A stochastic model for correlated protein motion

Article *in* Physics Letters A · January 2006

CITATION
1

READS
33

3 authors, including:

Wael Karain
Birzeit University
13 PUBLICATIONS **70** CITATIONS
SEE PROFILE

Basem Ajarmah
Al Istiqlal University (Palestinian Academy fo…
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com
SCIENCE DIRECT
Physics Letters A 354 (2006) 497–500
PHYSICS LETTERS A
www.elsevier.com/locate/pla

A stochastic model for correlated protein motions

Wael I. Karain [a,*], Nael I. Qaraeen [b], Basem Ajarmah [b]

[a] *Birzeit University, Department of Physics, PO Box 14, Birzeit, Ramallah, Palestinian Territory*
[b] *Birzeit University, Computer Science Department, PO Box 14, Birzeit, Ramallah, Palestinian Territory*

Received 19 November 2005; received in revised form 19 January 2006; accepted 31 January 2006

Available online 9 February 2006

Communicated by J. Flouquet

Abstract

A one-dimensional Langevin-type stochastic difference equation is used to find the deterministic and Gaussian contributions of time series representing the projections of a Bovine Pancreatic Trypsin Inhibitor (BPTI) protein molecular dynamics simulation along different eigenvector directions determined using principal component analysis. The deterministic part shows a distinct nonlinear behavior only for eigenvectors contributing significantly to the collective protein motion.

Keywords: Times series analysis; Principal component analysis; Protein collective motion; Langevin equation

1. Introduction

Recently, a method for extracting the deterministic and stochastic contributions from time series has been developed [1,2] and applied to analyze many interesting physical systems [3–7]. It is based on the application of the Langevin differential equation [8]

$$\frac{d}{dt}X(t) = g\big(X(t),t\big) + h\big(X(t),t\big)\Gamma(t) \tag{1}$$

where $X(t)$ is the state of the system at time t, the nonlinear function g provides the deterministic change, the function h gives the amplitude of the stochastic part, and $\Gamma(t)$ provides an uncorrelated white noise term with an average value of zero. This equation applies only to stationary, Markovian systems [9]. Time series resulting from molecular dynamics simulations offer another application for this new technique, where the deterministic and probabilistic contributions can be determined directly from the time series. This is a significant application because atomic motions in protein molecules play an important role in the function of the protein. This is especially true of correlated motions of groups of atoms in the protein molecule. Principal component analysis (PCA) is a standard tool used to investigate correlated protein motions and is also known as the essential dynamics (ED) method [10,11]. It was first reported in [12] and has been used extensively since then in studying protein motions [13–16]. ED is able to separate random protein motions from correlated motions. In other words, small-amplitude Gaussian fluctuations are separated from the essential subspace of large an-harmonic motions. The PCA technique consists of the diagonalization of the covariance matrix given by

$$C_{ij} = c_{ij} \tag{2}$$

with

$$c_{ij} = \frac{1}{T}\sum_{t}\big(x_i(t) - \langle x_i(t)\rangle\big)\big(x_j(t) - \langle x_j(t)\rangle\big) \tag{3}$$

where T is the total number of configurations $t = 1, 2, 3, \ldots, T$, $x_i(t)$ are the position coordinates with $i = 1, 2, 3, \ldots, 3N$, N is the number of C^{α} atoms, $\langle x_i(t)\rangle$ is the ensemble average over all configurations [15]. The diagonalization produces eigenvectors and eigenvalues. Each eigenvector defines a direction of concerted atomic motion in the $3N$-dimensional space. The eigenvalue gives the total mean square fluctuation of the protein motion along that eigenvector. The trajectory can be projected along any of the resulting eigenvectors. This results in a time

* Corresponding author.
E-mail address: wqaran@birzeit.edu (W.I. Karain).

0375-9601/$ – see front matter
doi:10.1016/j.physleta.2006.01.073

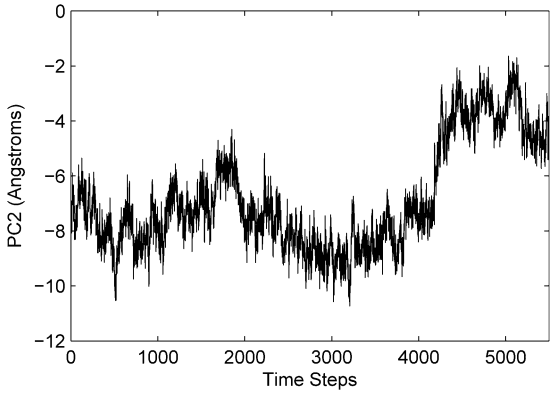

Fig. 1. A typical time series showing the trajectory projection along the second eigenvector PC2.

series showing the fluctuations of the projection for a certain direction (Fig. 1). In this work a set of time series resulting from projecting a 4 ns molecular dynamics trajectory of the protein BPTI along a set of directions given by eigenvectors found using PCA are analyzed.

2. Computational methods

The molecular dynamics simulation and analysis were performed using the program not another molecular dynamics (NAMD) [17] and visual molecular dynamics (VMD) [18]. The starting X-ray structure for the simulation (6PTI) was obtained from the protein data bank [19]. The principal components analysis (PCA) was performed using the MATLAB package (The MathWorks, Natick, MA 01760-2098, USA). Each time series consisted of 8000 points with a time step of 0.5 picoseconds. The set of time series investigated in this work consisted of two groups: the 'significant' group containing the first five eigenvectors with the five largest eigenvalues, making up 80% of all protein motion, and the 'insignificant' group containing the 10th, 20th, and 50th eigenvectors contributing a very small percentage of the correlated protein motions.

A difference form of the Langevin equation (1) can be used to model the system:

$$X(t+\tau) = X(t) + g\big(X(t);\tau\big) + h\big(X(t);\tau\big)\Gamma(t) \tag{4}$$

where τ is the lag time. Eq. (4) gives the differential equation (1) in the limit $\tau \to 0$. In this work, however, there is a minimum τ where Eq. (4) can be used to model the system. A similar limit was reported in a study of heart rate fluctuations [7]. The technique for finding $g(x)$ and $h(x)$ from Eq. (4) is the following: the projection values in the time series are divided into bins, with the middle value of the bin, x, being the representative value. Given any value $X(t)$ within the bounds of a bin, the future value $X(t+\tau)$ is stored. All future values for a bin form a distribution. The average value for this distribution is $x + g(x)$. The deviation value for the distribution gives $h(x)$ [2].

The time series investigated were stationary. The Markovian property for each series was checked by comparing the one-step and the two-step conditional probability densities. Agreement of the two would point to a lack of memory effects [2].

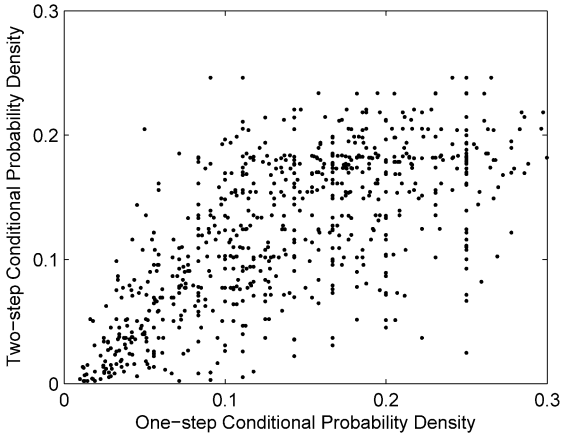

Fig. 2. A scatter plot showing the one-step and two-step conditional probability densities. The plot shows large agreement between the two functions.

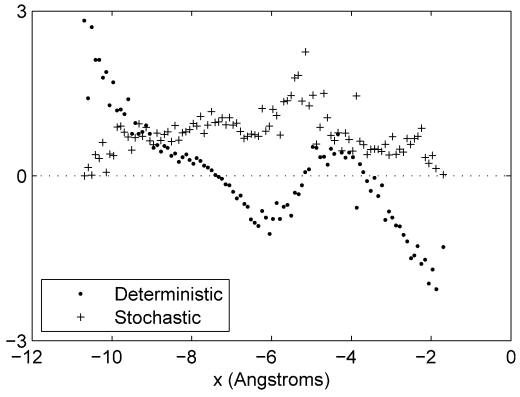

Fig. 3. The deterministic and stochastic contributions to the projection of the time series along PC2.

3. Analysis and discussion

The behavior of the second eigenvector (PC2) is discussed in this section. It showed a bi-stable behavior reported in the literature for a time series dealing with heart rate fluctuations [7]. The time lag used was 100 steps, or 50 picoseconds. The number of bins used to define the distributions was 100. To insure that the system is Markovian, the one-step and two-step conditional probability densities were calculated using histograms and compared (Fig. 2). The two functions agreed significantly, pointing to the lack of memory effects.

The deterministic function $g(x)$ and the stochastic function $h(x)$ are shown in Fig. 3.

The function $g(x)$ shows three fixed points where it crosses the x-axis. The two points on the outside are stable points. A change of the system in either direction attracts the system back to the stable point. In particular, any change away from the region bounded by the two stable points is met by a steep increase or decrease by the system to bring it back to that region. The inside fixed point is unstable. A change in either direction is encouraged by the deterministic contribution. The random

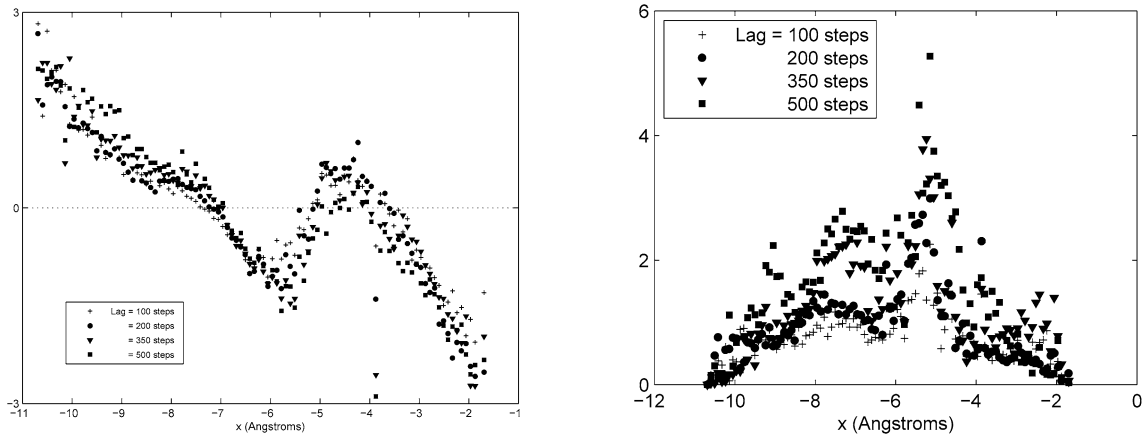

Fig. 4. (a) The deterministic function $g(x)$ versus lag time. (b) The stochastic function $h(x)$ versus lag time.

function $h(x)$ is mainly flat except for a pronounced maximum in the region surrounding the unstable fixed point. A local maximum is also apparent at one of the stable fixed points. One can conclude that the system is more deterministic at the stable fixed points, while being more stochastic in nature near the unstable point. The maximum stochastic value near the unstable point could signal a large random 'kick' that might transfer the system from the unstable point to one of the stable points. The stable points signify two conformations of the protein molecule, and it appears that the system oscillates between the two, using the unstable fixed point as an intermediary step. The deterministic function $g(x)$ shows considerable independence off the lag time (Fig. 4(a)). The stochastic function $h(x)$ shows a similar behavior in general, but a steady increase in the peaks centered at the unstable fixed point and one of the stable fixed points is apparent for large time steps.

A representative value of $g(-6)$ was plotted for lag times from 1 picoseconds to 500 picoseconds (10 steps to 1000 steps) (see Fig. 5). It is clear from the graph that the function is almost independent off the lag time in this range, except for the region of small lag times. This divergence may indicate the presence of measurement noise [20].

The behavior of the deterministic function $g(x)$ was compared for the five elements of the 'significant' eigenvector set, and the three elements of the 'insignificant' set (Fig. 6).

The elements of the significant set show nonlinear behavior of varying complexity. The first eigenvector (case 1) shows one clear stable fixed point. The second fixed point is not well pronounced. It could actually point to two fixed points. One explanation could be that the space around these two poorly defined fixed points is not well explored in the time series. The third eigenvector (case 3) shows one clear stable fixed point. The local minimum to the left of this fixed point could also point to another unexplored region. The protein seems to fluctuate randomly about an average value of zero. The fourth eigenvector (case 4) shows multiple fixed points. Again, some of these fixed points are not very well represented in the time series. The fifth eigenvector (case 5) also shows multiple fixed points, and closely resembles the behavior of the fourth eigenvector. The

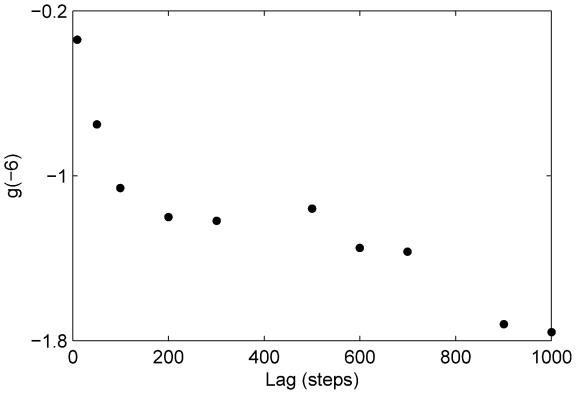

Fig. 5. A representative value of $g(-6)$ as a function of lag time.

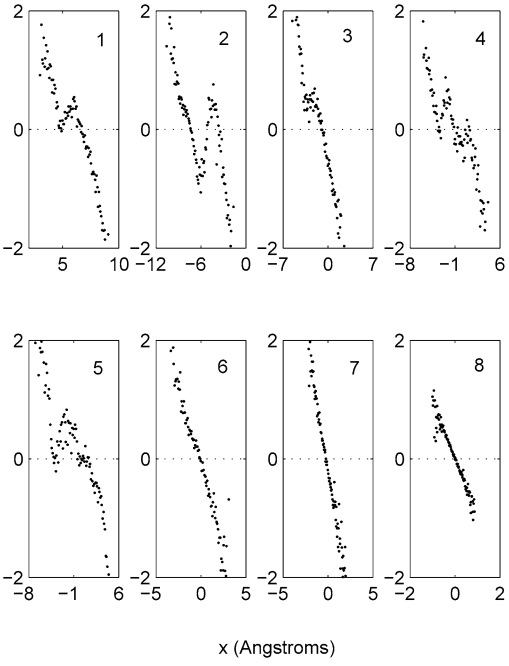

Fig. 6. The deterministic function $g(x)$ for the elements of the significant and insignificant sets. The cases 1–5 are the first five eigenvectors, respectively. The cases 6–8 are the 10th, 20th, and 50th eigenvectors, respectively.

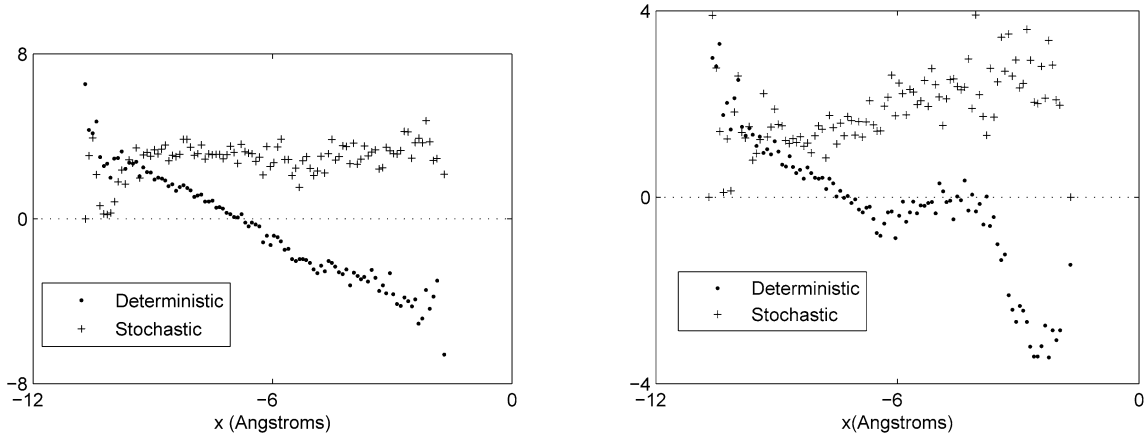

Fig. 7. The deterministic and stochastic contributions for the time series broken up into equal segments and remixed randomly, with segment lengths (a) 30 steps, (b) 300 steps.

cases 6–8 all show one stable fixed point without any nonlinear behavior. The behavior of the stochastic function $h(x)$ for the first five cases shares a common trait of maxima in the nonlinear region of the function $g(x)$. The last three cases show no such behavior with the function being flat throughout.

To insure that this behavior is not a statistical artifact, the time series for PC2 was broken up into equally sized segments, and then remixed randomly (Fig. 7).

For a time segment that is shorter than the lag time (Fig. 7(a)), the nonlinear behavior described above (Fig. 3) disappears. For segment lengths longer than the time lag length (Fig. 7b), the dynamic behavior stays practically the same. This proves that the behavior of the system is inherent in its dynamics.

4. Conclusion

The deterministic and stochastic contributions to time series of correlated protein motions show a rich and complex nonlinear behavior only for projections along directions that contribute a significant percentage of the protein correlated motions. This behavior provides another tool in the analysis of protein motions, and can provide clues to how proteins jump between different conformations. The method reported in this work can also be used to differentiate between significant and insignificant principal components through the behavior of their deterministic and stochastic contributions.

References

[1] S. Siegert, R. Friedrich, J. Peinke, Phys. Lett. A 234 (1998) 275.
[2] R. Friedrich, S. Siegert, J. Peinke, St. Lueck, M. Seifert, M. Lindemann, J. Raethjen, G. Deuschl, G. Pfister, Phys. Lett. A 271 (2000) 217.
[3] J. Gradisek, I. Grabec, S. Siegert, R. Friedrich, Mech. Syst. Signal Process. 16 (5) (2002) 831.
[4] T.D. Frank, P.J. Beek, R. Friedrich, Phys. Lett. A 328 (2004) 219.
[5] P. Sura, S.T. Gille, J. Mar. Res. 61 (2003) 313.
[6] S. Kriso, R. Friedrich, J. Peinke, P. Wagner, Phys. Lett. A 299 (2002) 287.
[7] T. Kuusela, T. Shepherd, J. Hietarinta, Phys. Rev. E 67 (2003) 061904.
[8] I. Nobuyuki, W. Shinzo, Stochastic Differential Equations and Diffusion Processes, North-Holland, Amsterdam, 1981.
[9] M. Kijima, Stochastic Processes with Applications to Finance, Chapman & Hall, 2003.
[10] G. Basu, A. Kitao, F. Hirata, N. Go, J. Am. Chem. Soc. 116 (1994) 6307.
[11] S. Hayward, A. Kitao, F. Hirata, N. Go, J. Mol. Biol. 234 (1993) 1207.
[12] A. Garcia, Phys. Rev. Lett. 68 (1992) 2696.
[13] D. Van Aalten, A. Amadei, A. Linssen, V. Eijsink, G. Vriend, H. Berendsen, Proteins: Struct. Funct. Genet. 22 (1995) 45.
[14] R. Laatikanien, J. Saarela, K. Tuppurainen, T. Hassinen, Biophys. Chem. 73 (1998) 1.
[15] A. Amadei, A. Linssen, H. Berendsen, Proteins: Struct. Funct. Genet. 17 (1993) 412.
[16] M.A. Balsera, W. Wriggers, Y. Oono, K. Schulten, J. Phys. Chem. 100 (1996) 2567.
[17] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten, J. Comput. Phys. 151 (1999) 283.
[18] W. Humphrey, A. Dalke, K. Schulten, J. Mol. Graph. 14 (1996) 33.
[19] A. Wlodawer, J. Nachman, G.L. Gilliland, W. Gallagher, C. Woodward, J. Mol. Biol. 198 (1987) 469.
[20] M. Siefert, A. Kittel, R. Friedrich, J. Peinke, Europhys. Lett. 61 (2003) 466.