J Math Chem (2015) 53:151–161
DOI 10.1007/s10910-014-0416-z

ORIGINAL PAPER

Finding protein thermostability and spin-coupling constant using Bayesian statistics

Wenjin Zhou · Allison M. Rossetto

Received: 7 August 2014 / Accepted: 19 September 2014 / Published online: 17 October 2014

Abstract This paper outlines the use of Bayesian statistics to find the thermostability and spin-coupling constant of protein. Thermostability is an important factor in protein efficacy; modeling it lets us find the mutation temperature of a protein. This is important since the temperature affects protein function. The spin-coupling constant provides high-level structure information about bond angles and rotation in a protein. We have used Bayesian statistics (MCMC) to find the unknown parameters for these two models. Predictive models using the parameters found with this method show good results.

Keywords Bayesian statistics · MCMC · Thermostability · Spin-coupling · Proteins

1 Introduction

Bayesian statistics is a useful method for finding multiple unknown parameters in a mathematical equation. The method lets us find unknown parameters given a mathematical model with some experimental or starting data [1,5]. Here we discuss the application of Bayesian statistics to find the unknown parameters in the models of protein thermostability and the spin-coupling constant of proteins.

Thermostability is an important factor in the efficacy of proteins. Knowing the mutation temperature of a protein is useful since thermostability affects function. Mutants, wild types, or homologous proteins may also have different mutation temperatures that

W. Zhou (✉) · A. M. Rossetto
Department of Computer Science and Engineering, Oakland University,
Rochester, MI 48309-4401, USA
e-mail: wzhou@oakland.edu

A. M. Rossetto
e-mail: amrosset@oakland.edu

cause them to function at different temperatures. A group at Fudan University studied the thermal stability of proteins using neutron-scattering-spectra-associated index parameters β and heat capacity C_p at constant pressure as the two criteria [4]. This method, however, requires complex integration of the parameter. Bayesian statistics method presented here provides an easier method to determine the thermostability.

The spin-coupling constant J provides information about bond angles and rotation, which gives insight into the higher-level structure levels of the protein studied. The J constant can be found experimentally using NMR, but the NMR data provides only part of the information needed to construct a mathematical model.

Here we present two applications and a brief introduction to Bayesian statistics. First, we give a brief survey of Bayesian statistical methods. Next, we outline the problem of determining the thermostable cutoff temperature for the wild-type thermostable catechol 2,3-dioxygenase (TC23O) and homologous catechol 2,3-dioxygenase (1MPY) using the Poisson distribution. We refer to catechol 2,3-dioxygenase by its pbd file name, 1MPY, rather than C23O to avoid confusion. In our results, we have identified the mutation-point temperature of TC23O at 310 K and the 1MPY mutation-point temperature at 240 K. This shows that the thermal stability of TC23O is better than that of 1MPY. We also obtain exact values for the six unknown parameters of the protein ubiquitin (Ub) for mathematical modeling of the spin-coupling constant J using Gaussian distribution. The final three sections discuss our results, discussion, and conclusions.

2 Methods

Bayesian statistics consists of the Bayesian formula and the Markov Chain Monte Carlo (MCMC) method to find unknown parameters in a problem using experimental or known data, and a mathematical model of the problem.

The Bayesian statistical view takes a known set of experimental data and the unknown parameters as random numbers, but each follows its own distribution. Suppose D is a vector expressing all experimental data $D_1, D_2, \ldots, D_n$, and the vector θ that expresses all unknown parameters. The relation of their probabilities is the Bayesian formula:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta} \tag{1}$$

The denominator of Eq. 1, however, is a constant when θ is independent. So, the formula can also be written as:

$$P(\theta|D) \propto P(\theta)P(D|\theta) \tag{2}$$

We call $P(\theta|D)$ the posterior probability, $P(\theta)$ the prior probability, and $P(D|\theta)$ the likelihood. Thus the mathematical expectation of a function $F(\theta)$ can be defined as:

$$E(\theta|D) = \int F(\theta)P(\theta|D)d\theta \tag{3}$$

In addition to the Bayesian statistics outlined above, the Markov Chain and Monte Carlo methods are necessary to obtain the unknown variables. This subsection gives a brief explanation of the Monte Carlo Method, the Markov Chain, and the combined Markov Chain Monte Carlo (MCMC) method.

The principle of the Monte Carlo method is relatively simple. According to the law of large numbers, the sample mean is close to the overall average when n is large. Thus we can use many random numbers with a known distribution to calculate numerical values for mathematical functions. For example, we can calculate the definite integral $\int_a^b f(x)dx$ using n uniformly distributed random numbers on the interval $[a, b]$ as $\int_a^b f(x)dx \to \frac{b-a}{n} \sum_{i=1}^{n} f(x_i)$ as $n \to \infty$. This method can be applied to many different kinds of problems, including Bayesian statistics.

A Markov Chain is a random process. It can also be thought of as a set of random variables. The next value or state at any point in the chain depends only on the current value or state; it is not affected by previous values. A very simple way of thinking about this is to consider a random process of putting beads on a string. You have many different containers with many colors of beads in them. Any bead you take out of a container is replaced with a new bead. If the last bead on the string is blue, you take the next bead out of a certain container. If that bead is green, your next bead may be from the same or a different container. The fact that the previous bead was blue has no effect on the bead you choose after the current one.

A mathematical way of stating this is that, given a chain of statuses $X_0, X_1, \ldots, X_{(t+1)}$, the probability of the next state is $P(X_{(t+1)}|X_t)$. This probability depends only on the state X_t and not on the value of t. More generally, n-step Markov Chains may be defined as $P(X_{(t+n)}|X_t)$ for some $n \in \mathbb{Z}^+$.

The Markov Chain depends on the initial state, denoted by π, and the transition matrix P_{ij} is defined as:

$$P_{ij} = P(X_{(t+n)} = a_j | X_t = a_i), \text{ where } n = [1, 2, 3, \ldots] \tag{4}$$

The state is generally represented by multiple variables, so the state values a_i, a_j can be vectors. The transition matrix P_{ij} has some important properties:

1. $P_{ij} \geq 0$
2. The sum of the elements of each row is 1. That is, $\sum_{j=1}^{n} P_{ij} = 1$.
3. P_{ij} is defined only for states separated by n discrete steps.

The Markov Chain steady state is independent of time and initial state π. A chain usually has several equilibrium distributions, but the initial state is not sensitive to the given conditions. Thus, the Markov Chain eventually forgets its initial state and converges to a unique stationary distribution of the steady state. Therefore, it will drop the m unstable states as expressed by the equation:

$$E(f(x)) = \frac{1}{n-m} \sum_{i=m+1}^{n} f(x_i) \tag{5}$$

Typical m values are 5–10 % of the total number of states.

The MCMC method simply uses the random numbers produced by a Markov Chain for the Monte Carlo method, as expressed by the equation:

$$E(\theta|D) = \int F(\theta)p(\theta|D)d\theta \rightarrow \frac{1}{n-m}\sum_{i=m+1}^{n} F(\theta_i) \tag{6}$$

Two main sampling methods are used to generate the Markov Chain: Metropolis–Hastings sampling and Gibbs sampling. We used the Gibbs sampling method in our application.

2.1 Finding the thermostability criterion

We applied the concept of Bayesian statistics to the problem of determining the thermostability criterion of TC23O and its homologue 1MPY. The neutron-scattering-spectra-associated index β and the heat capacity C_p at constant pressure are two experimentally determinable criterias of thermostability. From the experimental data and molecular dynamics simulations, we can also obtain the mean-square displacement (MSD) data for the hydrogen atoms of TC23O and 1MPY. Figure 1 shows the MSD curve for TC23O and 1MPY. The MSD is calculated using the total number of hydrogen atoms N as:

$$MSD = \frac{1}{N}\sum_i [r_i(t) - r_i(0)]^2 \tag{7}$$

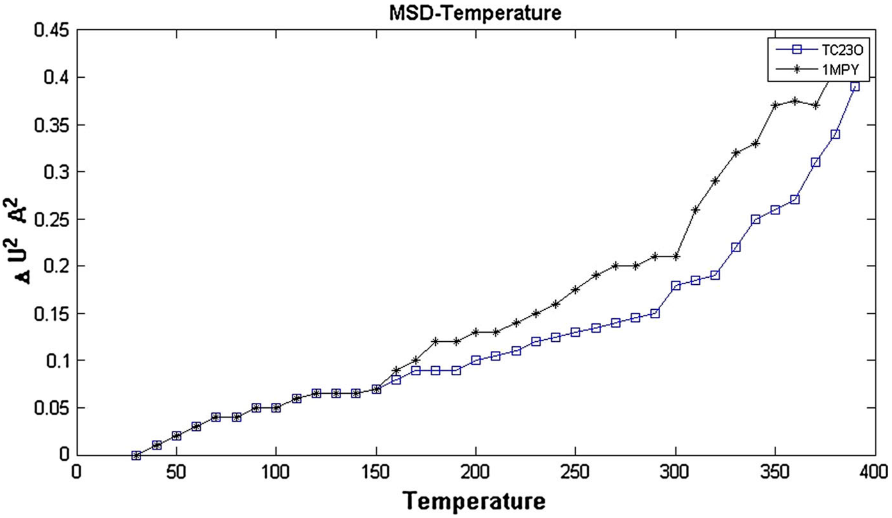

Fig. 1 The mean-square displacement (MSD) curve with temperature of all hydrogen atoms in TC23O and 1MPY. The kinetic transition temperature T_d appears at 310 and 245 K; (*line with open square*) TC23O hydrogen atom; (*line with filled diamond*) 1MPY hydrogen atom

Table 1 Temperature, MSD of TC23O hydrogen atoms, and MSD of 1MPY hydrogen atoms

T	MSD T230	MSD 1MPY	T	MSD T230	MSD 1MPY
30	0.000	0.000	220	0.110	0.140
40	0.010	0.010	230	0.120	0.150
50	0.020	0.020	240	0.125	0.160
60	0.030	0.030	250	0.130	0.175
70	0.040	0.040	260	0.135	0.190
80	0.040	0.040	270	0.140	0.200
90	0.050	0.050	280	0.145	0.200
100	0.050	0.050	290	0.150	0.210
110	0.060	0.060	300	0.180	0.210
120	0.065	0.065	310	0.185	0.260
130	0.065	0.065	320	0.190	0.290
140	0.065	0.065	330	0.220	0.320
150	0.070	0.070	340	0.250	0.330
160	0.080	0.090	350	0.260	0.370
170	0.090	0.100	360	0.270	0.375
180	0.090	0.120	370	0.310	0.370
190	0.090	0.120	380	0.340	0.410
200	0.100	0.130	390	0.390	0.420
210	0.105	0.130	NA	NA	NA

We also used the MSD and temperature data to set up a Bayesian statistical analysis to determine thermostability (Table 1).

The Poisson distribution is often used for a number of incidence statistics. As we see from Fig. 1, the mean-square displacement increases significantly at a certain temperature. Our aim is to use the MCMC method to estimate and predict this mutation point. The key point is to assume that the system obeys one Poisson distribution before the mutant temperature and another Poisson distribution after the mutant temperature. The Poisson distribution is expressed as:

$$P(x) = e^{-\lambda}\frac{\lambda}{x}, \quad \text{where } \lambda > 0, \quad x = 0, 1, 2, \ldots \tag{8}$$

Thus, the Poisson distribution should have a different λ parameter after the mutation temperature. The model can be defined in two parts as:

$$\lambda(i) = e^{(b_1 + step(i-k)*b_2)} \tag{9}$$

$$step(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{10}$$

Bayesian Inference Using Gibbs (OpenBUGS) is an MCMC software that runs on Windows, Linux and MacOS. The OpenBUGS software uses the Gibbs sampling algorithm with Bayesian statistics [3]. It is coded in R scripting language, from the project founded by Ross Ihaka and Robert Gentleman [2]. The code has three parts: the model section, the data section, and the initialization section. We first provide the likelihood and then the prior values for our model with its distribution in the model section. The experimental data appears as a list in the data part. In the initialization section, we provide initial values for the unknown parameters.

The program OpenBUGS runs MCMC to fit parameters b_1, b_2, and k in the model. This particular application is fairly simple. We used a Poisson distribution, the model defined in Eq. 9 and Eq. 10, and the molecular dynamics data, and we initialized the model at $b_1 = 0$, $b_2 = 0$, and $k = 20$. Since the Poisson distribution argument must be a positive integer, we expand 100 times for MSD. This does not affect its nature. For the 1MPY example, the OpenBUGS procedure is as follows in Program 1.

Program 1 Bugs Code for finding *lamda*

```
    model {

        for (i in 1:n){
            y[i] ~dpois(mu[i])
            log(mu[i]) < -b[1] + step(i-k)*b[2]
        }

        for(j in 1:2){
            b[j] ~ dnorm(0.0, 1.0E-6)
        }

        k~dunif(1,n)
    }

list(n=37, y =c(0, 1, 2, 3, 4, 4, 5, 5, 6, 6.5,
6.5, 6.5, 7, 9, 10, 12, 12, 13, 13, 14, 15,
16, 17.5, 19, 20, 20, 21, 21, 26, 29, 32, 33,
37, 37.5, 37, 41, 42) )

list(b=c(0,0),k=20)
```

2.2 Spin-coupling constant of a protein residue

A similar approach to that outlined in the previous section can be used to determine the equation for the spin-coupling constant J. The spin-coupling constant J can be determined from NMR experiments. X-ray diffraction can provide the rotation angle ψ around a single C_α–C bond and the rotation angle ϕ around a single C_α–N bond. The mathematical model defining the relationship between J, ψ, and ϕ can be written as:

$$J_i = A + B\sin(\psi_i + \psi_0) - C\cos(2(\psi_i + \psi_0)) + D\cos(2(\phi_i + \phi_0)) \qquad (11)$$

where J_i, ψ_i, and ϕ_i are the ith experimental values. The remaining constants A, B, C, and D and the initial angles ψ_0 and ϕ_0 that must be found using Bayesian statistics.

Again, we used OpenBUGS to run MCMC. The MCMC process to fit the six parameters in the J equation (11) is a little more complex than that for the thermostability application. We used uniform distribution for the unknown parameters and a normal distribution for *J*. We also need two more variables, τ and σ, that are used to define the distribution for *J*. It must also be noted that we must run MCMC for at least a million iterations to produce accurate results. The code to run the MCMC in order to find all the unknown parameters is shown below in Program 2.

Program 2 Bugs program for finding Spin-Coupling equation constants.

```
model
    {
        for( i in 1 : N ) {
            Y[i] ~ dnorm(mu[i], tau)

            mu[i] <- A+B*sin(3.14159/180*(psi[i]+psi0))
            -C*cos(2*3.14159/180*(psi[i]+psi0))
            +D*cos(2*3.14159/180*(phi[i]+phi0))
        }
        tau ~ dgamma(0.001, 0.001)
        psi0 ~dunif(100, 200)
        phi0 ~dunif(10,50)
        A  ~dunif(100, 200)
        B ~dunif(0, 5)
        C ~dunif(0,5)
        D ~dunif(0,5)
}

list(Y=c(143.8650, 138.2660, 142.8440, 138.7770, 143.5620, 138.5590,
141.5180, 141.6620, 143.3900, 142.0190, 141.0070, 142.9980, 139.6710,
142.9670, 141.8230, 141.2770, 142.8900, 141.6830, 141.7750, 147.1500,
145.3120, 147.3090, 145.8280, 147.0240, 142.7480, 141.7890, 145.0790,
139.5500, 142.7460, 146.1250, 141.3430, 139.6180, 139.3690, 140.6760,
140.4850, 141.7890, 135.8790, 141.0070, 140.6840, 139.4270, 138.2540,
144.1360, 146.6530, 139.8290, 135.2640, 139.0350, 137.9060, 144.3100,
131.6650, 144.1040, 142.3370, 140.7250, 141.1930, 141.5730, 139.0840,
142.3110, 142.9790, 141.2520, 141.8340 ),

phi = c(-93.0660, -129.9030, -111.9240, -115.5030, -89.5330, -95.0710,
-72.9520, -95.3990, -102.5400, -112.7690, -107.8610, -96.4490, -128.5930,
-108.1840, -136.9390, -116.5520, -74.2110, -82.1290, -63.9300, -56.9940,
-62.4970, -61.9890, -68.6350, -63.8930, -63.9800, -58.8500, -94.1960,
-119.0680, -79.3690, -64.4160, -92.3300, -84.8100, -128.3790, -96.1020,
-122.5630, -143.5870, 50.8410, -112.7020, -75.3370,-82.5850, -103.3780,
-62.2770, -66.1410, -96.7640, 63.4000, -77.2660, -99.2820, -53.6520,
72.7130, -74.1720, -116.8690, -97.8620, -104.8100, -105.2350, -113.0780,
-92.2360, -122.6380, -81.2270, -99.0580),

psi=c(132.5650, 159.7930, 134.7380, 106.6430, 121.4910, 170.6230,
-11.4640, 4.0580, 131.5780, 128.1320, 136.5680, 138.2780, 149.1180,
114.3720, 167.8460, 142.2070, 143.6580, 161.5920, -37.8350, -40.2400,
-43.2520, -34.1490, -37.5370, -42.3070, -40.0600, -38.7980, -28.1910,
-11.0430, 119.2440, -18.4380, -10.5110, 130.1390,  107.6240, 131.0830,
130.9270, 128.4580,  44.0300, 140.0160, 133.8270, 164.1810, 166.0220,
-32.9340, -31.4130,  -0.2260, 35.7980, 112.5620, 169.7320, 139.5180,
17.9760, 157.1250, 122.3190, 150.7280, 129.1740, 116.8930, 133.5510,
138.0120, 96.9040, 147.2140,  93.8800),

 N=59
)
list( tau=0.05, A=150.0, B=3.0, C=5.0, D=5.0, phi0=50.0, psi0=150)
```

3 Results and discussion

3.1 Thermostability

The results of the MCMC are shown in Table 2. The data shows very small error and standard deviation for the parameters. Figure 2 shows the density functions of the parameters.

Figure 3 shows the change in λ with respect to temperature of the proteins, this change can be called the *knee point*. The change in the Poisson distribution, which indicates the mutation temperature, can easily be seen at 240 K for 1MPY and 310 K for TC23O.

The change of the parameters is consistent with Fig. 1, showing that the thermal stability of TC23O is better than that of 1MPY. The Poisson distributions have two different parameters before and after the mutation points: TC23O has λ values of 7.524 and 24.260 while 1MPY has λ values of 7.152 and 27.750. There is a transition period between the two Poisson distribution parameters of a protein. It is usually

Table 2 MCMC node statistics of thermostability equation parameters

Parameter	Mean	SD	MC error	Median	Starting sample	Total samples
b[1]	1.962	0.1136	9.850E−4	1.969	10,001	90,000
b[2]	1.360	0.1035	6.660E−4	1.357	10,001	90,000
k	21.130	1.4220	1.211E−2	21.190	10,001	90,000

Table shows MCMC estimation results for thermostability equation parameters. The mean values can be used as estimates for each parameter in the equation. The first ten thousand samples are discarded so that only the values of the stable system are included in the mean values

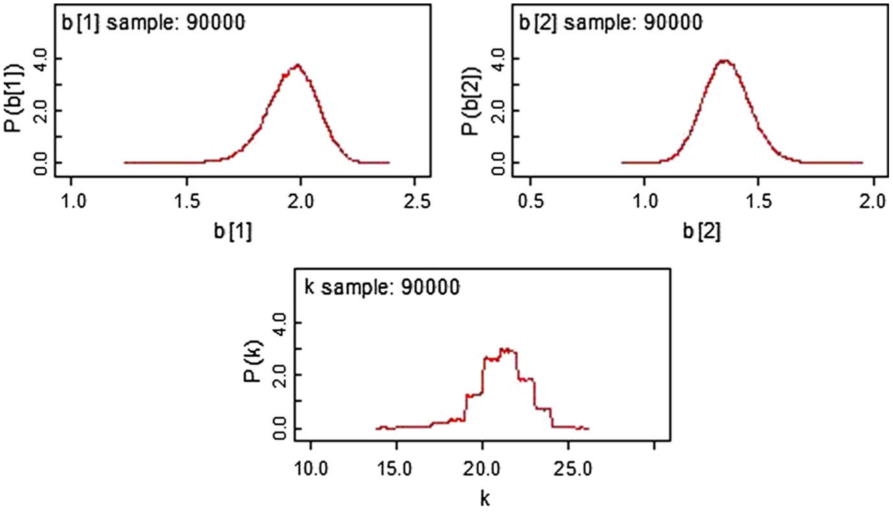

Fig. 2 Density functions for *b1*, *b2*, and *k* for 90,000 samples

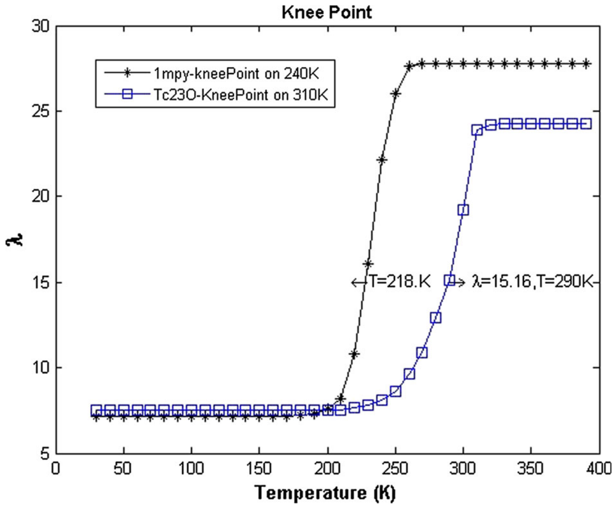

Fig. 3 Knee point change in λ versus temperature (K) for TC23O and 1MPY

Table 3 MCMC node statistics of the spin-coupling equation parameters. The first one hundred thousand samples are discarded so only the values of the stable system are included in the mean values

Parameter	Mean	SD	MC error	Median	Starting sample	Total samples
A	140.500	0.7491	1.858E−3	140.400	100,001	900,000
B	1.218	0.4312	9.780E−4	1.215	100,001	900,000
C	3.604	0.6043	1.464E−3	3.617	100,001	900,000
D	1.407	0.6808	1.676E−3	1.354	100,001	900,000
phi0	30.770	10.0800	2.473E−2	30.450	100,001	900,000
psi0	142.500	5.1230	1.261E−2	142.100	100,001	900,000

believed that the protein remains in a harmonic potential well and the system has only a single conformational dynamic when the temperature is below the mutation point. Then, as the temperature increases, the heat allows smaller molecules to vibrate. From the mutation temperature (i.e. the kinetic phase transition temperature), it can be seen that the transition period can be reached by 1MPY at 218 K and by TC23O at 290 K, if required to achieve the same λ value as in Fig. 3 ($\lambda = 15.16$). Thus, for the same number of conformational dynamics, TC23O needs more heat. This indirectly reveals that TC23O has higher thermal stability than its homologous 1MPY. This Bayesian statistical method for determining of protein thermal stability is simpler than the neutron-scattering-spectra-associated index β parameters, because the parameter β must be taken through a complex integration.

3.2 The spin-coupling constant

The results of the MCMC for the J constant equation are shown in Table 3.

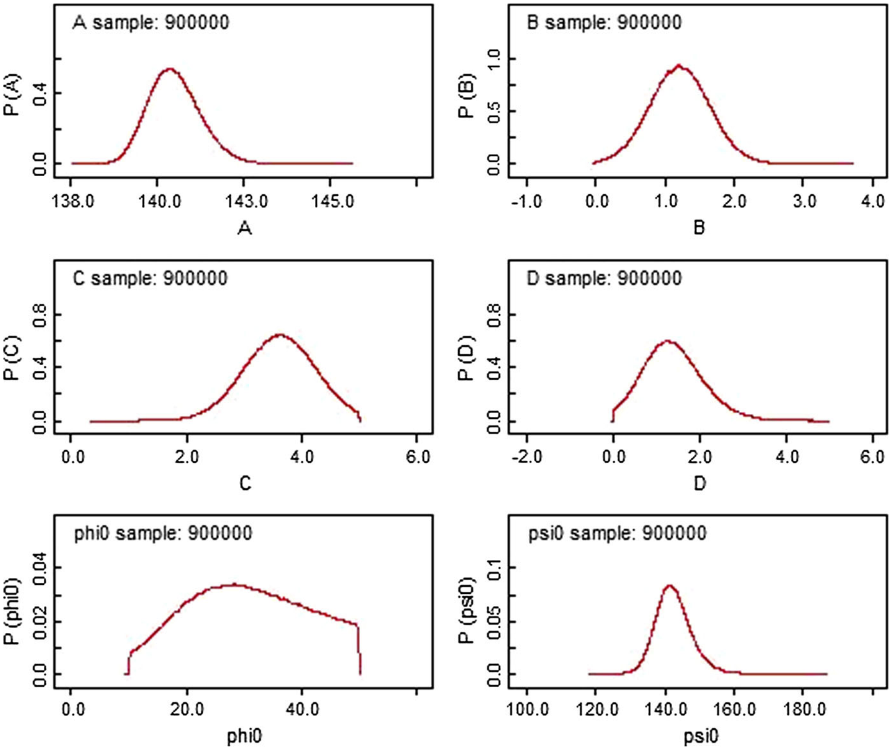

Fig. 4 Density distributions of the parameters A, B, C, D, ψ_0, and ϕ_0 from the J constant equation

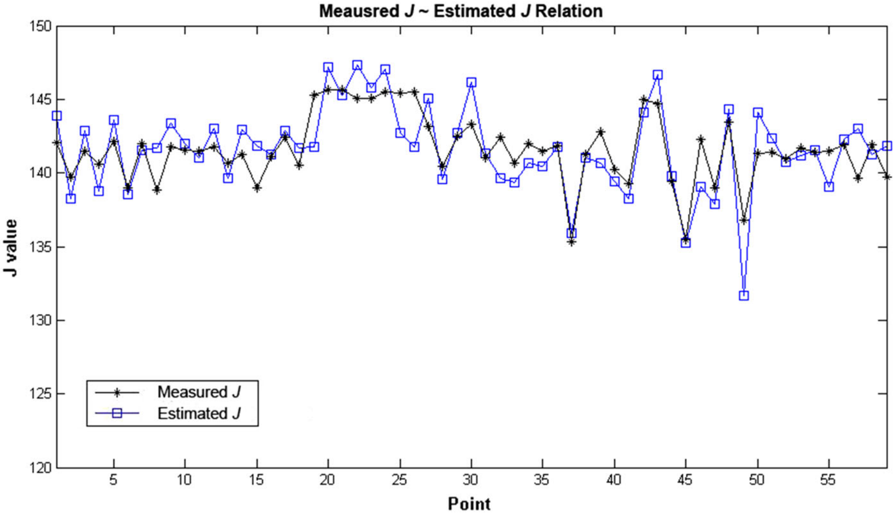

Fig. 5 Measured spin-coupling constant J (*blue open square*) and estimated J (*black filled diamond*) values generated using MATLAB

We can use these to give our full mathematical model:

$$J_i = 140.5 + 1.218\sin(\psi_i + 142.5) - 3.604\cos(2(\psi_i + 142.5)) \\ +1.407\cos(2(\phi_i + 30.77)) \quad (12)$$

Figure 4 shows the density distribution of the parameters. Most parameter distributions showed the expected normal distributions. Only the distribution of ϕ_0 is not a normal distribution, it is more flattened than a normal distribution. This accounts for some of the offset predicted in the measured results. Comparison of the estimated spin-coupling J values from our model with the measured J values in Fig. 5 shows our results are reasonably good predictions.

4 Conclusion

Problems often arise for which a set of experimental data and a mathematical model are available, but the mathematical model has several unknown variables. Bayesian statistics is a good method for finding unknown parameters in biological models, as shown by the protein applications discussed here. The mathematical models used to predict thermostability and the spin-coupling constant can be easily applied to other proteins. This would be particularly helpful for drug targeting, since a deep understanding of target proteins is required.

Acknowledgments The authors would like to thank Professor Linxiang Zhou for his help with this work.

References

1. M.H. Chen, Q.M. Shao, J.G. Ibrahim, *Monte Carlo methods in Bayesian computation* (Springer, New York, NY, 2000)
2. R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (2014)
3. A. Thomas, B.O. Hara, U. Ligges, S. Sturtz, Making BUGS Open. R News **6**, 12–17 (2006)
4. J.H. Zhang, L.L. Zhang, L.X. Zhou, Thermostability of protein studied by molecular dynamics simulation. J. Biomol. Struct. Dyn. **21**(21), 657–662 (2004)
5. L.X. Zhou, *Bayesian Statistics and Its Application* (Fudan Press, China, 2010)