MoDisc User's Manual

Ranjan Srivastava, Ph.D.

Last modified - June 11, 2007

1 Summary

MoDisc is a Bayesian-based model discrimination application for the identification of the most probable model out of a pool of models given a set of experimental data. MoDisc does not carry out simulations. Rather, it allows analysis of simulation results that have already been generated, in conjunction with experimental data, to evaluate model quality.

As an example of a typical usage of MoDisc, consider a researcher who is studying signal transduction systems in some organism. Assume that researcher postulated three different hypotheses regarding the signaling mechanism. Further assume that the researched had collected some experimental data about the process, but not enough to definitively identify which hypothesis was correct. In such a scenario, the researcher might use MoDisc to help evaluate the most probable hypothesis in the following manner. First the researcher would need to translate his or her hypotheses into some kind of mathematical model, such as using a system of ordinary differential equations to describe the signaling phenomena. Then the researcher would need to carry out parameter estimation for each of the three models based on the collected experimental data, as well as any data from the literature that might be usable. The next to step would be to carry out simulations using each of the three models and collecting the resulting simulation data. At this point the simulation data and the actual experimental data may be fed into MoDisc, and the most probable model/hypothesis will be determined.

It is important to note, however, that MoDisc may be used for far more than just the cell signaling example provided here. Types of analyses may range from identification of the most probable kinetic model, whether deterministic, stochastic, or a combination of both, to the determination of the most probable objective function for metabolic flux analysis. MoDisc is capable of running on the Windows, Linux, and the OS X platforms.

Software, documentation, and updates are freely available at http://www.engr.uconn.edu/~srivasta/modisc.html.

2 Overview

2.1 Introduction

Model development is a useful tool for understanding a wide range of phenomena in the sciences and engineering. Within the area of biological sciences, this approach is becoming more important as scientists strive to keep up with the rapid influx of data being generated. In attempting to make sense of biological phenomena, several models of a system may be postulated. The question then becomes how to discriminate among the models to determine which is most likely.

A method for identifying the most probable model of a chemical reaction network based on experimental data was developed by Stewart and colleagues [4, 5]. This approach, termed "model discrimination," is a Bayesian-based method in which the probability of a model, given a set of experimental data, may be calculated and compared to other potential models. The model with the highest value is considered the most probable model to describe the system.

Model discrimination has wide applicability for use with biological systems. Types of systems that may be analyzed include, but are not limited to, models of transcriptional regulatory networks, intracellular kinetic models, or models of viral dynamics [1, 3]. It should be noted that whether these models are deterministic, stochastic, or a mixture of both paradigms, model discrimination analysis may still be carried out without any hinderance. Model discrimination has also found utilization beyond the study of kinetic models. Recently this methodology has been used to identify the most probable objective function for use in metabolic flux analysis [2].

2.2 Model Discrimination Theory

Stewart's Method of model discrimination is based upon Bayesian analysis. A full derivation of the technique is provided in [4] and [5]. However a brief description of the principal points are provided here.

Stewart's Method is based on calculating the posterior probabilities of competing models relative to each other. The model with highest probability is considered the most likely choice relative to the other models. Calculation of posterior probabilities are based on the following proportionality,

$$p(M_j \mid \boldsymbol{Y}) \propto p(M_j) 2^{-\frac{p_j}{2}} \mid \hat{\mathbf{v}}_{\boldsymbol{j}} \mid^{-\frac{\nu_e}{2}} \tag{1}$$

where M_j is the j^{th} model, $\boldsymbol{Y}$ is the matrix of experimental results, and $p(M_j \mid \boldsymbol{Y})$ is the posterior probability of model M_j given $\boldsymbol{Y}$. Additionally, p_j is the number of independent parameters, ν_e is the number of degrees of freedom, and $\mid \hat{\mathbf{v}}_{\boldsymbol{j}} \mid$ is the determinant function. The elements of the determinant function are given by

$$v_{ik}(\boldsymbol{\theta_j}) = \sum_{u=1}^{n}[Y_{iu} - \mathcal{F}_{ji}(\xi_u, \boldsymbol{\theta_j})][Y_{ku} - \mathcal{F}_{jk}(\xi_u, \boldsymbol{\theta_j})] \qquad i, j = 1, \ldots, q \tag{2}$$

in which n is the the total number of events evaluated, q is the number of different chemical species monitored, and $\mathcal{F}_{ji}(\xi_u,\boldsymbol{\theta_j})$ is the model prediction. $\boldsymbol{\xi}$ is the vector of the number of different independent conditions (i.e. temperature) tested, and $\boldsymbol{\theta_j}$ is the vector of parameters providing the best fit of the model to the experimental data.

Normalizing the results of Equation 1 for any given model to the sum of the results for all the models is referred to as the *probability share* and is shown in Equation 3,

$$\pi_j(M_j \mid \boldsymbol{Y}) = \frac{p(M_j \mid \boldsymbol{Y})}{\sum\limits_k p(M_k \mid \boldsymbol{Y})} \tag{3}$$

The model with the highest probability share is considered most probable.

3 MoDisc Usage

3.1 Software Installation

To carry out installation of MoDisc, it is first necessary to download and install the LispWorks Common Lisp Personal Edition software. The software may be freely downloaded from http://www.lispworks.com/downloads/index.html, along with documentation of how to install the software. LispWorks is available for Windows, Linux, and OS X.

Once the LispWorks Personal Edition is installed, the MoDisc code may be downloaded. MoDisc is available at http://www.engr.uconn.edu/~srivasta/modisc.html. The link for the software is found on the left-hand side menu bar. You may also directly download it from http://www.engr.uconn.edu/~srivasta/modisc.html/modisc.zip.

Why Lisp? For those with experience with various mathematical software packages, such as *Mathematica* or MATLAB, one might ask “Why not implement MoDisc in one of those languages?” The reason is simple. We wanted to make this tool available to as large a group as possible. Other packages, such as those already mentioned, require a license. The code we provide may be used after downloading a free copy of the Lispworks Personal Edition. A further benefit of the Lispworks platform is that it runs on the three major operating systems, Windows, Linux, and OS X.

The reason for using Common Lisp over C, Fortran, or Java was primarily a matter of preference. Overall, we felt this was the best way to ensure that the most possible people who wanted to use the software actually could use it.

3.2 Input File

To use MoDisc, experimental data and model information may be entered via an input file. The file may be in the form of a spreadsheet, such as an Excel file, such as shown in Figure 1, or as a tab delimited text file. The input file is organized into a series of blocks of information for MoDisc.

The first block is for the comments sections. Comments may be entered within a "begin-comment" and "end-comment" section.

The next set of of blocks are for the details of the model, as well as the results of the simulation. Model simulation results may be entered by starting a section called "begin-model" followed by the name of the model. The number of parameters, the number of variables, and the number of degree of freedoms are entered next. The following row consists of a list of the variables used in the simulation. Simulation results for the specific model are then entered. Finally the block is closed by ending it with an "end-model" row. This procedure is repeated for each of the remaining models.

To enter the experimental data, a new block is started by entering "begin-experiment" in a new row. The number of variables are then entered, followed by a row consisting of the independent and dependent variables measured. Finally the experimental data is entered, followed by an "end-experiment" row. The file may then be saved as a tab delimited text file.

3.3 Running MoDisc

To run MoDisc, first launch LispWorks. Under the LispWorks menu bar, choose "File → Open" and select the "modisc.lsp" file. This will result in an editor window being launched containing the modisc source code. Select the editor window using your mouse. Then go to the menu bar and select "Buffers → Compile." Finally, at the "CL-USER 1 >" prompt in the original LispWorks window, type "(modisc)" (make sure to include the parenthesis). At this point, you will be prompted for your input file. After reading the input file in, MoDisc will return the probability share of each of the models.

3.4 MoDisc In Action - A Reaction Kinetics Example

A reaction kinetics example adapted from [5] using two different models is provided here. In this example, it is known that the system consists of three chemical species, $a1$, $a2$, and $a3$. However, the exact reaction mechanism is not known, and two hypotheses are put forward. In the first proposed reaction scheme, Model-1, it hypothesized that the chemical species $a1$, $a2$, and $a3$ follow a series of sequential irreversible reactions, as represented by

$$a1 \overset{k_1}{\rightarrow} a2 \overset{k_2}{\rightarrow} a3 \tag{4}$$

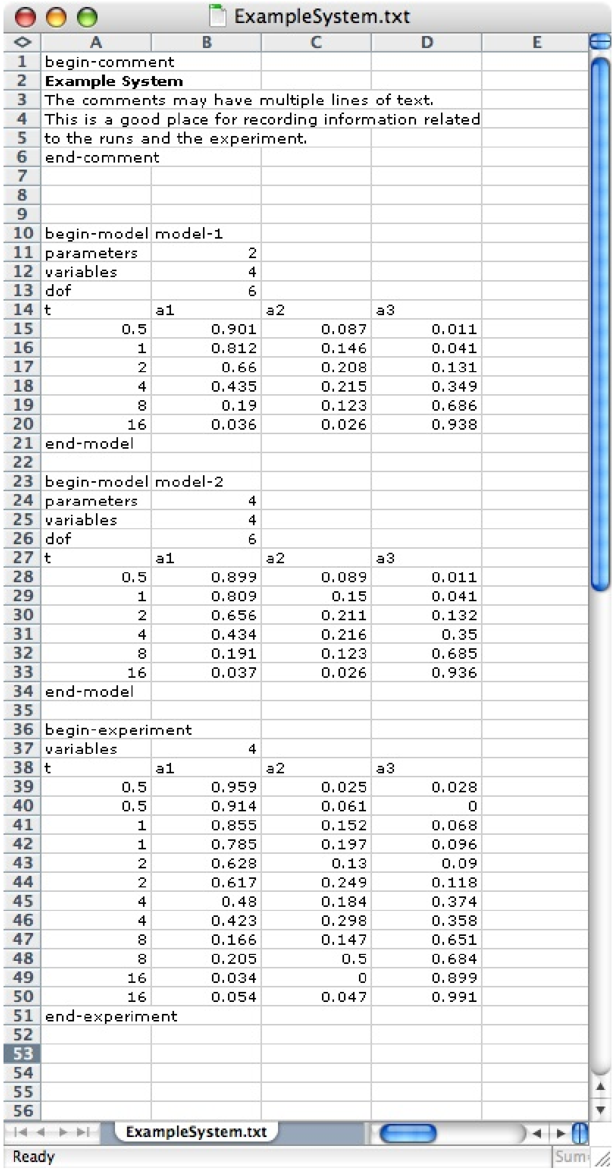
The second proposed model, Model-2, is similar to Model-1. However, in this case, all the reactions are considered to be reversible,

$$a1 \underset{k_3}{\overset{k_1}{\rightleftarrows}} a2 \underset{k_4}{\overset{k_2}{\rightleftarrows}} a3 \tag{5}$$

In this example for the sake of illustration, the “experimental data” was generated from Model-1 with noise, representing experimental error, added.

Both models were simulated via ordinary differential equations, where parameters were fitted to the “experimental” data. Time, represented by t, was the independent variable. Simulation and experimental results were entered into an Excel file, shown in Figure 1. Such a file may be used as an input file for the MoDisc program. Note that replicate experimental results were also included. MoDisc was then used to calculate which of these two models was most probable. In this case, the first model was selected as most probable with a probability share of 0.65. The probability share of the second model was 0.35. This result should not be surprising, given that the “experimental” data was generated artificially from Model-1 to begin with.

It should additionally be pointed out that although only two models were compared in this particular case, any number of models may be compared in actuality.



	A	B	C	D	E
1	begin-comment				
2	**Example System**				
3	The comments may have multiple lines of text.				
4	This is a good place for recording information related				
5	to the runs and the experiment.				
6	end-comment				
7					
8					
9					
10	begin-model	model-1			
11	parameters	2			
12	variables	4			
13	dof	6			
14	t	a1	a2	a3	
15	0.5	0.901	0.087	0.011	
16	1	0.812	0.146	0.041	
17	2	0.66	0.208	0.131	
18	4	0.435	0.215	0.349	
19	8	0.19	0.123	0.686	
20	16	0.036	0.026	0.938	
21	end-model				
22					
23	begin-model	model-2			
24	parameters	4			
25	variables	4			
26	dof	6			
27	t	a1	a2	a3	
28	0.5	0.899	0.089	0.011	
29	1	0.809	0.15	0.041	
30	2	0.656	0.211	0.132	
31	4	0.434	0.216	0.35	
32	8	0.191	0.123	0.685	
33	16	0.037	0.026	0.936	
34	end-model				
35					
36	begin-experiment				
37	variables	4			
38	t	a1	a2	a3	
39	0.5	0.959	0.025	0.028	
40	0.5	0.914	0.061	0	
41	1	0.855	0.152	0.068	
42	1	0.785	0.197	0.096	
43	2	0.628	0.13	0.09	
44	2	0.617	0.249	0.118	
45	4	0.48	0.184	0.374	
46	4	0.423	0.298	0.358	
47	8	0.166	0.147	0.651	
48	8	0.205	0.5	0.684	
49	16	0.034	0	0.899	
50	16	0.054	0.047	0.991	
51	end-experiment				
52					
53					
54					
55					
56					

Figure 1: Model results and experimental data for MoDisc may be entered via a spread sheet. A reaction kinetics example using two different models is shown here and described in detail in Section 3.4 . Note that more than two models at a time may be compared.

References

[1] R. Jain, A.L. Knorr, J. Bernacki, and R. Srivastava. Investigation of Bacteriophage MS2 Viral Dynamics Using Model Discrimination Analysis and the Implications for Phage Therapy. *Biotechnol Prog*, 22(6):1650–8, 2006.

[2] A.L. Knorr, R. Jain, and R. Srivastava. Bayesian-based selection of metabolic objective functions. *Bioinformatics*, 23(3):351 – 357, 2007.

[3] A.L. Knorr and R. Srivastava. Evaluation of HIV-1 kinetic models using quantitative discrimination analysis. *Bioinformatics*, 21(8):1668–77, 2005.

[4] W.E. Stewart, T.L. Henson, and G.E.P. Box. Model Discrimination and Criticism with Single-Response Data. *AIChE Journal*, 42(11):3055–3062, 1996.

[5] W.E. Stewart, Y. Shon, and G.E.P. Box. Discrimination and goodness of fit of multiresponse mechanistic models. *AIChE Journal*, 44(6):1404–1412, 1998.