iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences

Author:

Gaeta, Bruno; Malming, Harald R.; Jackson, Katherine J.L.; Bain, Michael E.; Wilson, Patrick; Collins, Andrew M.

Publication details:

Bioinformatics
v. 23
Chapter No. 13
pp. 1580-1587
1367-4803 (ISSN)

Publication Date:

2007

Publisher DOI:

http://dx.doi.org/10.1093/bioinformatics/btm147

Sequence Analysis

iHMMune-align: Hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences

Bruno A. Gaëta[1,2], Harald R. Malming[2], Katherine J. L. Jackson[1], Michael E. Bain[2], Patrick Wilson[3] and Andrew M. Collins[1,*]

[1]School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, NSW 2052 Australia

[2]School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052 Australia

[3]Molecular Immunogenetics Program, the Oklahoma Medical Research Foundation, Oklahoma City, OK 73104 USA

ABSTRACT

Motivation: Immunoglobulin heavy chain (IGH) genes in mature B lymphocytes are the result of recombination of IGHV, IGHD and IGHJ germline genes, followed by somatic mutation. The correct identification of the germline genes that make up a variable VH domain is essential to our understanding of the process of antibody diversity generation as well as to clinical investigations of some leukemias and lymphomas.

Results: We have developed iHMMune-align, an alignment program that uses a hidden Markov model (HMM) to model the processes involved in human immunoglobulin heavy chain gene rearrangement and maturation. The performance of iHMMune-align was compared to that of other immunoglobulin gene alignment utilities using both clonally-related and randomly selected IGH sequences. This evaluation suggests that iHMMune-align provides a more accurate identification of component germline genes than other currently available IGH gene characterisation programs.

Availability: iHMMune-align cross-platform Java executable is freely available to academic users and can be downloaded from http://www.cse.unsw.edu.au/~binftools/iHMMuneAlign.zip.

1 INTRODUCTION

Antibody production is critical to our defences against microbial invaders. In order to respond to the incredible diversity of foreign antigens, antibodies (immunoglobulins) must be produced with specificities of equal diversity. The diversity of the repertoire is created by recombination, for functional immunoglobulin genes are created by the joining of a number of genes. During the early development of each B cell, a functional heavy chain variable domain is created by the essentially random recombination of three germline genes (IGHV, IGHD, IGHJ) that are each selected from a separate set. There are between 38 and 45 functional IGHV genes per haploid genome (Lefranc, 2001; Li*,* *et al.*, 2002), 23 unique functional IGHD sequences (Lee*, et al.*, 2006) and 6 functional IGHJ sequences (Ravetch*, et al.*, 1981; Ruiz*, et al.*, 1999)

During the recombination process, the ends of the joining genes are trimmed by unknown exonucleases, and as many as ten nucleotides are frequently removed from the IGHD and IGHJ gene ends. Non-template encoded nucleotides (N nucleotides) are also added, between the recombining genes, by the enzyme terminal deoxynucleotidyl transferase (TdT) (Basu*, et al.*, 1983). This enzyme is biased towards the addition of guanine, but the process is an essentially random one that can result in the addition of as many as 25 nucleotides between the joining genes.

During an immune response, additional diversity is generated by the process of somatic mutation, which principally occurs within the germinal centres of the lymph nodes. During this process, antigen-selected B cells accumulate mutations in their immunoglobulin genes. Such mutations are high frequency events, occurring at an estimated rate of 10^{-3} mutations per base pair per generation or approximately one immunoglobulin gene mutation per B cell division. Clonal expansion, following antigen selection, therefore gives rise to a clone of diverse sequences whose antibodies are encoded by the same germline genes, but which have diverged from one another through the mutation process. In fact many sequences can be seen, after large clonal proliferations, which have accumulated thirty or more mutations within the 370 or so nucleotides that make up the rearranged V-D-J gene. Mutations are typically concentrated in the complementarity determining regions (CDRs) which encode the variable domain loops that contact the antigen.

Similarities between germline genes, the effects of exonuclease and TdT activity, and the somatic mutation process together can make it difficult to identify the germline genes within a rearranged V-D-J gene. This is particularly true in the case of IGHD genes, which range in length from just 11 nucleotides to 37 nucleotides. After processing of the IGHD gene ends by exonucleases, there may sometimes be very little remaining from the germline sequence. Nevertheless there are many reasons why researchers and clinicians attempt to identify the genes involved in the immunoglobulin rearrangements.

Many studies have reported biases in antibody gene usage in conditions including rheumatoid arthritis (Huang*, et al.*, 1998) and other autoimmune diseases (Dorner and Lipsky, 2001), as well as in many leukemias and lymphomas such as MALT lymphoma (Yoshida*, et al.*, 2006), multiple myeloma (Kosmas*, et al.*, 1999) and chronic lymphocytic leukaemia (Tobin*, et al.*, 2002; Widhopf*, et al.*, 2004). All such studies require the accurate identification of germline genes. The identification of somatic point mutations has been critical to the study of the mutation process (Neuberger*, et al.*, 2005; Zheng*, et al.*, 2005),

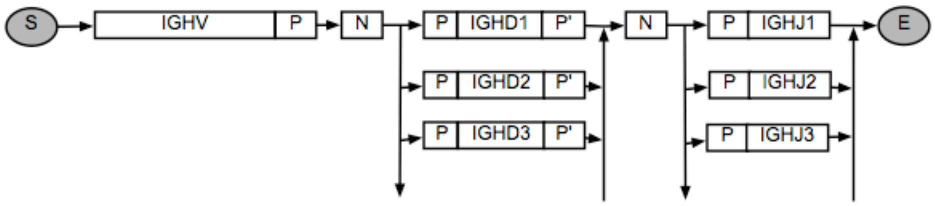

Fig 1: iHMMune-align HMM topology overview. The HMM models one IGHV gene and all the possible IGHD and IGHJ genes, together with junction states corresponding to N- and P- addition.

and this too begins with the alignment of mutated sequences against the germline gene repertoire. The analysis of mutations is also of clinical importance. For example, the enumeration of somatic point mutations is an important prognostic indicator in chronic lymphocytic leukemia (Damle*, et al.*, 1999; Hamblin*, et al.*, 1999). Finally, studies of the processes that generate immunoglobulin diversity - particularly nucleotide removals and additions – are impossible without reliable identification of germline genes and of the ends of the processed genes.

A number of different programs have been developed for aligning immunoglobulin gene sequences and identifying their germline components. IMGT/V-QUEST+JCTA (Bleakley*, et al.*, 2006) integrates IMGT/V-QUEST (Giudicelli*, et al.*, 2004) and IMGT/JunctionAnalysis (Yousfi Monod*, et al.*, 2004) and is based on dynamic programming sequence alignment with additional alignment steps at the gene junctions. JOINSOLVER (Souto-Carneiro*, et al.*, 2004) focuses on the third CDR of heavy chains (CDR3$_H$) which includes the IGHD gene and its junction with the IGHV and IGHJ genes. It uses specific sequence motif searches to delimit regions to be aligned (without gaps) to candidate germline genes. SoDA (Volpe*, et al.*, 2006) uses a variation on the dynamic programming sequence alignment algorithm that takes into account the variation around the IGHV-IGHD and IGHD-IGHJ junctions resulting from the competing effects of nucleotide addition and exonuclease action.

The mutation model underlying an alignment utility can have significant impact on its ability to identify the component germline genes of a rearranged sequence. Simple mutation models that do not take into account the sequence and location dependence of somatic mutation and other diversity generation processes often result in unlikely alignments with the 3' end of the variable region containing many more mismatches that would be expected based on the number of mutations observed at the 5' end. To address this issue, we have developed iHMMune-align, an application that incorporates explicit models of the various antibody generation processes in the form of probability distributions along a hidden Markov model (HMM) of the variable region of the heavy chain gene, and generates an alignment of the rearranged sequence with its most likely component germline genes. Its development was particularly designed to improve IGHD gene identification. IHMMune-align therefore presently focuses on alignment of heavy chain variable regions, though it can easily be adapted to the alignment of immunoglobulin light chains and T cell receptors. The current version is based on data gathered from human immunoglobulin gene sequence data and is therefore appropriate only for the alignment of human sequences, although the approach can also be readily adapted to other species provided sufficient data is available for these species

In this report, we provide a detailed description of iHMMune-align, and compare its performance against that of IMGT/V-QUEST+JCTA, JOINSOLVER and SoDA. It is difficult to evaluate the accuracy of immunoglobulin gene alignment software because it is almost always impossible to be certain of the germline genes that contributed to a sequence. Somatic point mutations of one sequence can arguably make it look like another sequence. For this reason, it is particularly difficult to be certain of IGHV alleles that might be used, and to distinguish between the short, highly similar IGHD genes. To test the performance of the various utilities, we have used sets of clonally-related sequences. These sets were generated from cDNA from tonsillar B cell, and each set therefore represents a clonal expansion from a single V-D-J rearrangement. Although there might be argument about the components of the original sequence, the performance of a program can be gauged by the proportion of sequences within each set for which the same IGHV, IGHD and IGHJ genes are identified. Further measures of the accuracy of the programs are provided by comparisons of alignments of an additional 662 cDNA-derived sequences. We conclude that iHMMune-align is the most accurate program for human heavy chain gene alignment, and as it is based upon a hidden Markov model, its performance can be expected to further improve as additional data are built into the model.

2 APPROACH

iHMMune-align proceeds through the following steps:

2.1 V-gene pre-alignment

The program starts with dynamic programming local alignment (Gotoh, 1990; Smith and Waterman, 1981) of the rearranged sequence with the human IGHV germline repertoire obtained from IMGT (Lefranc, 2005; Lefranc*, et al.*, 2005). This step allows the identification of the best-matching IGHV gene and also the estimation of the amount of somatic mutation based on the frequency of mismatches in the resulting alignment.

2.2 HMM construction

A HMM is built using the topology shown in figures 1 and 2. This HMM incorporates a chain of match states modelling the IGHV gene identified in step 2.1, and parallel chains of match states representing all the possible IGHD and IGHJ germline genes (Fig. 1). The IGHV, IGHD and IGHJ sections of the HMM are joined by match states representing nucleotides resulting from N- and P- addition, and delete states representing the effect of nucleotide removal at the junctions through exonuclease action (Fig 2).

The HMM initially represents rearranged but unmutated sequences, with emission probabilities for the IGHV, IGHD and IGHJ gene match states set at 1 for the nucleotide observed at this position in the germline gene, and 0 for the other 3 nucleotides. Emission probabilities for match states corresponding to nucleotides inserted by P-addition and transition probabilities at the IGHV-IGHD and IGHD-IGHJ junctions (which model exonuclease action in the V-D-J recombination process) are set based on frequencies for these events observed in a set of unmutated rearranged sequences (Jackson*, et al.*, 2004). Emission probabilities for match states corresponding to N-addition nucleotides are based on experimentally determined nucleotide addition propensities for terminal deoxyribonucleotide transferase (Basu*, et al.*, 1983).

2.3 Adjustment of emission probabilities

The emission probabilities in all match states of the model are then re-calculated to model the process of somatic mutation. The base probability of mutation is extrapolated from the mutation frequency observed in the initial IGHV- region pre-alignment, which provides an estimate of the number of rounds of mutation the sequence has undergone. This probability is then adjusted to take into account the position of the putative mutation, the local sequence context and the effect of antigen selection.

Sequence position: the probability of somatic hypermutation has been observed to decrease with distance from the 5' end of the rearranged gene, with a distribution fitting an exponential decay of the form $A_{N+i} = A_i.e^{-0.0024i}$ where A_N is the mutation propensity at position N (Rada and Milstein, 2001). iHMMune-align adjusts the probability of mutation along the sequence accordingly.

Local sequence context: iHMMune-align offers a choice of two models to represent the sequence-dependence of somatic hypermutation. The "Hotspot" model is based on the observation that sequence mutability occurs preferably at specific DNA motifs (RGYW, WRCY, WAN). iHMMune-align increases the probability of mutation at hotspots defined by these motifs, in proportion to the frequency of mutation at these motifs relative to other sequences (Martin and Scharff, 2002). When the user selects this model, iHMMune-align adjusts the probability of change away from the germline sequence by a factor of 32/6 for the second position of tetranucleotides fitting the RGYW consensus (and the third position of its reverse complement WRCY), based on the observation that these hotspots each contribute to 1/6 of observed mutations and each represent 1/32 of all possible tetranucleotides. Following the same reasoning, the probability of mutation for the central A of the WAN motif is adjusted by a factor of 8/3 and the probability of mutation of non-hotspot sequences is adjusted to take into account their lower than expected mutability.

The "Trinucleotide" model assigns a mutability score to each possible trinucleotide based on observed mutation frequencies after correction for position along the sequence (Collins*, et al.*,

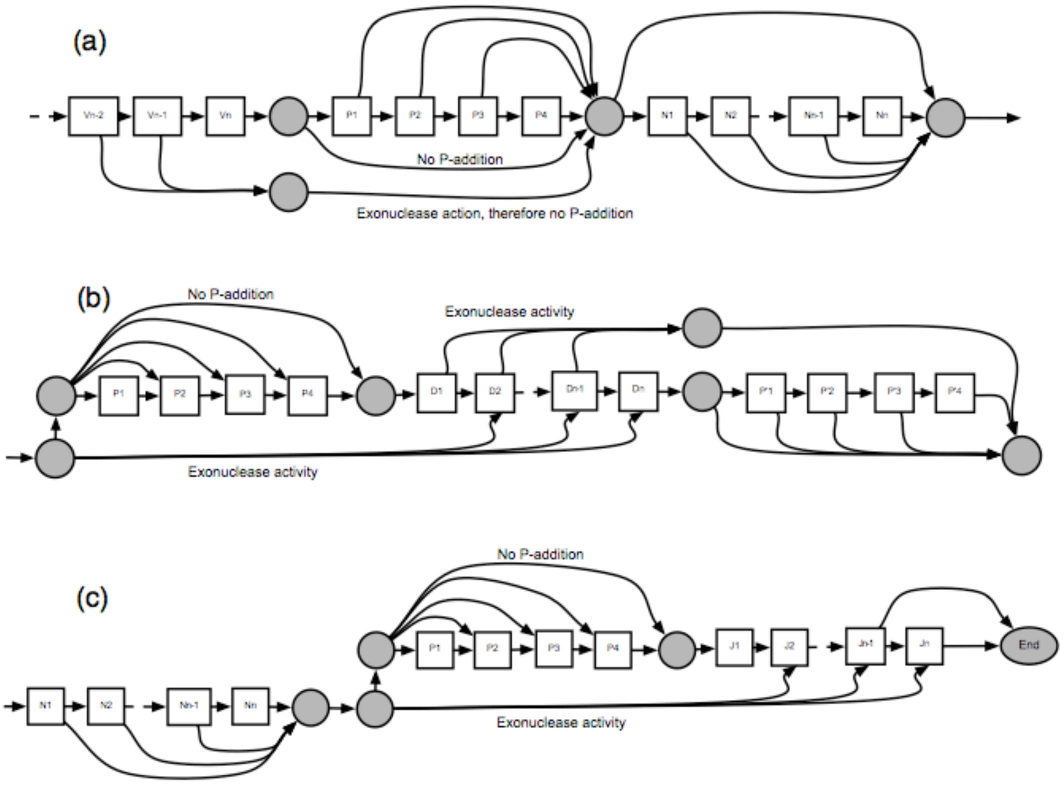

Fig 2: Detailed view of sections of the iHMMune-align HMM graph. (a) 3' end of IGHV region and adjacent N region (b) one IGHD gene model and adjacent P-addition states (c) IGHD-IGHJ N-region and one IGHJ gene model. Grey circles represent non-emitting states

2004) and adjusts emission probabilities along the model in proportion to this mutability score.

Antigen selection: over the course of the humoral immune response, B cells producing immunoglobulins with good affinity for antigen are encouraged to proliferate and those accumulating mutations that result in non-functional or less effective immunoglobulins are eliminated. As a result of this selection process, mutations in the CDRs that are involved in antigen binding are favoured over those in the framework regions that are responsible for the folding of the variable domain, and whose mutations are more likely to disrupt the overall structure of the antibody. To reflect this process, iHMMune-align adjusts the probability of mutation in the CDRs by a factor of 1.5, estimated from comparison of multiple immunoglobulin heavy chain gene sequences (Collins*, et al.*, 2004).

These factors are used to calculate the probability of mutation p_M at each match state of the HMM. Emission probabilities for each IGHV, IGHD and IGHJ gene state are set at 1- p_M for the germline sequence nucleotide and $p_M/3$ for each of the other 3 nucleotides.

2.4 Sequence alignment

The HMM is finally aligned with the rearranged, mutated sequence, using the Viterbi algorithm (Rabiner, 1989). The program outputs the alignment corresponding to the optimal path along the HMM and reports the germline genes corresponding to this optimal path. In cases where the IGHD gene has been heavily mutated or truncated by exonuclease action, its identification can be problematic and iHMMune-align reports the matching gene only when it meets an additional criterion based on number of consecutive perfect sequence matches in the IGHD gene alignment.

3 RESULTS

3.1 Implementation

iHMMune-align was implemented in the Java language, using the BioJava libraries (www.biojava.org), and has been successfully run under Microsoft Windows, MacOS X and Linux. The initial Smith-Waterman alignment is performed using the program Jaligner (Moustafa, 2006). iHMMune-align is accessed through a graphical user interface that allows selecting up to 50 sequences to align, and changing the IGHV, IGHD and IGHJ germline gene reference repertoires. A command line version is also available from the authors on request and has been used with Perl scripts to align batches of up to 10,000 sequences. The default values used to calculate the various transition probabilities at the IGHV-IGHD and IGHD-IGHJ junctions are based on frequencies observed in a large sequence set (Jackson*, et al.*, 2004) and are suitable in most cases. However these values can be modified by experts when aligning sequences that are known to have different characteristics – for example fetally-derived sequences that have been reported to undergo less N-nucleotide addition (Benedict*, et al.*, 2000). Users can also select between the commonly accepted Hotspot mutation model and the Trinucleotide model used routinely in our own analyses. Mutation model choice did not affect alignment in our evaluations but both models were included for expert users. Other user-modifiable parameters include the output format (HTML or spreadsheet-compatible) and the criterion for IGHD gene acceptance (5-mer or 8-mer). Using the 5-mer criterion, iHMMune-align requires 5 consecutive perfect matches to accept a D-gene identification. The more stringent 8-mer criterion used by default in our laboratory requires 8 or 9 consecutive matches in a row, or 10-11 matches with one mismatch, or 12 matches with 2 mismatches. These rules are based upon modelling the likelihood that randomly generated N nucleotides will match IGHD genes (Collins*, et al.*, 2004).

While not as fast as IMGT/V-QUEST+JCTA, the program requires only about 5 seconds to align a query sequence using a 2GHz Intel Core Duo processor. The program requires approximately 5Mb of disk space to install and runs best with the java heap and stack sizes set to a minimum of 512Mb each.

3.2 Evaluation using clonally-related sequence sets

Ideally, evaluation of the accuracy of iHMMune-align predictions requires a benchmark set of rearranged immunoglobulin gene sequences of known germline gene composition. Since the germline gene composition cannot be known with confidence except for relatively unmutated sequences, we used sets of clonally-related sequences known to be derived from the same V-D-J rearrangements. Two sets were derived from tonsillar IgD class-switched B cells (Zheng*, et al.*, 2004), and have previously been described. The two sets consisted of 57 and 99 unique sequences for which all programs being tested could produce an alignment. Sequences were aligned against the germline IGHV, IGHD and IGHJ gene repertoires using iHMMune-align, IMGT/V-QUEST+JCTA (Giudicelli*, et al.*, 2004; Yousfi Monod*, et al.*, 2004) JOINSOLVER (Souto-Carneiro*, et al.*, 2004) and SoDA (Volpe*, et al.*, 2006). For this comparison iHMMune-align alignments were performed using both the Trinucleotide and Hotspot mutation models, with both models resulting in the same alignments. Other programs were accessed through their websites, using default program parameters. At the time of testing all programs used the same version of the IMGT germline gene repertoire.

The results of the analysis with the 57-sequences set are shown in Table 1. The four programs generally agreed on the longer IGHV and IGHJ genes (with variation with respect to predicted alleles), but differed with respect to IGHD gene alignment. Out of the four programs, iHMMune-align was the most consistent in its germline gene identification and predicted the same IGHD gene for all but one of the sequences.

Results for the 99-sequences set are presented in Table 2. This set included heavily mutated sequences that can present a challenge for any alignment program. IMGT/V-QUEST+JCTA, iHMMune-align and SoDA all selected the same V-D-J gene set (IGHV4-34*01 IGHD6-6*01 IGHJ6*02) for the majority of sequences in the set, while JOINSOLVER predicted an inverted IGHD gene in the majority of its alignments. The best performing program with regard to consistent predictions was iHMMune-align, which predicted the consensus gene set for 72 out of 99 sequences. With regard to the identification of the IGHD gene, iHMMune-align did not make a prediction for 12 out of the 99 sequences, as their IGHD gene alignments did not satisfy the 8-mer criterion, and iHMMune-align predicted a non-consensus IGHD gene in only 8 out of the 87 remaining sequences.

Table 1: Number of alignments in agreement, and numbers of alternative alignments seen when 57 clonally-related sequences were aligned using four alignment utilities

	Main alignment[a]	No. of alternative alignments				
		IGHV		IGHD	IGHJ	
		Gene	Allele	Gene	Gene	Allele
IMGT/V-QUEST+JCTA	19[b]	0	27	20	0	4
iHMMune-align	39[c]	0	4	1	0	14
JOINSOLVER	36[c]	0	6	4	0	15
SoDA	23[d]	0	14	19	0	12

[a] Number of alignments to the most commonly identified rearrangement for that program. This includes alternative alignments for six of the sequences where receptor revision has led to the replacement of the original IGHV gene with IGHV4-61*01

[b] IGHV4-34*12 (IGHV4-61*01) IGHD3-10*01 IGHJ3*02

[c] IGHV4-34*01 (IGHV4-61*01) IGHD7-27*01 IGHJ3*02

[d] IGHV4-34*04 (IGHV4-61*01) IGHD3-10*01 IGHJ3*02

Table 2: Number of alignments in agreement, and number of alternative alignments seen when 99 clonally-related sequences were aligned using four alignment utilities

	Main alignment[a]	No. of alternative alignments				
		IGHV		IGHD	IGHJ	
		Gene	Allele	Gene	Gene	Allele
IMGT/V-QUEST+JCTA	46[b]	0	36	10	17	0
iHMMune-align	72[b]	0	7	8	0	0
JOINSOLVER	37[c]	0	13	55	17	0
SoDA	61[b]	0	16	25	0	0

[a] No. of alignments to most commonly identified rearrangement for that program

[b] IGHV4-34*01 IGHD6-6*01 IGHJ6*02

[c] IGHV4-34*01 IGHDIR*01R IGHJ6*02

3.3 Evaluation using a random sequence sample

Clonally-related sequences can provide an effective benchmark set as they are known to be derived from the same V-D-J rearrangement and alignment utilities can therefore be evaluated based on their ability to predict the same V-D-J combination for every sequence in the set. However only few of these sequence sets are available. Further testing required the use of randomly sampled immunoglobulin sequences of unknown V-D-J composition.

A set of 662 rearranged immunoglobulin gene sequences that have previously been used in an evaluation of alignment software (Volpe*, et al.*, 2006) were collected from the EMBL database (Kanz, *et al.*, 2005). Germline genes were predicted for each of these sequences using each of the 4 alignment programs, with the same parameter settings as for evaluation with clonally-related sequences. Total agreement was only seen with 104 of the sequences. Sequences were removed from analysis if one or more utilities failed to identify an alignment, and Table 3 summarises the extent of consensus that was seen between the programs. Analysis of the ‘odd program out’, where consensus was seen between three of the four programs, highlights the differing performances of the programs, but does not allow firm conclusion to be drawn regarding accuracy since the original germline rearrangements are not known and different programs may agree more on the basis of similarity of algorithm than on actual predictive accuracy.

Table 3: Level of agreement between four alignment programs IMGT/V-QUEST+JCTA (IMGT), iHMMune-align (iHMM), JOINSOLVER (JS) and SoDA, in the alignment of 662 human immunoglobulin heavy chain sequences.

	IGHV		*IGHD*	*IGHJ*	
	Gene	Allele	Gene	Gene	Allele
Agreement[a]	357		291	322	
IMGT disagrees[b]	1	104	15	24	75
iHMM disagrees[b]	1	4	9	8	11
JS disagrees[b]	2	16	21	1	0
SoDA disagrees[b]	3	0	7	13	1
No agreement[c]	0	37	53[d]	12	58
TOTAL	525		396	525	

[a] All four programs agree

[b] Three programs agree, but one program disagrees

[c] There is no consensus by three or more programs

[d] Differences between IGHD1-1, IGHD1-7 and IGHD1-20 were not scored where alignments were of equal length

The quality of iHMMune-align alignments is further supported by analysis of the distribution of mutations between associated IGHV, IGHD and IGHJ genes. The shortest IGHD gene alignments (mean length of 13.3 nucleotides) were seen with JOINSOLVER and did not include mismatches. The longest were seen with IMGT/Junction Analysis, with a mean length of 17.9 nucleotides having mean 2.1 mutations. iHMMune-align alignments had a mean length of 15.4 nucleotides and 0.6 mutations, while SoDA alignments had a mean length of 16.1 nucleotides and 1.0 mutations.

JOINSOLVER and SoDA both allow alignments to inverted IGHD sequences. The appropriateness of this was tested by an examination of alignment lengths. The average length of 86 inverted SoDA alignments was 8.1 nucleotides, compared to an average 15.0 nucleotides for the 550 other SoDA IGHD alignments. The most common inverted SoDA alignment was to one or other of the IGHD2-2 alleles. The average length of these 37 alignments was 6.5 nucleotides, while the mean length of the 48 regular IGHD2-2 alignments was 20.5. The longest inverted alignment was 15 nucleotides, and this included 3 mismatches. This was an alignment to IGHD3-22*01. The inverted germline IGHD3-22*01 gene aligns well against many IGHD genes in the regular orientation, including a single mismatch in ten nucleotides to IGHD2-21*01 and six mismatches in 21 nucleotides to IGHD3-3*02. It should therefore not be surprising, for example, if a mutated IGHD3-3*02 gene occasionally aligns well to an inverted IGHD3-22*01 sequence.

As a further evaluation of IGHD and IGHJ gene alignments, we considered sequences that aligned perfectly against germline IGHV genes according to each alignment program. We then examined the level of mutations in the associated IGHD and IGHJ genes. Since the probability of somatic mutation decays exponentially with position along the sequence (Rada and Milstein, 2001), it is expected that sequences with no mutations in the IGHV gene should have very few or no mutations in their IGHD and IGHJ genes. The results are presented as Table 4, which shows lower levels of mutation in the iHMMune-align output. This is probably a measure of the ability of the different programs to correctly identify the ends of the genes, for most of the apparent mutations identified by the other programs were at the gene ends. No data are presented for JOINSOLVER because its algorithm does not allow for any IGHD mutations.

Table 4: Numbers of mismatches seen in IGHD and IGHJ alignments, in sequences that aligned to germline IGHV genes with no mismatches.

No. of mutations	iHMMune-align		SoDA		IMGT/ V-QUEST+JCTA	
	IGHD	IGHJ	IGHD	IGHJ	IGHD	IGHJ
0	148	147	117	145	17	42
1	8	9	31	15	12	10
2	0	0	8	1	9	16
3	0	0	4	0	10	0
4	0	0	0	0	20	0
5	0	0	10	0	0	0
TOTAL	156		161		68	

4 DISCUSSION

The identification and alignment of component germline genes in rearranged and mutated immunoglobulin gene and cDNA sequences is important not only for understanding the mechanisms for generating antibody diversity but also as part of many clinical investigations. We have developed iHMMune-align, an application that incorporates an explicit model of V-D-J recombination and somatic mutation processes in the form of a hidden Markov model. The use of an HMM allows modelling immunoglobulin gene-specific processes that are not represented when using standard pairwise sequence alignment techniques originally developed for aligning homologous sequences and modelling general evolutionary processes. The current version of iHMMune-align focuses on the alignment of human immunoglobulin heavy chains (IGH), although the algorithm can be applied to light chains and other organisms provided sufficient training data are available.

iHMMune-align includes an initial Smith-Waterman alignment step for identifying the IGHV gene. The IGHV region is relatively long (around 300 bases) and diverse, with 50 genes and 217 alleles having been reported to be functional.(Lefranc, 2001). Incorporating all possible IGHV genes in the HMM would therefore be too computationally expensive. The V-REGION is long enough to be readily identifiable by pairwise alignment, and all 4 utilities tested were in relatively good agreement regarding their IGHV gene assignment, although variations were observed with regard to predicted alleles. The pre-alignment of the IGHV gene also allows iHMMune-align to estimate the relative amount of somatic mutation over the sequence which is then used to calibrate the emission probabilities of the HMM.

The evaluation of immunoglobulin gene alignment accuracy is not straightforward as no "gold standard" sequence benchmark of known V-D-J composition is available. Germline gene composition can be inferred by expert visual inspection only for relatively unmutated sequences that are trivial to align by any approach, but not for more problematic mutated sequences where utilities will differ most in their predictions. Others have used simulated rearranged immunoglobulin genes as test sequences (Volpe*, et al.*, 2006), but this approach does not take into account the known (and unknown) mechanisms of immunoglobulin diversity generation and is unlikely to result in biologically significant conclusions. We propose the use of sets of clonally-related sequence sets for evaluating immunoglobulin gene alignment utilities. Although the V-D-J composition of these sequences is unknown, all sequences in a set are derived from a single rearrangement. An "ideal" alignment utility able to identify the original germline genes should therefore predict the same V-D-J composition for all sequences in the set. Two such sets were available to us, each displaying different characteristics in the CDR3 region (one uses a short IGHD gene, the other a longer IGHD gene). Together these sets provide a good estimate of the performance of different alignment approaches with highly mutated sequences, and complement evaluations with larger, more diverse sets of relatively unmutated sequences used in our and other studies. More comprehensive evaluation of immunoglobulin gene alignment methods will require the availability of additional sets of clonally-related sequences, but in their absence, we believe the present study represents the most thorough attempt to date to develop an objective benchmark for immunoglobulin gene alignment accuracy.

A comparison of iHMMune-align with the current standard application, IMGT/V-QUEST+JCTA, and JOINSOLVER and SoDA, two relatively new alignment programs, highlights its excellent performance. IHMMune-align predicted the V-D-J composition of the two sets of clonally-related sequences with higher accuracy than the other programs. Program quality was also evaluated by visual inspection of alignments, especially over the short IGHD gene and its junction with IGHV and IGHJ genes. iHMMune-align IGHD gene alignments were generally longer and contained fewer mismatches than the alignments generated by other programs. iHMMune-align models explicitly the factors known to affect somatic mutation and as a result avoids unlikely alignments that postulate a much larger number of mutations at the 3' end of the gene relative to its 5' end.

Alignment quality is also a function of the germline gene repertoire used by the program. All four programs were evaluated using the same release of the IMGT repertoire. However both JOINSOLVER and SoDA allow alignment to inverted IGHD genes, whose use is controversial (Corbett*, et al.*, 1997) and which are not included in the default repertoires used by IMGT programs and iHMMune-align. For every sequence in the test set where JOINSOLVER or SoDA predicted an inverted IGHD gene, iHMMune-align was able to generate a likely alignment with an IGHD gene in standard orientation.

One potential weakness of iHMMune-align algorithm is that it allows for nucleotide insertions or deletions at the gene junctions but not within germline genes. Some nucleotide insertions and deletions have been observed at small frequency in the somatic mutation process. However the need to maintain the immunoglobulin protein framework means that frameshift mutations in the coding regions are eliminated during clonal expansion and antigen selection and only in-frame insertions and deletions of 3 or 6 nucleotides are tolerated and have been reported in CDR1 and CDR2 (Wilson*, et al.*, 1998). Of the four algorithms tested, only SoDA allows for insertions and deletions within genes, but in our testing the overwhelming majority of insertions were single nucleotide insertions in the IGHV genes. Insertions were seen in 31 IGHD alignments, but again, all but one of these insertions were of a single nucleotide.

iHMMune-align is currently available as a Java application. Planned program developments include a web interface to facilitate access to the program for the casual user. The nature of the algorithm suggests that the accuracy of iHMMune-align can be improved with additional training data. We are currently focussing on improving the modelling of exonuclease removal, which should improve the reliability of short IGHD alignments.

REFERENCES

Basu, M., Hegde, M.V. and Modak, M.J. (1983) Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase, *Biochem Biophys Res Comms*, **111**, 1105-1112.

Benedict, C.L., Gilfillan, S., Thai, T.H. and Kearney, J.F. (2000) Terminal deoxynucleotidyl transferase and repertoire development, *Immunol Rev*, **175**, 150-157.

Bleakley, K., Giudicelli, V., Wu, Y., Lefranc, M.-P. and Biau, G. (2006) IMGT standardization for statistical analyses of T cell receptor junctions: The TRAV-TRAJ example, *In Silico Biology*, **6**, 0051.

Collins, A.M., Ikutani, M., Puiu, D., Buck, G.A., Nadkarni, A. and Gaeta, B. (2004) Partitioning of rearranged Ig genes by mutation analysis demonstrates D-D fusion and V gene replacement in the expressed human repertoire, *J Immunol*, **172**, 340-348.

Corbett, S.J., Tomlinson, I.M., Sonnhammer, E.L., Buck, D. and Winter, G. (1997) Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination, *J Mol Biol*, **270**, 587-597.

Damle, R.N., Wasil, T., Fais, F., Ghiotto, F., Valetto, A., Allen, S.L., Buchbinder, A., Budman, D., Dittmar, K., Kolitz, J., Lichtman, S.M., Schulman, P., Vinciguerra, V.P., Rai, K.R., Ferrarini, M. and Chiorazzi, N. (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia., *Blood*., **94**, 1840-1847.

Dorner, T. and Lipsky, P.E. (2001) Immunoglobulin variable-region gene usage in systemic autoimmune diseases, *Arthritis Rheum*, **44**, 2715-2727.

Giudicelli, V., Chaume, D. and Lefranc, M.P. (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis, *Nucleic Acids Res.*, **32**, W435-440.

Gotoh, O. (1990) Optimal Sequence Alignment Allowing for Long Gaps, *Bulletin of Mathematical Biology*, **52**, 359-373.

Hamblin, T.J., Davis, Z., Gardiner, A., Oscier, D.G. and Stevenson, F.K. (1999) Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia, *Blood*, **94**, 1848-1854.

Huang, S.C., Jiang, R., Hufnagle, W.O., Furst, D.E., Wilske, K.R. and Milner, E.C. (1998) VH usage and somatic hypermutation in peripheral blood B cells of patients with rheumatoid arthritis (RA), *Clin. Exp. Immunol.*, **112**, 516-527.

Jackson, K.J., Gaeta, B., Sewell, W. and Collins, A.M. (2004) Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire, *BMC Immunol*, **5**, 19.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2005) The EMBL Nucleotide Sequence Database, *Nucleic Acids Res.*, **33**, 1.

Kosmas, C., Stamatopoulos, K., Stavroyianni, N., Belessi, C., Viniou, N. and Yataganas, X. (1999) Molecular analysis of immunoglobulin genes in multiple myeloma, *Leukemia & Lymphoma*., **33**, 253-265.

Lee, C.E., Gaeta, B., Malming, H.R., Bain, M.E., Sewell, W.A. and Collins, A.M. (2006) Reconsidering the human immunoglobulin heavy-chain locus: 1. An evaluation of the expressed human IGHD gene repertoire, *Immunogenetics*, **57**, 917-925.

Lefranc, M.P. (2001) Nomenclature of the human immunoglobulin heavy (IGH) genes, *Exp Clin Immunogenet*, **18**, 100-116.

Lefranc, M.P. (2005) IMGT, the international ImMunoGeneTics information system(R): a standardized approach for immunogenetics and immunoinformatics, *Immunome Res*, **1**, 3.

Lefranc, M.P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clement, O., Chaume, D. and Lefranc, G. (2005) IMGT, the international ImMunoGeneTics information system, *Nucleic Acids Res*, **33**, D593-597.

Li, H., Cui, X., Pramanik, S. and Chimge, N.O. (2002) Genetic diversity of the human immunoglobulin heavy chain VH region, Immunol. Rev., 190, 53-68.

Martin, A. and Scharff, M.D. (2002) AID and mismatch repair in antibody diversification, *Nat Rev Immunol*, **2**, 605-614.

Moustafa, A. (2006) JAligner: Open source Java implementation of Smith-Waterman, http://jaligner.sourceforge.net.

Neuberger, M.S., Di Noia, J.M., Beale, R.C., Williams, G.T., Yang, Z. and Rada, C. (2005) Somatic hypermutation at A.T pairs: polymerase error versus dUTP incorporation, *Nature Reviews. Immunology*, **5**, 171-178.

Rabiner, L.R. (1989) A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, **77**, 257-286.

Rada, C. and Milstein, C. (2001) The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially, *EMBO J*, **20**, 4570-4576.

Ravetch, J.V., Siebenlist, U., Korsmeyer, S., Waldmann, T. and Leder, P. (1981) Structure of the human immunoglobulin μ locus: characterization of embryonic and rearranged J and D genes, *Cell*, **27**, 583-591.

Ruiz, M., Pallares, N., Contet, V., Barbie, V. and Lefranc, M.P. (1999) The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments, *Exp. Clin. Immunogenet.*, **16**, 173-184.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195-197.

Souto-Carneiro, M.M., Longo, N.S., Russ, D.E., Sun, H.-w. and Lipsky, P.E. (2004) Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER, *J. Immunol.*, **172**, 6790-6802.

Tobin, G., Thunberg, U., Johnson, A., Thorn, I., Soderberg, O., Hultdin, M., Botling, J., Enblad, G., Sallstrom, J., Sundstrom, C., Roos, G. and Rosenquist, R. (2002) Somatically mutated Ig V(H)3-21 genes characterize a new subset of chronic lymphocytic leukemia., *Blood*., **99**, 2262-2264.

Volpe, J.M., Cowell, L.G. and Kepler, T.B. (2006) SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations, *Bioinformatics*, **22**, 438-444.

Widhopf, G.F., Rassenti, L.Z., Toy, T.L., Gribben, J.G., Wierda, W.G. and Kipps, T.J. (2004) Chronic lymphocytic leukemia B cells of more than 1% of patients express virtually identical immunoglobulins, *Blood*, **104**, 2499-2504.

Wilson, P.C., de Bouteiller, O., Liu, Y.J., Potter, K., Banchereau, J., Capra, J.D. and Pascual, V. (1998) Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes, *J Exp Med*, **187**, 59-70.

Yoshida, M., Okabe, M., Eimoto, T., Shimizu, S., Ueda-Otsuka, K., Okamoto, M., Ishii, G., Ueda, R., Chan, J.K.C., Nakamura, S. and Inagaki, H. (2006) Immunoglobulin V-H genes in thymic MALT lymphoma are biased toward a restricted repertoire and are frequently unmutated, *J. Pathol.*, **208**, 415-422.

Yousfi Monod, M., Giudicelli, V., Chaume, D. and Lefranc, M.P. (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs, *Bioinformatics*, **20 Suppl 1**, I379-I385.

Zheng, N.-Y., Wilson, K., Wang, X., Boston, A., Kolar, G., Jackson, S.M., Liu, Y.-J., Pascual, V., Capra, J.D. and Wilson, P.C. (2004) Human immunoglobulin selection associated with class switch and possible tolerogenic origins for C delta class-switched B cells, *J. Clin. Invest.*, **113**, 1188-1201.

Zheng, N.Y., Wilson, K., Jared, M. and Wilson, P.C. (2005) Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation, *J. Exp. Med.*, **201**, 1467-1478.