BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 16 2010, pages 2060–2061
doi:10.1093/bioinformatics/btq328

Gene expression

Advance Access publication June 18, 2010

CplexA: a *Mathematica* package to study macromolecular-assembly control of gene expression

Jose M.G. Vilar[1,2,*] and Leonor Saiz[3,*]

[1]Biophysics Unit (CSIC-UPV/EHU) and Department of Biochemistry and Molecular Biology, University of the Basque Country, PO Box 644, 48080 Bilbao, [2]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain and [3]Department of Biomedical Engineering, University of California, 451 E. Health Sciences Drive, Davis, CA 95616, USA

Associate Editor: Trey Ideker

ABSTRACT

Summary: Macromolecular assembly coordinates essential cellular processes, such as gene regulation and signal transduction. A major challenge for conventional computational methods to study these processes is tackling the exponential increase of the number of configurational states with the number of components. CplexA is a *Mathematica* package that uses functional programming to efficiently compute probabilities and average properties over such exponentially large number of states from the energetics of the interactions. The package is particularly suited to study gene expression at complex promoters controlled by multiple, local and distal, DNA binding sites for transcription factors.

Availability: CplexA is freely available together with documentation at http://sourceforge.net/projects/cplexa/

Contact: j.vilar@ikerbasque.org; lsaiz@ucdavis.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on April 20, 2010; accepted on June 13, 2010

1 INTRODUCTION

The study of the cellular behavior from the molecular components often requires approximations in terms of chemical reactions. However, there are many instances, such as combinatorial macromolecular assembly, that cannot be efficiently described in terms of chemical reactions (Saiz and Vilar, 2006). Macromolecular complexes are typically made of smaller building blocks with a modular organization that can be combined in a number of different ways. The result of each combination is a specific molecular species. Therefore, there are potentially as many reactants as the number of possible ways of arranging the different elements, which grows exponentially with the number of the constituent elements.

Several approaches have been developed to tackle this exponentially large multiplicity in the number of states. They involve a diversity of methodologies that range from stochastic configuration sampling (Le Novere and Shimizu, 2001; Saiz and Vilar, 2006) to automatic generation of all the underlying equations (Hlavacek *et al*., 2006). The complexity of the general problem makes each of these approaches work efficiently only on a particular type of problems, be it conformational changes, multi-site phosphorylation or oligomerization (Borisov *et al*., 2006; Bray and Lay, 1997; Saiz and Vilar, 2006).

The package CplexA focuses on macromolecular assembly on a template. The prototypical example is a complex promoter, where DNA provides a flexible template for the assembly of transcription factors. CplexA provides mathematical tools to infer the probability of having a given set of configurations. In the case of a promoter, it would be the probability of having a pattern of transcription factors bound, which can be used to infer the resulting transcription rate in a way that can be integrated with other software to study the dynamics of cellular networks (Shapiro *et al*., 2003).

This type of systems has traditionally been studied by writing a table with entries for each state and the corresponding free energies and associated probabilities, which are used to compute average quantities such as effective transcription rates (Ackers *et al*., 1982). As the number of states increases exponentially, the approach becomes impracticable. In this type of systems, however, it is possible to take advantage of the unambiguous structures that macromolecular complexes typically have on a template and use 'table-centric' equivalent mathematical approaches that are able to capture this complexity in simple terms (Saiz and Vilar, 2006).

2 METHODS

The mathematical approach underlying CplexA is discussed in detail in Saiz and Vilar (2006). It specifies the system by a set of N state variables, $S=\{s_1,\ldots s_i,\ldots s_N\}$, that can be either 1 to indicate that a property is present (e.g. binding or conformation) or 0 to indicate that it is not. The free energy, $\Delta G(S)$, and a configuration pattern, $\Gamma(S)$, can generally be expressed as a function of these state variables. The probability of the configuration pattern is obtained from

$$\bar{\Gamma}=\frac{\sum_S \Gamma(S)e^{-\Delta G(S)/RT}}{\sum_S e^{-\Delta G(S)/RT}} \quad (1)$$

by computing the thermodynamic average over all 2^N possible values of S.

3 APPLICATION

The package CplexA provides the function AvConf[Γ,ΔG, S] that computes the thermodynamic average $\bar{\Gamma}$. Figure 1 illustrates the use of CplexA with *Mathematica 7* to compute the effective transcription rate at the P_{RM} promoter of Phage λ (Saiz and Vilar, 2006). This system consists of two sets of three contiguous binding sites for the CI dimer. The two sets, known as left and right operators, are 2 kb

*To whom correspondence should be addressed.


```
In[8]:= S:={sOL1, sOL2, sOL3, sOR1, sOR2, sOR3, sL}

In[9]:= ΔGλ:=sOR1(-12.7-RT Log[n])+sOR2(-10.7-RT Log[n])+
          sOR3(-10.2-RT Log[n])-3 sOR1 sOR2-3 sOR2 sOR3+
          3 sOR1 sOR2 sOR3 + sOL1(-13.8-RT Log[n])+
          sOL2(-12.1-RT Log[n])+sOL3(-12.4-RT Log[n])-
          2.5 sOL1 sOL2-2.5 sOL2 sOL3+2.5 sOL1 sOL2 sOL3+
          sL(21-21.2 sOL1 sOL2 sOR1 sOR2-3 sOL3 sOR3);

In[10]:= Γrm:=
           (1-sOR3)(45+((240-45)sL+(460-45)(1-sL))sOR2);

In[11]:= AΓrm=AvConf[Γrm,ΔGλ,S];

In[12]:= Show[plotPRM,Plot[AΓrm/.{RT->0.6,n->NF[nn]},
           {nn,0,2.5},PlotStyle->{Red,Thick}]]
```

Out[12]=

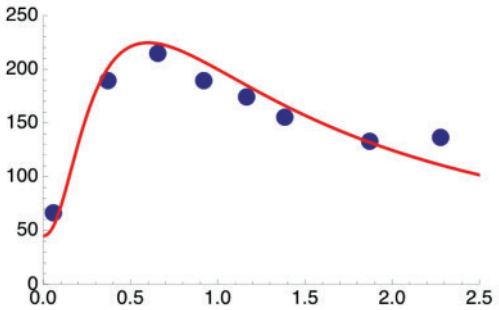

Fig. 1. Use of CplexA with *Mathematica 7* to compute the effective transcription rate at the P_{RM} promoter of Phage λ. The graph in line 'Out[12]' shows the computed average transcription (solid red line) as a function of the normalized CI monomer concentration. The filled circles correspond to the experimentally measured activity of the P_{RM} promoter (Dodd *et al*., 2004; for details, see Supplementary Material).

apart from each other. CI dimers bound at different operators can interact with each other by looping the intervening DNA. In this case, just a few lines of code, from lines 'In[8]' to 'In[11]' in Figure 1, can achieve the same results as a table with entries for each of the 128 states. The state of the system, S, is described by six binding and one looping state variables in line 'In[8]'. The free energy, $\Delta G(S)$, in line 'In[9]' includes, in a very compact manner, binding to each of the six sites as a function of the dimer concentration, interactions between neighboring dimers, DNA looping and the formation of octamers and tetramers between dimers bound at different sets of binding sites. The transcription rate, $\Gamma(S)$, as a function of the binding and looping state is given in line 'In[10]'. Its average value, $\bar{\Gamma}$=AvConf[Γ,ΔG, S], closely matches the experimental data on the transcriptional activity of the promoter (Dodd *et al*., 2004).

CplexA also provides the function DGTable[ΔG, S], which constructs a table with the free energy and statistical weight (Boltzmann factor) of each state that has a non-zero probability.

4 IMPLEMENTATION

The critical issue in the implementation of the function AvConf[Γ,ΔG, S] is dealing with the combinatorial explosion in the number of states. Using state variables overcomes the combinatorial explosion in the specification of the problem but not in the sum over all the states, which still grows as 2^N. A fundamental advantage of using a computer algebra system, such as *Mathematica*, over imperative programming languages, such as Fortran, C or Java, is that it allows for the direct manipulation of functions. In CplexA, the implementation of the sum over all possible values of S in the numerator and denominator of Equation (1) is performed in N steps, rather than in 2^N, by using the backwards recursion

$$f_{N-1}(s_1,\ldots s_{N-1}) = f_N(s_1,\ldots s_{N-1},0) + f_N(s_1,\ldots s_{N-1},1).$$

Starting this recursion with the functions

$$f_N(s_1,\ldots s_N) \equiv \Gamma(S)e^{-\Delta G(S)/RT} \text{ and } f_N(s_1,\ldots s_N) \equiv e^{-\Delta G(S)/RT}$$

leads to the sought values of the sums as

$$\sum_S \Gamma(S)e^{-\Delta G(S)/RT} = f_0 \text{ and } \sum_S e^{-\Delta G(S)/RT} = f_0,$$

respectively, after the N steps of the recursion have been performed. With this method, the actual computational complexity depends on the specific form of $f_N(s_1,\ldots s_N)$ and does not necessarily increase proportionally to the number of states. For instance, for a linear array of binding sites with next-neighbor interactions, the CPU time needed to compute the average occupancy for the case of 40 sites is only a factor ~8 higher than that needed for 20 sites, whereas the number of states increases by a factor $\sim 10^6$ (Supplementary Material).

Funding: Ministerio de Ciencia e Innovacion (FIS2009-10352); University of California, Davis.

Conflict of Interest: none declared.

REFERENCES

Ackers,G.K. *et al*. (1982) Quantitative model for gene-regulation by lambda-phage repressor. *Proc. Natl Acad. Sci. USA*, **79**, 1129–1133.

Borisov,N.M. *et al*. (2006) Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains. *Biosystems*, **83**, 152–166.

Bray,D. and Lay,S. (1997) Computer-based analysis of the binding steps in protein complex formation. *Proc. Natl Acad. Sci. USA*, **94**, 13493–13498.

Dodd,I.B. *et al*. (2004) Cooperativity in long-range gene regulation by the lambda CI repressor. *Genes Dev*., **18**, 344–354.

Hlavacek,W.S. *et al*. (2006) Rules for modeling signal-transduction systems. *Sci. STKE*, **2006**, re6.

Le Novere,N. and Shimizu,T.S. (2001) STOCHSIM: modelling of stochastic biomolecular processes. *Bioinformatics*, **17**, 575–576.

Saiz,L. and Vilar,J.M.G. (2006) Stochastic dynamics of macromolecular-assembly networks. *Mol. Syst. Biol*., **2**, 2006.0024.

Shapiro,B.E. *et al*. (2003) Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics*, **19**, 677–678.