BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 14 2014, pages 2073–2075
doi:10.1093/bioinformatics/btu154

Sequence analysis

Advance Access publication March 28, 2014

CorMut: an R/Bioconductor package for computing correlated mutations based on selection pressure

Zhenpeng Li[1], Yang Huang[1], Yabo Ouyang[1], Yang Jiao[1], Hui Xing[1], Lingjie Liao[1], Shibo Jiang[2], Yiming Shao[1,*] and Liying Ma[1,*]

[1]State Key Laboratory for Infectious Disease Prevention and Control, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Beijing 102206 and [2]Key Laboratory of Medical Molecular Virology (Ministries of Education and Health), Shanghai Medical College and Institute of Medical Microbiology, Fudan University, Shanghai 200032, China

Associate Editor: John Hancock

ABSTRACT

Summary: Correlated mutations constitute a fundamental idea in evolutionary biology, and understanding correlated mutations will, in turn, facilitate an understanding of the genetic mechanisms governing evolution. CorMut is an R package designed to compute correlated mutations in the unit of codon or amino acid mutation. Three classical methods were incorporated, and the computation results can be represented as correlation mutation networks. CorMut also enables the comparison of correlated mutations between two different evolutionary conditions.

Availability and implementation: CorMut is released under the GNU General Public License within bioconductor project, and freely available at http://bioconductor.org/packages/release/bioc/html/CorMut.html.

Contact: mal@chinaaids.cn or yshao08@gmail.com

Received on January 22, 2014; revised on March 7, 2014; accepted on March 13, 2014

1 INTRODUCTION

In genetics, the Ka/Ks ratio (or ω, dN/dS) is the ratio of the number of non-synonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) (Hurst, 2002), which can be used as an indicator of selective pressure acting on a protein-coding gene. The selection pressure may reflect a change in the function of a gene or a change in environmental conditions that forces the organism to adapt. For example, the mutations to confer resistance of a highly variable virus, e.g. HIV or HCV, to new antiviral drugs may be expected to undergo positive selection in a patient population treated with the drugs against the corresponding virus.

The concept of correlated mutations is a fundamental idea in evolutionary biology. The amino acid substitution rates are expected to be limited by functional constraints (Tourasse and Li, 2000). Given the functional constraints operating on genes, a mutation in one position can be compensated by an additional mutation. As a result, mutation patterns can be formed by correlated mutations responsible for specific conditions.

Here, we developed an R/Bioconductor package to detect the correlated mutations among positive selection sites by combining Ka/Ks ratio and correlated mutations analysis. CorMut is suitable for computing the correlated mutations of highly variable genomes, such as HIV and HCV.

2 IMPLEMENTATION

CorMut provides functions for computing Ka/Ks for individual sites or specific amino acids and detecting correlated mutations among them. The computation of Ka/Ks for an individual site or specific amino was based on the model of Chen *et al.* (2004). Traditionally, Ka/Ks computation has been based on sequence alignment, but this model was revised to compute the selection pressure for individual sites or specific amino mutation. The Ka/Ks of a specific amino acid substitution (X2Y) for a codon is computed as follows:

$$\frac{Ka}{Ks}=\frac{\frac{N_Y}{N_S}}{\frac{n_{Y,t}f_t+n_{Y,v}f_v}{n_{S,t}f_t+n_{S,v}f_v}}$$

where N_Y and N_S are the count of X→Y mutations at that codon and the count of synonymous mutations observed at that codon, respectively. Then, N_Y/N_S is normalized by the ratio expected under a random mutation model where selection pressure is absent, which is represented as the denominator of the formula. In the random mutation model, f_t and f_v indicate the transition and transversion frequencies, respectively. They are measured from the entire dataset according to the following formulas: $f_t = N_t/n_tS$ and $f_v = N_v/n_vS$, where S is the total number of samples, N_t and N_v are the numbers of observed transition and transversion mutations, respectively, and n_t and n_v are the number of possible transitions and transversions in the focused region (simply equal to its length L and 2L, respectively).

CorMut incorporates three classical methods to detect correlated mutations, including conditional selection pressure, mutual information and Jaccard index. Conditional selection pressure, also known as conditional Ka/Ks, indicates the Ka/Ks of one mutation (X) in the presence of non-synonymous mutations (Y). It can be used to evaluate the influence of X on Y. Mutual information can be used to measure the mutual dependence of two

*To whom correspondence should be addressed.

random variables, making it possible to measure the correlated mutations between two positions (Cover and Thomas, 2012). Jaccard index measures the similarity between two variables, and it has been widely used to measure correlated mutations (Myers and Pillay, 2008; Reuman *et al.*, 2010; Rhee *et al.*, 2007).

The computation for correlated mutations consists of two steps: First, the positive selection sites are detected using the selection pressure-based method. Second, the mutation correlations are computed among the positive selection sites using the three methods described above. It should, however, be noted that CorMut can also be used to compute correlated mutations without considering selection pressure at all. Meanwhile, CorMut facilitates the comparison of the correlated mutations between two conditions by means of a correlated mutation network.

3 FUNCTIONS AND SAMPLES

CorMut enables the processing of multiple sequence alignment files. Because sequences may have raw bases, the *seqFormat* function will replace the raw bases with common bases and delete the gaps according to the reference sequence. As one raw base has several common bases, a base causing an amino acid mutation will be randomly chosen. Once this task is completed, CorMut provides various functions to compute the correlated mutations. For the conditional Ka/Ks and mutual information methods, functions are provided to compute correlated mutations in the codon and amino units. However, for Jaccard index method, the computation is only provided for the amino unit.

CorMut provides a *biCompare* function to compare the correlated mutations between two conditions by means of a correlated mutation network. The *biCompare* results can be visualized by a plot method. Here, 300 HIV protease sequences of treatment-naive and treatment groups derived from an HIV drug-resistant database will be used as examples.

```
>library (CorMut)
>examplefile   =   system.file("extdata",
"PI_treatment.aln", package = "CorMut")
>examplefile02  =  system.file  ("extdata",
"PI_treatment_naive.aln",
package="CorMut")
>example = seqFormat(examplefile)
>example02 = seqFormat(examplefile02)
>biexample = biCkaksAA(example02,example)
>result = biCompare(biexample)
>plot(result)
```

As shown in Figure 1, only positive selection amino acid mutations in treatment-naive or treatment groups were displayed. Square nodes indicate the distinct amino acid mutations of the treatment-naïve group, which means that these nodes will be non-positive selection in the treatment group. Circle nodes indicate that distinct positive amino acid mutations appeared in the treatment group. The plot function also has an option for displaying the unchanged positive selection nodes in both conditions. If plot Unchanged is FALSE, then the unchanged positive selection nodes in both conditions will not be displayed.

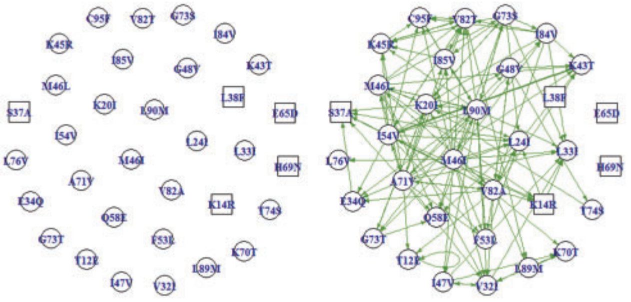

Fig. 1. The network represents correlated mutations between treatment-naive and treatment groups. Square nodes indicate the distinct amino acid mutations of treatment-naive group, while circle nodes indicate the distinct positive amino acid mutations that appeared in the treatment group. The edge between two nodes indicates the significant correlation between two mutations

Many software tools are available for correlated mutation analysis, such as CMAT (Jeong and Kim, 2012), HelixCorr (Fuchs *et al.*, 2007) and OMES-KASS (Kass and Horovitz, 2002), but each of these tools has a singular focus. For example, HelixCorr was specifically designed for the transmembrane parts of membrane proteins. In contrast, CorMut is multifunctional and can be applied to highly variable genomes. One particular advantage of CorMut is pre-filtering mutations based on selection pressure methods, which enables user to focus on the mutations of evolutionary interest. CorMut also displays the correlated mutations in network view and intuitively compares the correlated mutations between two conditions by such correlated mutation network.

4 CONCLUSION

The CorMut package provides simple and flexible functionality to compute and display the correlated mutations between codons or amino mutations. The representation of correlation mutations as a network and the ability to compare correlated mutations between two conditions make it possible to understand the genetic mechanisms governing evolution.

Funding: The State Key Laboratory for Infectious Disease Prevention and Control (grant number 2011SKLID102); the National Nature Science Foundation of China (grant numbers 30872232 and 81172733).

Conflict of Interest: none declared.

REFERENCES

Chen,L. *et al.* (2004) Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.*, **78**, 3722–3732.

Cover,T.M. and Thomas,J.A. (2012) *Elements of Information Theory*. John Wiley & Sons, United States.

Fuchs,A. *et al.* (2007) Co-evolving residues in membrane proteins. *Bioinformatics*, **23**, 3312–3319.

Hurst,L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.*, **18**, 486.

Jeong,C.S. and Kim,D. (2012) Reliable and robust detection of coevolving protein residues. *Protein Eng. Des. Sel.*, **25**, 705–713.

Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.

Myers,R.E. and Pillay,D. (2008) Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J. Virol.*, **82**, 9228–9235.

Reuman,E.C. *et al.* (2010) Constrained patterns of covariation and clustering of HIV-1 non-nucleoside reverse transcriptase inhibitor resistance mutations. *J. Antimicrob. Chemother.*, **65**, 1477–1485.

Rhee,S.Y. *et al.* (2007) HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput. Biol.*, **3**, e87.

Tourasse,N.J. and Li,W.H. (2000) Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.*, **17**, 656–664.