Published in final edited form as:

Nucleosides Nucleotides Nucleic Acids. 2010 June ; 29(4): 427–432. doi:10.1080/15257771003741174.

How to use the computing environment R to analyze ATP-induced ribonucleotide reductase R1 hexamerization data

T. Radivoyevitch¹

T. Radivoyevitch: txr24@case.edu

¹ Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106, USA

Abstract

R is an object oriented free and open source statistical computing environment. The R package Combinatorially Complex Equilibrium Model Selection is being developed to meet the analysis challenges of ribonucleotide reductase (RNR). An example of its use is given here. This example involves ATP-induced R1 hexamerization dynamic light scattering data that suggests that R1 hexamers have two types of *a*-sites, one that binds ATP and another that does not (here R1 is the large subunit of RNR).

Keywords

R; Combinatorial Complexity; equilibrium model selection; RNR

INTRODUCTION

Ribonucleotide reductase (RNR) R1 monomers have 5 possible catalytic site states (empty or filled with 1 of 4 NDPs), 5 possible selectivity (*s*-) site states (empty or filled with ATP, dATP, dTTP or dGTP), 3 possible activity (*a*-) site states (empty or filled with ATP or dATP) (1), and, possibly (2), two *h*-site states (empty or bound by ATP). Based on this, R1 monomers could potentially exist in any of $5 \times 5 \times 3 \times 2 = 150$ states and R1 hexamers could exist in any of $150^6 (= \sim 10^{13})$ states. If RNR R2 dimers are also considered, the number of possible RNR complexes is even greater. As each such complex has a complete dissociation constant associated with it that could potentially be either independently estimated or hypothesized to be infinite (in which case the concentration of the corresponding complex is hypothesized to be approximately zero), the number of plausible R1 equilibrium models is on the order of $2^{10^{13}}$. And if binary dissociation constants are also hypothesized to equal each other, the number of models is greater still. RNR modeling difficulties thus arise because: 1) the number of RNR complexes is much greater than the number of RNR reactants (i.e. RNR is combinatorially complex), and 2) the number of models is much greater than the number of complexes (i.e. model spaces grow combinatorially in the number of model parameters of the full model that defines the model space, e.g. see `topology` in the R code below). The R package Combinatorially Complex Equilibrium Model Selection, `ccems`, attempts to deal with such complexities. Through the use of a high level `ccems` function, large complicated RNR model spaces can be generated and fitted to RNR data. This is illustrated in this short paper.

MATERIALS AND METHODS

The R package `ccems` is available from the Comprehensive R Archive Network (CRAN) (3). A 32-core ROCKS (4) cluster was used to execute the code. The ATP induced R1

hexamerization Dynamic light scattering (DLS) data in Fig. 1 of (2) was digitized using plotDigitizer (5).

RESULTS

The following R script explores the hypothesis that *h*-sites are not needed to explain the ATP induced R1 hexamerization DLS data in Fig. 1 of (2). To this end, it uses *ccems* to automatically generate 3410 models, the vast majority of which include terms that correspond to occupied *h*-site complexes, and fit them to the DLS data.

```
library(ccems) # loads the ccems package into R
topology <- list(
  heads=c("R1X0", "R2X2", "R4X4", "R6X6"), # X = ATP and R = R1
  sites=list( # s-sites are already filled only in (j>1)-mers
    a=list( # a-site
      m=c("R1X1"), # monomer
      d=c("R2X3", "R2X4"), # dimers
      t=c("R4X5", "R4X6", "R4X7", "R4X8"), # tetramers
      h=c("R6X7", "R6X8", "R6X9", "R6X10", "R6X11", "R6X12") # hexamers
    ), # tails of a-site threads serve as heads of h-site threads
    h=list( # h-site
      m=c("R1X2"), # 1-mer
      d=c("R2X5", "R2X6"), # 2-mers
      t=c("R4X9", "R4X10", "R4X11", "R4X12"), # 4-mers
      h=c("R6X13", "R6X14", "R6X15", "R6X16", "R6X17", "R6X18") #6-mers
    )
  )
g=mkg(topology)
dd=subset(RNR,(year= 2002)&(fg= 1)&(X>0),select=c(R, X, m, year)) # get
data
names(dd)[1:2]=c("RT", "XT") # changes names to indicate totals
cpus=c("localhost"=4, "compute-0-0"=4, "compute-0-1"=4, "compute-0-2"=4,
  "compute-0-3"=4, "compute-0-4"=4, "compute-0-5"=4, "compute-0-6"=4)
top10=ems(dd,g,cpusPerHost=cpus,maxTotalPs=3,ptype="SOCK",
  KIC=100, topN=10, transform= "none")
```

In this code, the main work for the user is to specify the full model via the structure *topology*. In this structure, nodes in the *heads* field are complexes that are always reached from free reactants via a single complete dissociation constant edge, and nodes in the *a* and *h* lists of the *sites* field are complexes that can also be reached from head nodes using binary dissociation constants, i.e. in addition to being reached from free reactants using single complete dissociation constants. In *topology*, complexes are represented by strings where single characters represent reactants (*X* = ATP and *R* = R1) and subsequent integers indicate the number of reactant instances in the complex, i.e. complexes with the same reactant numbers are not distinguished. The structure *topology* specifies a model where *s*-sites are never filled in monomers and always prefilled in oligomers: in *heads*, the number following *X* is 0 when the number following *R* is 1, and these numbers are equal otherwise. It also assumes that *a*-sites fill after *s*-sites and that *h*-sites fill last.

The `ccems` function `mkg` interprets topology and converts it into a generic full model object. This new list of lists object is much bigger than `topology` and includes among other things C codes and R functions that are needed to solve the underlying models (6). This structure can be seen by typing `g` at the R prompt after the assignment `g=mkg(topology)` has been made.

The line of R code that follows `mkg` assigns the relevant subset of the `ccems` dataframe `RNR` to the object `dd`. Several other `RNR` datasets are included in `RNR` and `ccems` also includes thymidine kinase 1 literature data in its dataframe `TK1`.

The next line specifies the names and cpu numbers of the computers used. The names shown are defaults for a ROCKS linux cluster and should be changed to match the cluster at hand. This vector of named integers is used in the next line which calls the function `ems`. In addition to passing `ems` the data `dd` and the generic full model object `g`, this line specifies the maximum number of parameters of the model space that is to be automatically generated and fitted, the type of parallelization method used (e.g. “SOCK”, see the R package `snow` [small network of workstations] for details), and the initial conditions of the dissociation constant parameters (if these were not specified a default of 1 would be used and fewer models would converge; nonlinear least squares is used in the optimizations). Other options in this line state that the data will not be transformed (i.e. to stabilize the variance) and that the html output (see Table 1) is a list of the top 10 (as ranked by the Akaike Information Criterion (7)) fitted models.

After the first line of code which loads the R package `ccems` into R, `ccems` help can be obtained by placing a question mark in front of the `ccems` object. For example, help regarding the datasets `TK1` and `RNR` that are built-in to `ccems` can be obtained by typing `?TK1` and `?RNR` at the R command line prompt and help regarding the `ccems` functions `mkg` and `ems` can be obtained by typing `?mkg` and `?ems`.

DISCUSSION

The top 10 models out of 3410 automatically generated and fitted to recent ATP induced R1 hexamerization DLS data (2) do not include model terms with higher powers of ATP [=X] than 9 (see Table 1). Since most models in this space include at least one occupied *h*-site term, this suggests that under the experimental conditions of this dataset, $\sim 1/2$ of the *a*-sites in R1 hexamers are not bound by ATP. This conclusion is consistent with R1 hexamers appearing qualitatively as in Figure 1 where R1 hexamerization partitions *a*-sites into two groups, those that bind ATP and those that do not. It is tempting to speculate that those *a*-sites that do not bind ATP are reserved for dATP binding to allow rapid negative feedback as might be needed to quench an *s*-site mediated dATP positive feedback loop that threads through the dNTP supply system as $dATP \rightarrow dCTP \rightarrow dUMP \rightarrow dTTP \rightarrow dGTP \rightarrow dATP$. Assuming *h*-sites fill only after *a*-sites fill, that only half of the *a*-sites are filled implies that the dataset analyzed does not support the existence of *h*-sites. The R script above shows how `ccems` can be used to answer a complex question. A limitation of `ccems` is that it currently generates model spaces automatically only for two reactant scenarios; a software upgrade that will handle three reactant scenarios is currently being planned.

Acknowledgments

This work was supported by the National Cancer Institute grant K25 CA104791.

References

1. Thelander L, Reichard P. Reduction of ribonucleotides. *Annu Rev Biochem* 1979;48:133–158. [PubMed: 382982]

2. Kashlan OB, Scott CP, Lear JD, Cooperman BS. A comprehensive model for the allosteric regulation of mammalian ribonucleotide reductase. Functional consequences of ATP- and dATP-induced oligomerization of the large subunit. *Biochemistry* 2002;41:462–474.
3. Comprehensive R Archive Network. <http://cran.case.edu/>
4. Rocks. <http://www.rocksclusters.org/wordpress/>
5. Plot Digitizer. <http://plotdigitizer.sourceforge.net/>
6. Radivoyevitch T. Equilibrium model selection: dTTP induced R1 dimerization. *BMC Systems Biology* 2008;2:15. [PubMed: 18248678]
7. Burnham, KP.; Anderson, DR. Model Selection and Multimodel Inference: A Practical-Theoretic Approach. Springer-Verlag; 2002.

Figure 1.

A hypothesized structure of R1 hexamers. In this picture R1 hexamer formation splits two identical *a*-sites into two different types. This could account for ATP binding to only 3 of 6 hexamer *a*-sites.

Table 1

HTML output file of R script.

RX Model Space

CPU time using 32 cpu(s) is 22.8 hours

Model Space Size is 3410

[MS0] ... [MS5]: 0, 13, 288, 3109, 0, 0

Best AICs: 1e+06, 141.2, 142.7, 144.4, 1e+06, 1e+06

Parallelization method: SOCK

Model space search method: brute force

Model type: total concentration constraints

Model	Parameter	Initial Value	Optimal Value	Confidence Interval
1 #####J#####.3	R6X8	100.000 ¹³	63.101 ¹³	(59.878 ¹³ , 66.175 ¹³)
	SSE	840487.830	7726.693	
	AIC	211.573	141.234	
	Cpu	0.000	3.490	fit succeeded
2 #####J#####.91	R2X4	100.000 ⁵	432.997 ⁵	(311.064 ⁵ , 601.845 ⁵)
	R6X8	100.000 ¹³	62.796 ¹³	(59.878 ¹³ , 66.175 ¹³)
	SSE	778605.476	6873.411	
	AIC	213.608	142.660	
	Cpu	0.000	5.269	fit succeeded
3 #####J#####.4	R6X9	100.000 ¹⁴	70.367 ¹⁴	(66.686 ¹⁴ , 74.228 ¹⁴)
	SSE	841284.881	8624.411	
	AIC	211.588	142.883	
	Cpu	0.000	2.945	fit succeeded
4 #####J#####.92	R2X4	100.000 ⁵	347.719 ⁵	(270.426 ⁵ , 445.858 ⁵)
	R6X9	100.000 ¹⁴	69.748 ¹⁴	(66.212 ¹⁴ , 73.175 ¹⁴)
	SSE	869203.983	6997.670	
	AIC	215.259	142.929	
	Cpu	0.000	4.443	fit succeeded
5 #####J#####.128	R4X7	100.000 ¹⁰	134.979 ¹⁰	(116.746 ¹⁰ , 156.022 ¹⁰)
	R6X9	100.000 ¹⁴	69.828 ¹⁴	(66.212 ¹⁴ , 73.700 ¹⁴)
	SSE	361186.068	7009.212	
	AIC	202.086	142.954	
	Cpu	0.000	6.679	fit succeeded
6 #####J#####.189	R6X9	100.000 ¹⁴	70.158 ¹⁴	(66.686 ¹⁴ , 73.700 ¹⁴)
	R2X5	100.000 ⁶	497.972 ⁶	(403.429 ⁶ , 611.960 ⁶)
	SSE	NaN	7046.696	
	AIC	NaN	143.034	
	Cpu	0.000	5.392	fit succeeded
7 #####J#####.188	R6X9	100.000 ¹⁴	69.061 ¹⁴	(65.273 ¹⁴ , 73.175 ¹⁴)
	R1X2	100.000 ²	1783.464 ²	(992.275 ² , 3294.468 ²)
	SSE	753147.502	7066.620	

Nucleosides Nucleotides Nucleic Acids. Author manuscript; available in PMC 2011 June 1.