Bioinformatics, 31(19), 2015, 3213–3215
doi: 10.1093/bioinformatics/btv326
Advance Access Publication Date: 25 May 2015
Applications Note

Sequence analysis

IgSimulator: a versatile immunosequencing simulator

Yana Safonova[1,2,*], Alla Lapidus[1,2] and Jennie Lill[3]

[1]Center of Algorithmic biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, Russia, [2]Algorithmic Biology Laboratory, St. Petersburg Academic University, St. Petersburg, Russia and [3]Department of Protein Chemistry, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 11, 2014; revised on May 18, 2015; accepted on May 19, 2015

Abstract

Motivation: The recent introduction of next-generation sequencing technologies to antibody studies have resulted in a growing number of immunoinformatics tools for antibody repertoire analysis. However, benchmarking these newly emerging tools remains problematic since the gold standard datasets that are needed to validate these tools are typically not available.

Results: Since simulating antibody repertoires is often the only feasible way to benchmark new immunoinformatics tools, we developed the IgSimulator tool that addresses various complications in generating realistic antibody repertoires. IgSimulator's code has modular structure and can be easily adapted to new requirements to simulation.

Availability and implementation: IgSimulator is open source and freely available as a C++ and Python program running on all Unix-compatible platforms. The source code is available from yana-safonova.github.io/ig_simulator.

Contact: safonova.yana@gmail.com

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

1 Introduction

The recent introduction of next-generation sequencing (NGS) technologies to antibody studies have resulted in a growing number of immunoinformatics tools for antibody repertoire analysis (Boyd *et al.*, 2009; Weinstein *et al.*, 2009; Shugay *et al.*, 2014). However, benchmarking these newly emerging tools remains problematic since the gold standard datasets that are needed to validate these tools are typically not available. Since simulating antibody repertoires is often the only feasible way to benchmark new immunoinformatics tools, many immunoinformatics groups have developed their own antibody repertoire simulators of various complexity (Bonissone and Pevzner, 2015; Weinstein *et al.*, 2009). Lack of a publicly available immunoinformatics simulator makes it difficult to benchmark multiple antibody analysis tools and forces every new immunoinformatics researcher to invest efforts into development of a yet another simulator. We argue that the time has come to develop a comprehensive publicly available immunoinformatics simulator, similar to the widely used ART (Huang *et al.*, 2012) read simulator for NGS studies. Developing an immunoinformatics simulator is a more complex task than developing ART since it should reflect a complex process of forming antibody repertoires in a realistic statistical setting. These complications are further amplified by the fact that similar but distinct antibodies within a repertoire often differ from each other by only a small number of mutations, thus, making it difficult to distinguish them from sequencing errors. Moreover, antibody repertoires often feature a few large clusters (occurring as a result of *in vivo* clonal selection) accompanied by a large number of smaller clusters, presenting a challenge for the repertoire analysis algorithms that attempt to separate small clusters from often similar large clusters. Below we describe the IgSimulator tool that addresses various complications in generating realistic antibody repertoires. IgSimulator simulates repertoires and uses ART to simulate reads resulting from this repertoire.

2 Methods

2.1 Simulating individual antibodies

Antibody repertoires are generated by complex processes: V(D)J recombination, intergenic insertions and somatic hypermutations. IgSimulator models all these processes and generates repertoires based on the following features (Murphy, 2012). Distribution of the clone frequencies contains few overrepresented clones and plenty of low abundant clones. The lengths of exonuclease removals do not exceed 10 nt and are distributed uniformly. The lengths of N nucleotides, or non-genomic intergenic insertions, do not exceed 10 nt and are distributed uniformly. The lengths of P nucleotides, or inverted palindromic intergenic insertions, do not exceed 4 nt and are distributed uniformly. Frequencies of SHM in CDRs (per nucleotide) are higher than frequencies of SHM in FRs. SHMs are generated as a mix of motif based (Rogozin and Kolchanov, 1992) and random changes.

2.2 Simulating antibody repertoires

The repertoire simulation proceeds in the following five steps illustrated in Figure 1. IgSimulator generates a set of the *base antibody sequences* using simulation of the V(D)J recombination mechanisms, somatic deletions of start and end of the gene segments and insertions of the non-genomic P and N nucleotides (step 1 in Fig. 1). By default, IgSimulator uses the IMGT database of human Ig germline genes (Brochet *et al.*, 2008; user can provide different set of germline genes if necessary) and selects among equally likely V/D/J gene segments candidates to simulate the V(D)J recombination. IgSimulator randomly assigns the frequency to each base antibody sequence using the power law distribution (step 2 in Fig. 1). Evidence of this distribution was shown recently (Weinstein *et al.*, 2009). IgSimulator introduces the somatic mutations into each base antibody sequences resulting in *mutated antibody sequences* (step 3 in Fig. 1). To generate an antibody repertoire, IgSimulator randomly assigns the frequency to each mutated antibody sequence using the power law distribution (step 4 in Fig. 1). The frequencies represent the number of times each mutated antibody sequence will be present in the antibody repertoire. The resulting antibody repertoire is further subjected to an NGS read simulation using ART (step 5 on Fig. 1). To configure the diversity rate, IgSimulator uses the following parameters: # *base sequences*, # *mutated sequences* and *repertoire size*. A small difference between # *mutated sequences* and # *base sequences* leads to simulation of low abundant families of mutated antibodies. In contrast, big difference between # *mutated sequences* and # *base sequences* leads to simulation of highly repetitive repertoires that include large mutated groups (see Supplementary).

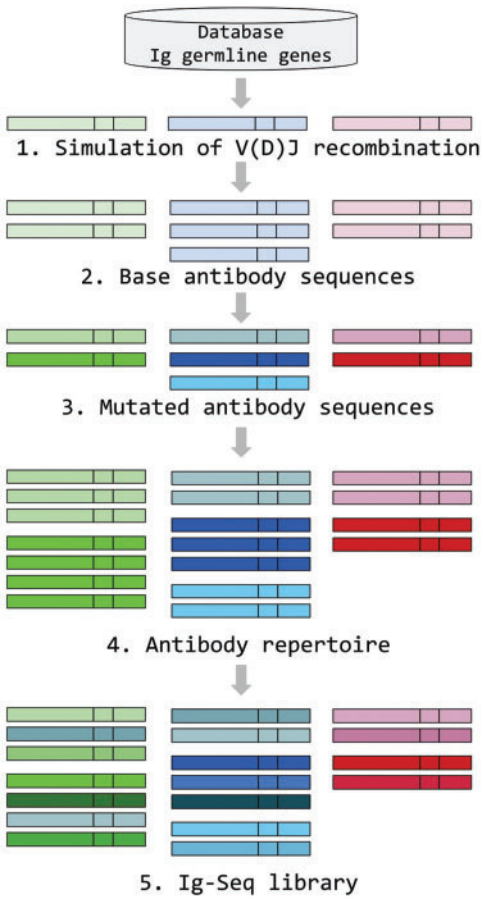

Fig. 1. IgSimulator pipeline for simulating an antibody repertoire

2.3 Simulating Ig-seq library

IgSimulator uses a simulated repertoire as a reference and runs ART to simulate an Ig-seq library with reads featuring realistic and technology-specific error profiles. The user has an option to generate reads with errors that follow the specifics of Illumina. IgSimulator selects parameters of ART in such a way that the following conditions hold: (i) reads (or read-pairs) in the simulated library cover the variable region of the antibody (in the case of Illumina reads, IgSimulator generates paired-end overlapping reads with insert size similar to the length of the variable region), and (ii) the coverage of antibody sequences is uniform. As a result, IgSimulator creates an idealized repertoire for simulated library that can be used for further repertoire studies and benchmarking various repertoire analysis tools.

3 Conclusions

Many immunoinformatics algorithms have been developed for the challenging problem of antibody repertoire analysis. Our antibody repertoire simulator IgSimulator will help scientists to assess various immunoinformatics software tools and to choose the best pipeline for their research. The initial version of IgSimulator addresses most but not all aspects of modeling the antibody repertoires. In the future, it will be extended to address the following issues: (i) a data-driven modeling of the distribution of mutation across CDRs and FRs based on analyzing multiple immunoinformatics datasets, (ii) probabilistic model for modeling SHM motifs (Bonissone and Pevzner, 2015), (iii) statistical inference methods for generating repertoires (Murugan *et al.*, 2012), (iv) addition of insertions and deletions to the set of SHMs (Klien *et al.*, 1998) and (v) biases in N addition such as the GC preference (Briney *et al.*, 2012).

Acknowledgements

The authors thank Dr. Pavel A. Pevzner and SPAdes team (Bankevich *et al.*, 2012) for productive collaboration, helpful comments and feedback on using our software.

Funding

Y.S. and A.L. are supported by Russian Science Foundation (grant No 14-50-00069).

Conflict of Interest: none declared.

References

Bankevich,A. *et al*. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol*., **19**, 455–477.

Bonissone,S. and Pevzner,P.A. (2015) Immunoglobulin classification using the colored antibody graph. In: Przytycka,T.M. (ed.), *Research in Computational Molecular Biology (RECOMB), Lecture Notes in Computer Science*, **Vol. 9029**, pp. 44–59.

Boyd,S. *et al*. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med*., **1**, 12–23.

Briney,B.S. *et al*. (2012) Frequency and genetic characterization of v(dd)j recombinants in the human peripheral blood antibody repertoire. *Immunology*, **137**, 56–64.

Brochet,X. *et al*. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*., **36**, W503–W508.

Huang,W. *et al*. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Klien,U. *et al*. (1998) Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev*., **162**, 261–280.

Murphy,K.P. (2012) *Janeway's immunobiology*. 8th edn. Garland Science, London.

Murugan,A. *et al*. (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. USA*, **109**, 161616.

Rogozin,I. and Kolchanov,N. (1992) Somatic hypermutagenesis in immunoglobulin genes. II. influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta*, **1171**, 11–18.

Shugay,M. *et al*. (2014) Towards error-free profiling of immune repertoires. *Nat. Methods*, **11**, 653–655.

Weinstein,J. *et al*. (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807–810.