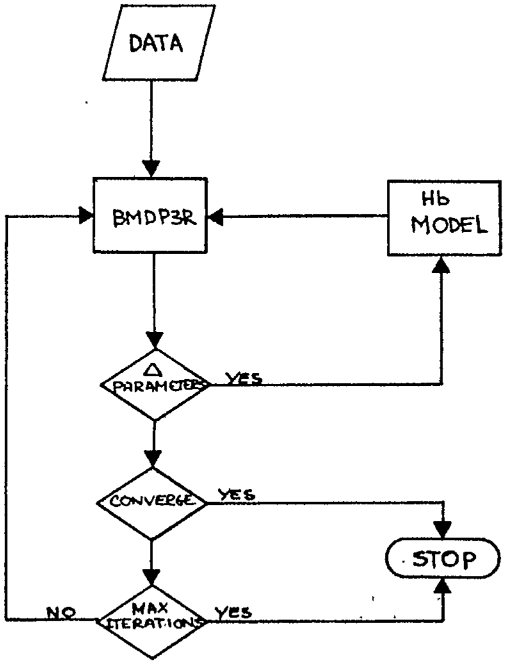
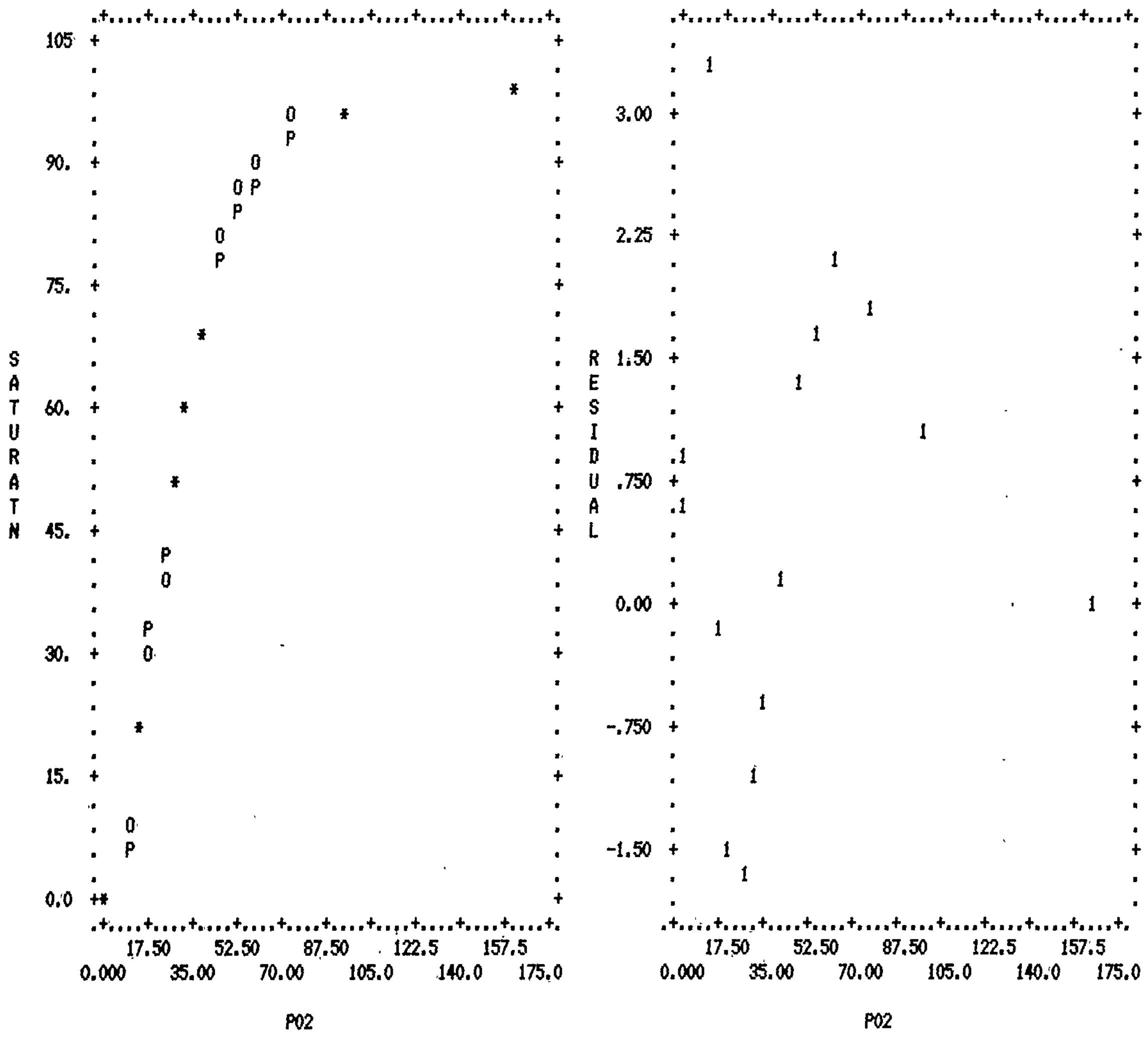


Figure 2. CHEMIST/BMDP3R Interface



```
RESIDUAL MEAN SQUARE            2.32481
DEGREES OF FREEDOM              14
PARAMETER       ESTIMATE        ASYMPTOTIC          TOLERANCE
                            STANDARD DEVIATION

P1              -10.233560      0.045683        1.0000000000
P2              -10.106700      0.021345        1.0000000000
P3              -20.345020      0.095634        1.0000000000
P4              -57.657883      0.039891        1.0000000000
```

PLOTS OF VARIABLE(1) VERSUS PREDICTED AND OBSERVED VARIABLE(2) AND VERSUS RESIDUALS.

procedure is satisfied or an arbitrary maximum number of iterations is obtained. Figure 3 lists the regression output from a typical model run.

5. Conclusions

The advantages of this approach to modeling are the ability to probe selected sections of a large model without disrupting the model's integrity and to quantify the behavior of the model with respect to data obtained from the real system. Probing and goaling permit data inclusion to estimate parameters or comparison of performance against different data sets of models. This procedure essentially includes an optimization procedure into or as an integral part of the model itself. The model need not be permanently altered, however, nor is a great deal of time and programming effort required to incorporate powerful statistical methods into model development and maintenance. This approach is not limited to the use of regression. Any of the statistical software may be used in the same sense although not for the same purposes.

Modelers have, for years, selected subroutines from mathematical libraries to save program development time. Now we have at our disposal extensive, well-documented and well-tested statistical libraries. Their use in model development and validation will be a great advantage as models proliferate and increase in complexity.

6. Literature Cited

Barr, A.J., J.H. Goodnight, J.P. Sall, W.H. Blair, D.M. Chilko. 1979. SAS User's Guide. SAS Institute. Raleigh, NC. pp 494.

DeLand, E.C. 1967. CHEMIST-The Rand Chemical Equilibrium Program. Rand Memorandum RM-5404-PR. ppxi+132.

Dixon, W.J., M.B. Brown. 1979. BMDP-79: Biomedical Computer Programs P-Series. University of California Press, Berkeley. pp xiii+880.

Fell, D. 1979. Computer Simulation Studies of the Mixing Technique and Nonlinear Optimization used in the Analysis of Oxyhemoglobin Dissociation. Math. Biosci., 46:59-69.

Monod, J., J. Wyman and J.P. Changeaux. 1965. On the Nature of Allosteric Transitions: A Plausible Model. J. Mol. Biol., 12:88-118.

Perutz, M.F. 1970. Stereochemistry of Cooperative Effects in Hemoglobin. Nature. 228:726-739.

Roughton, F.J.W., E.C. DeLand, J.C. Kernohan, J.W. Severinghaus. 1971. Some Recent Studies of the Oxyhaemoglobin Dissociation Curve of Human Blood under Physiological Conditions and the Fitting of the Adair Equation to the Standard Curve.

Seaton, B. and B. Lloyd. 1974. A Method for Obtaining Data and Equilibrium Constants for the Hemoglobin-Oxygen Equilibrium in vitro. Resp. Physiol., 20:191-207.

Severinghaus, J.W.. 1966. Blood-Gas Calculator. J. Appl. Physiol., 21:1108-1116.