BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2501–2502
doi:10.1093/bioinformatics/btt413

Structural bioinformatics

Advance Access publication July 17, 2013

deltaGseg: macrostate estimation via molecular dynamics simulations and multiscale time series analysis

Diana H. P. Low[1] and Efthymios Motakis[2,*]

[1]Institute of Molecular and Cell Biology, Epigenetics, Chromatin and Differentiation, 61 Biopolis Street, Proteos #03-06, Singapore 138673 and [2]Bioinformatics Institute, Genome and Gene Expression Data Analysis, 30 Biopolis Street, Matrix #07-01, Singapore 138671

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Binding free energy calculations obtained through molecular dynamics simulations reflect intermolecular interaction states through a series of independent snapshots. Typically, the free energies of multiple simulated series (each with slightly different starting conditions) need to be estimated. Previous approaches carry out this task by moving averages at certain decorrelation times, assuming that the system comes from a single conformation description of binding events. Here, we discuss a more general approach that uses statistical modeling, wavelets denoising and hierarchical clustering to estimate the significance of multiple statistically distinct subpopulations, reflecting potential macrostates of the system. We present the *deltaGseg* R package that performs macrostate estimation from multiple replicated series and allows molecular biologists/chemists to gain physical insight into the molecular details that are not easily accessible by experimental techniques.

Availability: *deltaGseg* is a Bioconductor R package available at http://bioconductor.org/packages/release/bioc/html/deltaGseg.html.

Contact: emotakis@hotmail.com

Received on April 9, 2013; revised on June 27, 2013; accepted on July 15, 2013

1 INTRODUCTION

Free energy calculations are extensively used to understand intermolecular interactions. Among the several free energy estimation methods (Bash *et al*., 1987), the MMGB(PB)SA (molecular mechanics with generalized Born and surface area solvation) has gained popularity because of its fast performance and its ability to predict trends in affinities (Kollman *et al*., 2000). Traditionally in MMGB(PB)SA, one performs molecular dynamics simulations (MDs) and selects a set of representative snapshots from a stretch of the trajectory that is deemed to be at equilibrium (or converged). The MDs are often replicated multiple times (typically from 3 to 6), each time with slightly different starting conditions. Of primary interest is the calculation of the relative free binding energy $\Delta G = G_{complex} - G_{receptor} - G_{ligand}$, which is estimated from the replicated MDs of the complex, receptor and ligand terms. The average ΔG across the set of selected snapshots is then reported as the net free energy, which is subject to large variation.

All average-based methods assume that the ΔG series originates from a single conformation description of binding events. They are designed to minimize autocorrelation and estimate the series on appropriate (de)correlation times (otherwise the statistical error would be underestimated) via the method of statistical inefficiency or block-averaging (Flyvbjerg and Petersen, 1989). It has been discovered that this approach is sensitive to the drift during the course of the simulation and, especially in long simulations, gives exponentially increased correlation times (Genheden and Ryde, 2011). The unappealing consequence of the monotonic relationship between the correlation and the simulation times is that longer simulations lead to less data points for reliable inference.

Our new *deltaGseg* methodology relaxes the restrictive 'single conformation' assumption and estimates multiple distinct subpopulations, reflecting potential macrostates upon experimental validation, within a series and its replicates (Zhou *et al*., 2012). Rigorous statistical modeling avoids the correlation time estimation without compromising the statistical validity of the estimates.

2 APPROACH

We present the main functions of the *deltaGseg* R package that implements and enriches with alternative options our original MMPBSA_segmentation algorithm (Zhou *et al*., 2012) for macrostate estimation in these steps: (i) data visualization and adjustment for preprocessing, (ii) data segmentation via change-point analysis for initial subpopulations estimation, followed by wavelets denoising for series estimation and (iii) hierarchical clustering for the macrostates estimation. Theoretical results (Nason, 2008) and *deltaGseg* diagnostic plots (function `diagnosticPlots`) show the estimated series and its insignificant autocorrelation coefficients. Our approach is not only limited to binding energies between protein-ligand systems, but can be also used to analyze different conformational spaces in the folding of peptides or protein by identifying different energetic states along the simulation.

3 METHODS

3.1 Visualization and preprocessing

This task is carried out by the `parseTraj` function. We load each time series $X_t^r(t = 1, \ldots T; r = 1, \ldots R)$—$t$ denotes the time and r the replicate—and, initially, we test whether each X_t^r is weakly (trend) stationary by estimating the augmented Dickey–Fuller P-value (Dickey and Fuller, 1979). If the assumption is not satisfied, we visualize and split the series. Each generated subseries is once again tested for stationarity and, if passes, it is further analyzed independently (see next steps). Our experience with

*To whom correspondence should be addressed.

MD simulation data indicates that weak stationarity holds for splitted series. If the problem persists, series differentiation is suggested by calculating of the first differences, $X^r[t] - X^r[t-1]$. Differentiation is appropriate for subseries with an apparent trend, which will be effectively removed.

3.2 Series segmentation and denoising

These tasks are performed by the `denoiseSegments` function. Each subseries X^r_t of Section 3.1 undergoes statistical multiple change-point analysis by the Binary Segmentation method of Auger and Lawrence (1989). Binary Segmentation automatically splits X^r_t into Q^r segments, $Q^r \in [2, maxQ]$. The user sets the initial parameter $maxQ$ (e.g. $maxQ = 15$ for a series of $T = 5{,}000$; case study below), and the algorithm estimates the significant Q^r segments at level $\alpha = 0.01$. Subsequently, the $X^r_{t,q}$ data of each segment $q^r, q^r = 1, \ldots, Q^r$, is modeled and smoothed by 1D wavelets via *wavethresh* (Nason, 2008). The user may adjust various modeling and smoothing parameters such as the degree of smoothness, the wavelet family (default is the Haar wavelet with step smoothing function) and the minimum length, L_q, for a segment q to be accepted ($L_q \leq 0.5 \times L_{X^r_t}$).

3.3 Data summarization and macrostate identification

Final estimation is carried out by the function `clusterSegments`. Here, the wavelet smoothed $X^r_{t,q}$ are summarized into a vector of quantiles V^r_q ('identity' vectors). All $\sum Q^r$ segments are simultaneously analyzed by feeding the Euclidean distances among the V_{q^r} into hierarchical clustering. The clustering can be performed either by the *pvclust* algorithm (Shimodaira, 2004) that assesses the significance of each cluster / subpopulation where potential macrostates can be derived from (see case study below) or by simple average linkage clustering and a user-defined number of subpopulation to be identified. Both options imply manual (subjective) selection of the final subpopulations but the *deltaGseg* interactive plots aid the user in this decision. The *pvclust* provides an additional statistical assessment measure.

4 CASE STUDY

We applied *deltaGseg* to the interaction data of the ErbB2 receptor, a member of the growth factor receptor tyrosine kinase family, and adenosine triphosphate, a natural ligand to ErbB2 (Pereira *et al.*, submitted for publication). Snapshots from three replicates (5000 time points each) were used (Fig. 1A). The data preprocessing algorithm suggested splitting the third replicate at $t = 2775$ (vertical dotted line). Segmentation, denoising and summarization were performed for each of the four subseries. The *pvclust* was used to estimate the significance of each cluster at each tree node (Fig. 1B).

We estimated six distinct subpopulations at significance level $\alpha = 0.01$. Structural characterization of these ensembles can determine if they represent few different binding modes between ErBB2 and adenosine triphosphate that could arise from the exploration of different conformational spaces as it has been observed for an inhibitor of these kinases (Huang and Rizzo, 2012). Ideally, every single population should be characterized by specific structural determinants which might be difficult to prove experimentally. Our approach has the advantage of applying statistical methods to identify subpopulations in the bound state of a system, providing a more efficient manner to characterize quantitatively residues that contribute either in a transient fashion or permanently in a given conformational space.

The analysis carried out here suggests the existence of two apparent binding modes with a difference of ~36 kcal/mol between them. Of the three replicates, the first extensively explores the looser conformation, the second preferentially explores the tighter binding mode and the third appears to oscillate between the two modes. When these different binding modes can be detected with experimental methods, they will be assigned to different macrostates that can be observed at the macroscopic level.

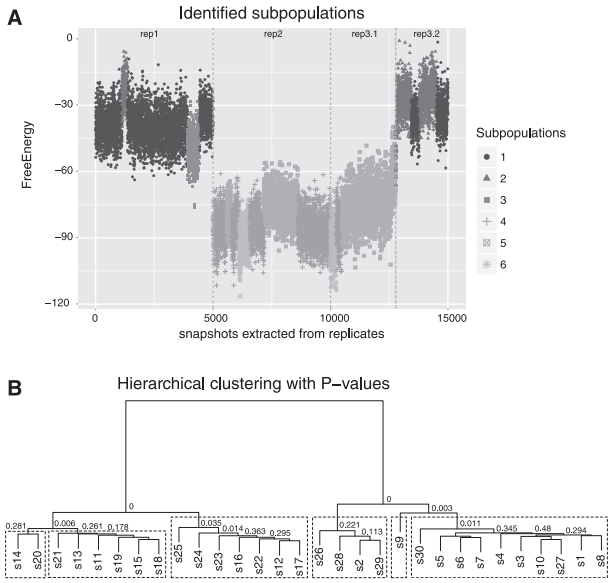

Fig. 1. (**A**) Three replicates of 5000 time points each grouped into six subpopulations. (**B**) The hierarchical clustering result with *P*-values

ACKNOWLEDGEMENT

We thank Dr Gloria Fuentes for providing the data and offering valuable advice and suggestions.

Conflict of Interest: none declared.

REFERENCES

Auger,I. and Lawrence,C. (1989) Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, **51**, 39–54.

Bash,P.A. *et al.* (1987) Free energy calculations by computer simulation. *Science*, **236**, 564–568.

Dickey,D. and Fuller,W. (1979) Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.*, **74**, 427–431.

Flyvbjerg,H. and Petersen,H.G. (1989) Error estimates on averages of correlated data. *J. Chem. Phys.*, **91**, 461–466.

Genheden,S. and Ryde,U. (2011) Comparison of the efficiency of the lie and mm/gbsa methods to calculate ligand-binding energies. *J. Chem. Theory Comput.*, **7**, 3768–3778.

Huang,Y. and Rizzo,R.C. (2012) A water-based mechanism of specificity and resistance for lapatinib with erbb family kinases. *Biochemistry*, **51**, 2390–2406.

Kollman,P.A. *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models acc. *Acc. Chem. Res.*, **33**, 889–897.

Nason,G. (2008) *Wavelet Methods in Statistics with R*. Springer, New York.

Shimodaira,H. (2004) Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.*, **32**, 2616–2641.

Zhou,W. *et al.* (2012) Macrostate identification from biomolecular simulations through time series analysis. *J. Chem. Inf. Model*, **52**, 2319–2324.