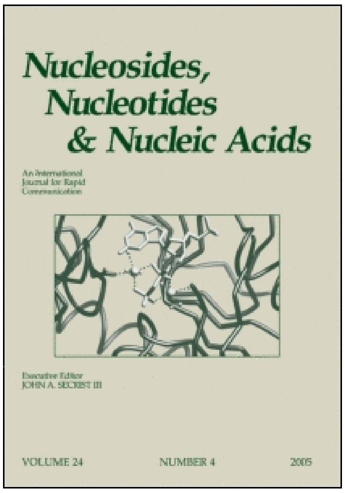

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

How to Use the Computing Environment R to Analyze ATP-Induced Ribonucleotide Reductase R1 Hexamerization Data

T. Radivoyevitch[a]
[a] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

Online publication date: 11 June 2010

To cite this Article Radivoyevitch, T.(2010) 'How to Use the Computing Environment R to Analyze ATP-Induced Ribonucleotide Reductase R1 Hexamerization Data', Nucleosides, Nucleotides and Nucleic Acids, 29: 4, 427 — 432
To link to this Article: DOI: 10.1080/15257771003741174
URL: http://dx.doi.org/10.1080/15257771003741174

Nucleosides, Nucleotides and Nucleic Acids, 29:427–432, 2010
Copyright © Taylor and Francis Group, LLC
ISSN: 1525-7770 print / 1532-2335 online
DOI: 10.1080/15257771003741174

HOW TO USE THE COMPUTING ENVIRONMENT R TO ANALYZE ATP-INDUCED RIBONUCLEOTIDE REDUCTASE R1 HEXAMERIZATION DATA

T. Radivoyevitch
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

□ *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 × 5 × 3 × 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 $2^\wedge(10^{13})$. And, if binary dissociation constants are also hypothesized to equal each other, the number of models is greater still. RNR

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

Address correspondence to T. Radivoyevitch, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA. E-mail: txr24@case.edu

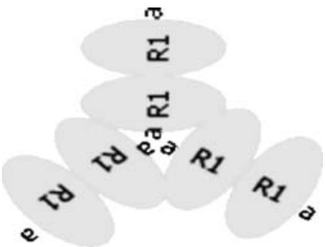

FIGURE 1 A hypothesized structure of R1 hexamers. R1 hexamer formation splits two identical *a*-sites into two different types; this could account for ATP binding to only 3 of 6 hexamer *a*-sites.

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 article.

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 Figure 1 of Kashlan et al.[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 Figure 1 of Kashlan et al.[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, that is, 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, that is, 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 and 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, $\sim\frac{1}{2}$ 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→dCTP→dUMP→dTTP→dGTP→dATP. Assuming *h*-sites fill only after *a*-sites fill, that only half of the *a*-sites are

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 IIIIIIIIIIJIIIIIIIIIIIIIIIIII.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 IIIIJIIIIIJIIIIIIIIIIIIIIIIII.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 IIIIIIIIIIIJIIIIIIIIIIIIIIII.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 IIIIJIIIIIIJIIIIIIIIIIIIIIII.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 IIIIIIIJIIIJIIIIIIIIIIIIIIII.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 IIIIIIIIIIIJIIIIJIIIIIIIIII.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 IIIIIIIIIIIJIIIJIIIIIIIIIIII.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	
	AIC	213.109	143.076	
	Cpu	0.000	4.692	fit succeeded
8 IIIIIIIIIJIIIIIIIIIIIIIIIIIIII.2	R6X7	100.000^12	55.510^12	(52.808^12, 58.362^12)
	SSE	838893.670	8762.456	
	AIC	211.545	143.121	
	Cpu	0.000	3.092	fit succeeded
9 IIIIIIJIIIJIIIIIIIIIIIIIIIIII.115	R4X6	100.000^9	121.981^9	(95.160^9, 155.158^9)
	R6X8	100.000^13	62.668^13	(59.419^13, 66.175^13)
	SSE	411302.845	7166.368	
	AIC	204.035	143.287	
	Cpu	0.000	10.117	fit succeeded
10 IIIIIIIIIIJIIIIJIIIIIIIIIIII.172	R6X8	100.000^13	62.505^13	(59.419^13, 66.175^13)
	R1X2	100.000^2	2765.892^2	(992.275^2, 7707.892^2)
	SSE	794114.082	7173.010	
	AIC	213.904	143.300	
	Cpu	0.000	5.191	fit succeeded

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.

REFERENCES

1. Thelander, L.; Reichard, P. Reduction of ribonucleotides. *Annu. Rev. Biochem.* **1979**, 48, 133–158.
2. Kashlan, O.B.; Scott, C.P.; Lear, J.D.; Cooperman, B.S. 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. Available from http://cran.case.edu/
4. Rocks. Available from http://www.rocksclusters.org/wordpress/
5. Plot Digitizer. Available from http://plotdigitizer.sourceforge.net/
6. Radivoyevitch, T. Equilibrium model selection: dTTP induced R1 dimerization. *BMC Systems Biology* **2008**, 2, 15.
7. Burnham K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach.* Springer-Verlag, New York, 2002.