Sequence analysis

SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements

Supriya Munshaw\(^1,2\) and Thomas B. Kepler\(^1,3,\ast\)
\(^1\)Center for Computational Immunology, \(^2\)Computational Biology and Bioinformatics Program, Duke University, P.O. Box 90090 and \(^3\)Department of Biostatistics and Bioinformatics, 2424 Erwin Road, Suite 1103, Durham, NC 27705, USA

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: The inference of pre-mutation immunoglobulin (Ig) rearrangements is essential in the study of the antibody repertoire produced in response to infection, in B-cell neoplasms and in autoimmune disease. Often, there are several rearrangements that are nearly equivalent as candidates for a given Ig gene, but have different consequences in an analysis. Our aim in this article is to develop a probabilistic model of the rearrangement process and a Bayesian method for estimating posterior probabilities for the comparison of multiple plausible rearrangements.

Results: We have developed SoDA2, which is based on a Hidden Markov Model and used to compute the posterior probabilities of candidate rearrangements and to find those with the highest values among them. We validated the software on a set of simulated data, a set of clonally related sequences, and a group of randomly selected Ig heavy chains from Genbank. In most tests, SoDA2 performed better than other available software for the task. Furthermore, the output format has been redesigned, in part, to facilitate comparison of multiple solutions.

Availability: SoDA2 is available online at https://hippocrates.duhs.duke.edu/soda. Simulated sequences are available upon request.

Contact: kepler@duke.edu

Received on September 15, 2009; revised on February 3, 2010; accepted on February 4, 2010

1 INTRODUCTION

B cells express immunoglobulin (Ig) molecules on their outer surface and secrete them into the extracellular space. Secreted Ig is known as antibody. The genes that encode for antibodies are generated by many diversifying mechanisms including combinatorial rearrangement of gene segments, addition of non-templated (n) nucleotides at the junctions, and somatic hypermutation. This circumstance presents the important challenge of inferring the components of the original rearrangement for any observed Ig gene. Because point mutations cause loss of information regarding the original rearrangement, there may be multiple plausible solutions. In this article, we present a Bayesian statistical method based on a Hidden Markov Model (HMM) that allows a complete statistical treatment of the problem by providing the posterior probability of each possible rearrangement.

Antibodies serve as effector molecules that neutralize microbes by binding to exposed antigens and targeting them to other components of the immune system, such as phagocytic cells and complement, that effect clearance. Ig genes generate diversity in two stages: an antigen-independent and an antigen-dependent stage. Antigen-independent diversity is generated in the bone marrow, where B cells originate, by combinatorial rearrangement of gene segments and junctional diversity. Combinatorial diversity is created in a number of ways. First, each antibody molecule comprises one heavy- and one light-chain protein. Both the light- and heavy-chain genes are encoded by gene segments that are genetically rearranged during a process known as V(D)J recombination (Sakano et al., 1980; Tonegawa, 1983). Heavy chains are made up of three gene segments—variable (VH), diversity (DH) and joining (JH) where as light chains only have a V and J segment. In humans, there are ~50 known functional VH segments, 27 known functional DH segments, and six known functional JH segments (LeFranc, 2001). This arrangement allows for ~8100 combinations in the heavy chain alone. Humans also have two light-chain loci, \(\kappa\) and \(\lambda\) (Frippiat et al., 1995). Only one of these loci is expressed per cell so that each antibody either has a \(\kappa\) light chain or a \(\lambda\) light chain. Humans have 44 functional \(\kappa\) genes, 5 \(\lambda\) genes, 33 \(\kappa\) V genes and 5 \(\lambda\) V genes (LeFranc, 2001) resulting in 220 possible \(\kappa\) chains and 165 possible \(\lambda\) chains. Thus this combinatorial rearrangement alone allows for greater than 3 million antibodies. Junctional diversity is the result of multiple recombination site choices for each recombination event and the addition of n nucleotides. n nucleotides are sometimes added at the junction by terminal deoxynucleotidyl transferase (TdT) between adjoining gene segments (Desiderio et al., 1984). Although TdT is believed to be expressed only in pro-B cells, the stage in which the heavy-chain rearrangement takes place (Desiderio et al., 1984), the presence of n nucleotides in light chains has also been seen in a few studies (Bridges, 1998). Antigen-dependent diversity is generated by somatic hypermutation in the periphery in a manner dependent on activation-induced cytidine deaminase (AID); during this process, mutations in the Ig genes are accumulated at rate of up to 10\(^6\) times the normal background rate (Muramatsu et al., 2000). B cells are subsequently selected for enhanced affinity for the eliciting antigen. It is estimated that these processes of diversification can generate \(~10^{12}\) different antibodies making it challenging to correctly identify the underlying
The inference of the recombination and mutation events that produced a given Ig gene is of great importance in the study of humoral immunity and has been tackled in many different ways. The goal of such inference is to identify each of the component gene segments used as well as the recombination sites, point mutations and n nucleotides. The aligned gene segments usually overlap, which is why alignments of the target gene to the individual gene segments cannot be treated as independent. Somatic mutations, n nucleotide addition and recombination site choice make this task more challenging. The short length of the DH gene segment makes it especially difficult to identify the CDR3 region of the heavy chain, which is the most diverse region in the antibody sequence. This leads to many possible gene segment combinations that can result in a given antibody gene. Hence, it is necessary to report all such rearrangements and assign a probability to each of the combinations, making it easy to compare all possible rearrangements. Several algorithms have been developed for inferring Ig gene segment composition. IMGT/V-QUEST is one of the first and most complete of these tools and has the ability to analyze both Ig and TCR sequences for a variety of organisms including human and mouse (Guidicelli et al., 2004). V-QUEST, however, is based on the BLAST algorithm; it does not guarantee finding the best alignment of two sequences (Altschul et al., 1990). Additionally, the implementation of the algorithm only allows for running a maximum of 50 sequences at a time. Another tool, JOINSOLVER, is based on the identification of conserved motifs in the target gene (Souto-Carneiro et al., 2004). Both JOINSOLVER and V-QUEST provide multiple gene segment possibilities but the implementation only provides junction analysis for the topmost choice. Somatic diversification analysis (SoDA) (Volpe et al., 2006) uses a 3D alignment algorithm that allows for insertions and deletions. The algorithm uses dynamic programming and is an extension of the Smith–Waterman local alignment algorithm (Smith and Waterman, 1981). The 3D alignment allows for a continuous alignment through all the states of the recombination. SoDA infers only a single highest-scoring alignment, and ignores other solutions that may have equal or nearly equal scores. SoDA's guarantee of optimality in the inferred rearrangement is obtained at the cost of computational effort; SoDA takes more CPU time than either JOINSOLVER or V-QUEST. A major shortcoming for all the programs above is that they do not provide a meaningful comparison of the different possible rearrangements. iHMMune-align (Gaeta et al., 2007) partially solves the problem and provides a probabilistic model using an HMM to infer the rearrangement. iHMMune-align uses the Viterbi algorithm (Rabine, 1989) to find the most probable path through the alignment matrix, but does not sum over paths or provide results on suboptimal alignments. This choice becomes an issue when selecting an appropriate DH gene segment for Ig heavy chains. The DH gene is the shortest of all gene segments, and is typically the most difficult to align. We have found Ig genes that present an equally good alignment with different DH genes (see Results section and Figure 5). iHMMune-align or SoDA gives only the solution with the highest score even if the highest score is not significantly better than the second highest score and so on (iHMMune-align does provide the option of viewing the top 10 VH gene alignments, but not DH). Among these four methods, only SoDA allows for gaps when performing alignments, although insertions and deletions are known to occur at non-negligible frequencies during somatic hypermutation (Smith et al., 1996), and alignment without gaps when gaps are present leads to dramatically erroneous inferences. The method we are introducing here is an update of SoDA— we call it SoDA2. It employs a probability mass function-based scoring to create a local-aligning gene segments and a probabilistic HMM for the inference of CDR3. The system calculates the posterior probability over all paths using a particular set of gene segments by the forward and backward algorithms. It then provides the alignment path with the highest posterior probability. If the sequence does not hold enough information to unambiguously select a gene segment, SoDA2 reports all alignments that do not differ significantly. We tested this method using a simulated dataset constructed from the statistics of observed rearrangements and compared these results with those obtained using existing methods. We also used two natural datasets, a set of clonally related Ig genes and a random set of sequences from NCBI. Each test indicates that SoDA2 provides the most thorough and accurate results among all programs in addition to providing the most statistically complete results.

2 METHODS

2.1 Determining the type of Ig

The first step consists of aligning the target sequence with a consensus-like sequence of the VH, Vκ and Vλ families to determine if the input sequence is a heavy, kappa or lambda chain. These consensus sequences are pre-created by separate alignments of the VH, Vκ and Vλ segments. We use the AHO numbering scheme (Hotezegger and Plückthun, 2001) which is based on the spatial alignment of known 3D structures of Ig domains. The gaps are placed to minimize the average deviation from the averaged structure of the aligned domain so that the position of the CDRs remains consistent. The consensus is represented by a probability mass function (pmf), a V × 5 matrix where V is the length of the V genes in this case (Kepler et al., submitted for publication). For each nucleotide position in the gene, we determine the frequency of use for each nucleotide state (including ‘gap’) at that position in the family. For the target antibody, we create a similar pmf using the quality scores of the input sequence. The quality score is proportional to the log probability of the estimated sequencing error and is provided by the user’s base-calling software (Ewing and Green, 1998). If quality scores are not provided, we treat the input sequence as well-determined and all mismatches as due to somatic mutation. The pmf at each position of the target sequence depends on the quality score, which varies at each position, and a mutation frequency µ which is assumed to be constant over positions. For each position, we then have the probability of observing the five bases (including a gap) at that position given the quality score and the mutation frequency. We use the pmf of the target antibody gene and the pmf of the VH, Vκ and Vλ sequences as the input alignment (Kepler et al., submitted for publication, Smith and Waterman, 1981).

2.2 V and J gene pre-alignment

Assume, for example, that our target sequence has been determined to be a heavy chain. We use the trace-back path generated by aligning the pmf of VH to our target and obtain the likelihood for each member of the VH family. The mutation frequency µ is recalculated after observing mismatches in the highest scoring alignment. All VH sequences with equally high likelihood alignments are then submitted to the HMM. The position of the invariant cysteine is determined.

The target sequence is then aligned past the invariant cysteine with all the appropriate JH segments, using the pmf-based alignment mentioned above. The JHs with the highest likelihood are selected for submission to the HMM. The target sequence is further trimmed at the invariant...
A functional Ig chain must have no stop codons and the invariant cysteine

We implemented a pair HMM with 10 non-silent states—match/mismatch

Transition probabilities between states are
determined based on empirical
data (Basu et al., 1983). Transition probabilities between states are
determined by fitting a negative binomial distribution to the recombination
site choice for VH, DH and JH and number of n nucleotides in the junctions
as determined in a set of 293 unmutated sequences rearranged sequences
(Fig. 2. Jackson et al., 2004). Figure 3 shows a detailed implementation of
the HMM with transition probabilities.

2.4 Algorithm
Once we have the appropriately trimmed target and germline sequences, we
calculate the log of the total probability of a proposed rearrangement using the
forward and backward algorithms (Durbin et al., 1998; Majoras, 2007). We
select the gene segments that lead to the highest posterior probabilities, and perform a posterior Viterbi algorithm with traceback (Fariselli et al., 2005)
to select the path with the highest posterior probability for each possible
rearrangement. We report the probability of the most probable path for each
of the equally probable gene segment sets. For a heavy chain, a DH gene
alignment of ≥3 n is flagged as ‘unreliable D alignment’. The functionality
of an antibody gene is determined as follows and reported with the results.
A functional Ig chain must have no stop codons and the invariant cysteine

3 RESULTS
3.1 Simulated datasets
We created simulated datasets of 100 sequences each with
mutation frequencies of 2.5, 5, 10 and 20%. Recombination site
choice and number of n nucleotides for these simulations were
drawn from a negative binomial distribution. To avoid any bias
forwards our HMM, the parameters for these simulations were
estimated using a set of 662 sequences obtained from Genbank.
Furthermore, rearrangements for these sequences were determined
Table 1. Number of correct rearrangements (correct gene and allele for VH, DH and JH segments) identified by each software out of 100 sequences tested at each mutation rate

<table>
<thead>
<tr>
<th>Software</th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoDA2</td>
<td>73</td>
<td>65</td>
<td>47</td>
<td>28</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>99</td>
<td>94</td>
<td>87</td>
<td>58</td>
</tr>
<tr>
<td>JOINSOLVER</td>
<td>90</td>
<td>98</td>
<td>85</td>
<td>72</td>
</tr>
<tr>
<td>SoDA2</td>
<td>76</td>
<td>92</td>
<td>49</td>
<td>27</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>87</td>
<td>90</td>
<td>82</td>
<td>52</td>
</tr>
<tr>
<td>SoDA2</td>
<td>79</td>
<td>92</td>
<td>77</td>
<td>68</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>87</td>
<td>90</td>
<td>86</td>
<td>42</td>
</tr>
<tr>
<td>SoDA2</td>
<td>79</td>
<td>92</td>
<td>77</td>
<td>68</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>87</td>
<td>90</td>
<td>86</td>
<td>42</td>
</tr>
</tbody>
</table>

Table 2. Number of correct rearrangements (correct gene and allele for VH, DH and JH genes) identified by each software out of 100 sequences tested at each mutation rate for the simulated sequences

<table>
<thead>
<tr>
<th>Software</th>
<th>2.5%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoDA2</td>
<td>97</td>
<td>95</td>
<td>94</td>
<td>87</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>99</td>
<td>94</td>
<td>97</td>
<td>93</td>
</tr>
<tr>
<td>JOINSOLVER</td>
<td>76</td>
<td>42</td>
<td>69</td>
<td>20</td>
</tr>
<tr>
<td>SoDA2</td>
<td>76</td>
<td>42</td>
<td>69</td>
<td>20</td>
</tr>
<tr>
<td>iHMMune-align</td>
<td>87</td>
<td>90</td>
<td>86</td>
<td>42</td>
</tr>
</tbody>
</table>

For sequences where SoDA2 failed to identify the correct rearrangement as the most probable one, we found a median difference of 0.67 in the natural log of the probability between the highest scoring rearrangement and the correct one at the 5% mutation rate. Thus, if allowed to include rearrangements with low differences (<1) in the natural log of the probability from the top scoring alignment, SoDA2 would have identified correct rearrangements for 22 additional sequences at the 5% mutation rate, yielding a possible 87% success rate.

3.2 Clonally related datasets

In order to test real biological data, we used two clonally related datasets that were used to test iHMMune-align (Gaeta et al., 2004) derived from tonsilar IgD class-switched B cells (Zheng et al., 2004). Because they are clonally related, sequences within a given set should have identical rearrangements and differ only by somatic mutation. We analyzed this dataset using VQuest, JOINSOLVER, SoDA and iHMMune-align to determine the number of times each of the programs resulted in the same rearrangement as was done by Gaeta et al. (2007). We ran the sequences through all the programs and found that iHMMune-align selected 47/57 identical rearrangements for the first group of sequences, while SoDA2 selected 34/57 identical rearrangements. IMGT/VQuest, JOINSOLVER and SoDA identified 37, 25 and 18 identical rearrangements, respectively. SoDA2 returned a minority DH gene segment in 17 cases, a minority JH allele in five cases, and a minority VH allele in four cases. In cases where SoDA2 failed to select the majority VH or JH gene segment, all the other programs, including iHMMune-align also failed to select the majority gene segment. It can be seen in these cases that mutation had obliterated the information necessary to make the correct inference. For the 17 cases where SoDA2 did not return the majority DH segment, the DH segment that was returned was typically judged more probable than the majority segment due to the balancing of n-nucleotide use and mutations. An example of this phenomenon is the inference for AF262199 (Fig. 4). In this case, the mutation frequency in the VH gene segment is ∼7%, SoDA2 selects IGHD1*01 and IGHD2*01 requiring two mutations (8.5% mutation frequency in CDR3) and seven n-nucleotides, while IGHD1*01 and IGHD2*01 requires two mutations (5.5% mutation frequency) and 10 n-nucleotides. For the second dataset of 99 sequences, both iHMMune-align and SoDA2 identified 68/99 identical rearrangements while IMGT/VQuest, JOINSOLVER and SoDA identified 56, 41 and 37 identical rearrangements, respectively.

3.3 Sequences from Genbank

We tested a set of 662 sequences collected from Genbank and previously used for testing iHMMune-align and SoDA (Genbank accession nos Z68345-487 and Z60363-770). Out of 662 sequences, 113 produced inferences on which all five programs agreed. There was no agreement from any of the programs on 140 sequences. This means that they either could not infer a rearrangement at all or they all differed in their inference. From the rest, SoDA2 agreed with the majority of the programs on 300 rearrangements (Table 3). These did not include those where SoDA and SoDA2 were the only
SoDA2: identification of immunoglobulin rearrangements

SoDA2 identifies immunoglobulin rearrangements more accurately than other programs. It performs particularly well on datasets where other programs disagree on the rearrangement. SoDA2 provides a statistically complete picture of the observed sequence’s origins.

4 CONCLUSION

The problem of inferring the correct rearrangement for antigen receptors is difficult due to the stochastic nature of the process, but the task is important for an increased understanding of the population somatic genetics of the immune response. In this paper we present a method based on an HMM that provides a statistical basis for identifying rearrangements of Ig genes. In addition to providing the posterior probability of the top rearrangement candidate, SoDA2 also provides the user with an option to see all rearrangements within a reasonable range of probabilities. This shows that allowing rearrangements within a reasonable range of probabilities in SoDA2 would give an accurate and thorough picture of the various rearrangements possible for a given Ig sequence. It is important to note that SoDA2 considers factors such as recombination site choices for each gene segment and numbers of n nucleotides at both junctions derived from empirical data in inferring rearrangements while alignment algorithms used by SoDA, VQuest and JOINSOLVER base their results on sequence similarity matrices which may not accurately represent the process of V(D)J recombination.

SoDA22: identification of immunoglobulin rearrangements

Table 3. Results from 662 sequences from Genbank, showing the performance of the five programs

<table>
<thead>
<tr>
<th>Number of rearrangements</th>
<th>All programs agree</th>
<th>All programs disagree</th>
<th>SoDA2 agrees with two or more programs</th>
<th>VQuest agrees with two or more programs</th>
<th>iHMMune-align agrees with two or more programs</th>
<th>JOINSOLVER agrees with two or more programs</th>
<th>SoDA agrees with two or more programs</th>
<th>SoDA2 agrees only with SoDA (no other programs agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All programs agree</td>
<td>113</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All programs disagree</td>
<td>140</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SoDA2 agrees with two or more programs</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQuest agrees with two or more programs</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iHMMune-align agrees with two or more programs</td>
<td>137</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOINSOLVER agrees with two or more programs</td>
<td>272</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SoDA agrees with two or more programs</td>
<td>244</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SoDA2 agrees only with SoDA (no other programs agree)</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If two or more programs displayed the same rearrangement (including the alleles), it was believed to be the majority rearrangement.

Fig. 4. (a) Top rearrangement as chosen by SoDA2 with a higher mutation frequency than the alternative, shown in (b). The different rearrangements represent a trade-off between mutation frequency and number of n nucleotides.

Fig. 5. The alignment of CDR3H of sequence by 1154693 using IGHD1-21*01 by (a) SoDA2, (b) IMGT/V-QUEST, JOINSOLVER and iHMMune-align, (c) SoDA. Rearrangements (b) and (c) were also provided by SoDA2 at a slightly lower probability.
method for the absolute comparison among these alternatives. This performance and thorough result reporting leads to a substantially longer computation time. SoDA2 takes ~15 s of real user time per set of V and J segment for a given heavy target sequence on a 64 bit machine with a 2.19 GHz processor and 4 GB RAM, but the investment of computational effort seems worthwhile.

ACKNOWLEDGEMENTS

Special thanks to William H. Majores, Institute for Genome Sciences and Policy, Duke University and Todd Wasson, Computational Biology and Bioinformatics Program, Duke University for very helpful discussions on implementing Hidden Markov Models.

Funding: We are grateful for the financial support of the Bill and Melinda Gates Foundation through grant number 38643 to Dr Barton F. Haynes.

Conflict of interest: none declared.

REFERENCES