In Silico Biology 6 (2006) 1–17 1
IOS Press

IMGT Standardization for Statistical Analyses of T Cell Receptor Junctions: The TRAV-TRAJ Example

Kevin Bleakley[a,*], Véronique Giudicelli[b], Yan Wu[b], Marie-Paule Lefranc[b,c] and Gérard Biau[a]

[a] *Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, CC 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France*
E-mail: {bleakley, biau}@math.univ-montp2.fr

[b] *IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France*
E-mail: {Veronique.Giudicelli, Marie-Paule.Lefranc, Yan.Wu}@igh.cnrs.fr@igh.cnrs.fr

[c] *Institut Universitaire de France*

Edited by E. Wingender; received 18 June 2006; revised and accepted 23 October 2006; published 23 November 2006

ABSTRACT: The diversity of immunoglobulin (IG) and T cell receptor (TR) chains depends on several mechanisms: combinatorial diversity, which is a consequence of the number of V, D and J genes and the N-REGION diversity, which creates an extensive and clonal somatic diversity at the V-J and V-D-J junctions. For the IG, the diversity is further increased by somatic hypermutations. The number of different junctions per chain and per individual is estimated to be 10^{12}. We have chosen the human TRAV-TRAJ junctions as an example in order to characterize the required criteria for a standardized analysis of the IG and TR V-J and V-D-J junctions, based on the IMGT-ONTOLOGY concepts, and to serve as a first IMGT junction reference set (IMGT®, http://imgt.cines.fr). We performed a thorough statistical analysis of 212 human rearranged TRAV-TRAJ sequences, which were aligned and analysed by the integrated IMGT/V-QUEST software, which includes IMGT/JunctionAnalysis, then manually expert-verified. Furthermore, we compared these 212 sequences with 37 other human TRAV-TRAJ junction sequences for which some particularities (potential sequence polymorphisms, sequencing errors, etc.) did not allow IMGT/JunctionAnalysis to provide the correct biological results, according to expert verification. Using statistical learning, we constructed an automatic warning system to predict if new, automatically analysed TRAV-TRAJ sequences should be manually re-checked. We estimated the robustness of this automatic warning system.

KEYWORDS: IMGT, T cell receptor, TRAV, TRAJ, variable gene, joining gene, junction, immunoglobulin, antibody, DNA rearrangement, IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT-ONTOLOGY, statistical learning, classification, k-Nearest Neighbors

INTRODUCTION

The diversity of the chains of the antigen receptors, immunoglobulins (IG) or antibodies and T cell receptors (TR) depends on several mechanisms: the combinatorial diversity, which is a consequence of the number of variable (V), diversity (D) and joining (J) genes in the IG and TR loci, and the N-REGION

*Corresponding author. Tel.: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; E-mail: bleakley@math.univ-montp2.fr.

Electronic publication can be found in *In Silico Biol.* **6**, 0051 <http://www.bioinfo.de/isb/2006/06/0051/>, 23 November 2006.

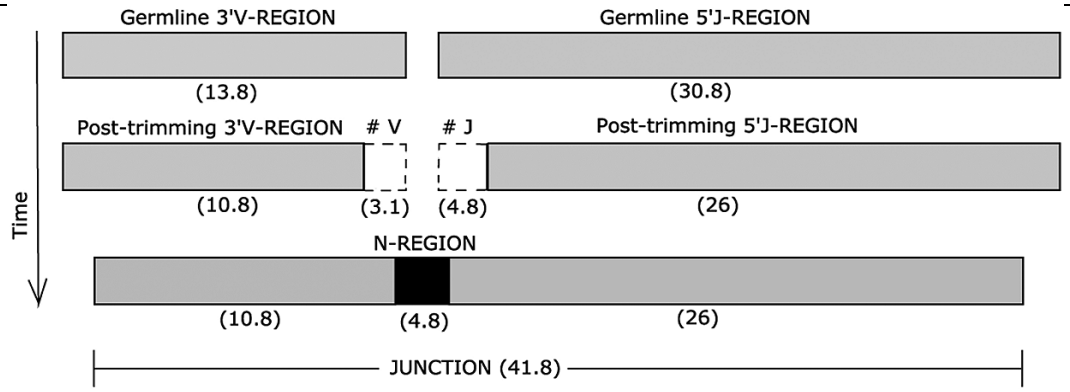

Fig. 1. The process of TRAV-TRAJ JUNCTION formation. The average length values in number of nucleotides are in parentheses. P-REGIONs (not shown) are only observed with an untrimmed 3'V-REGION and/or 5'J-REGION. The first nucleotide of the 3'V-REGION is the first nucleotide of the 2nd-CYS codon[1] [8]. The last nucleotide of the 5'J-REGION is the third nucleotide of the J-PHE[b] codon (for the TRAJ gene) [8].

diversity, which creates an extensive and clonal somatic diversity at the V-J and V-D-J junctions (for review [1,2]). For the IG, the diversity of the variable domains (V-J-REGION of the light chains and V-D-J-REGION of the heavy chains) is further increased by somatic hypermutations [2].

The number of different junctions per chain and per individual is estimated to be 10^{12} in humans and the only limiting factor seems to be the number of B cells (for the IG) and T cells (for the TR) which is genetically programmed in a given species. In the human genome, the IG and TR genes are localized in seven major loci, three IG (IGH, IGK and IGL) loci and four TR (TRA, TRB, TRG and TRD) loci [1,2]. We have chosen V-J rearrangements in the human TRA locus (TRAV-TRAJ junctions) as an example in order to characterize the required criteria for a standardized analysis of the IG and TR V-J and V-D-J junctions, based on the IMGT-ONTOLOGY concepts [3,4,5], and to serve as a first IMGT junction reference set. Indeed, rearranged TRAV-TRAJ sequences are an ideal starting point for antigen receptor junction sequence analysis because of their lack of somatic mutations [1], which means that automatic alignment with archived variable (V) and joining (J) germline genes is usually very precise. Moreover, in contrast to the TRB or TRD loci, there is no D gene involved in the rearrangement [1].

IMGT® the international ImMunoGeneTics information system® (http://imgt.cines.fr) [6], the widely acknowledged reference in immunogenetics and immunoinformatics [4,5], has developed tools for the analysis of the antigen receptor rearranged sequences, based on the IMGT Scientific Chart rules for the classification, description and numerotation of the IG and TR genes [3]. These tools, widely used and available on the website, comprise IMGT/V-QUEST [7], a software which automatically aligns a sequence with the closest (in terms of the proportion of correctly matched nucleotides in the alignment to the total number of nucleotides) germline V and J genes of the IMGT reference directory, and IMGT/JunctionAnalysis [8], a tool which predicts exactly what has happened at the junction of the V-J and V-D-J recombining genes. Indeed, several nucleotides may be trimmed at the ends of the 3'V-REGION[2] and 5'J-REGION of the respective genes and a number of nucleotides (N-REGION)

[1] 2nd-CYS and J-PHE are IMGT labels and thus are in capital letters. 2nd-CYS defines the second cysteine (Cys) of the disulfide bridge in the V-LIKE-DOMAIN. J-PHE defines the phenylalanine (Phe) of the J-REGION at position 118 in the IMGT Collier de Perles. Cys and Phe are abbreviations according to the IUPAC-IUB Joint Commission on Biochemical Nomenclature.

[2] IMGT® labels from the DESCRIPTION concept of IMGT-ONTOLOGY are in capital letters [3,4,5].

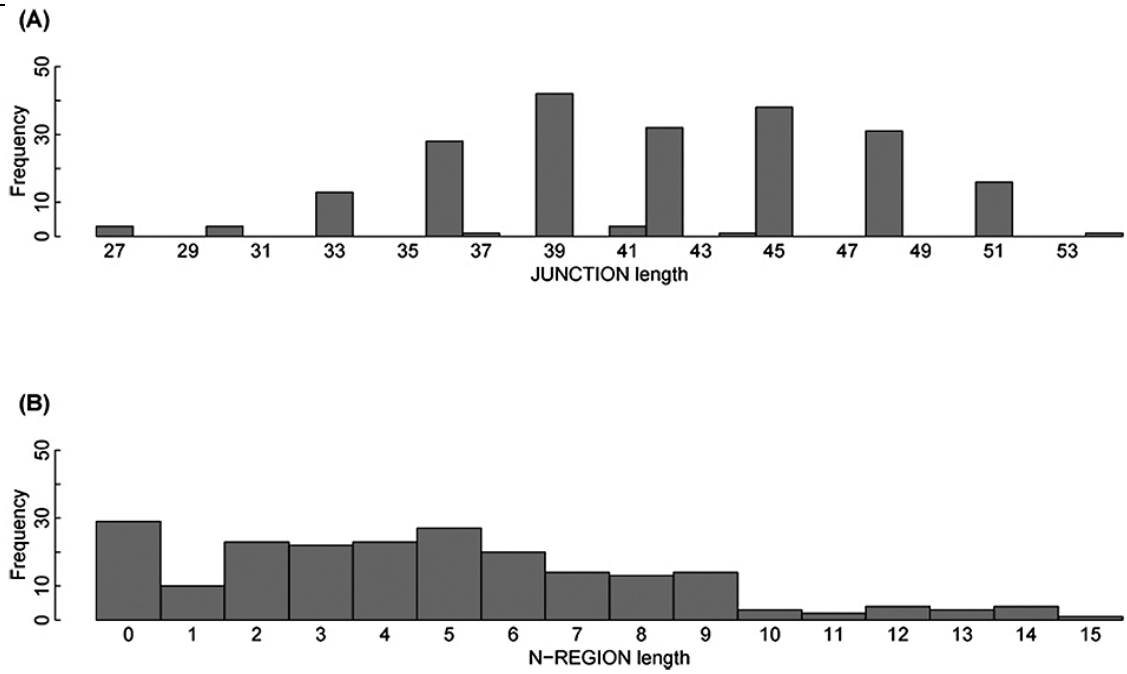

Fig. 2. Histograms of the JUNCTION length (A) and N-REGION length distribution (B) for the set of 212 human rearranged TRAV-TRAJ junction sequences.

may be added at random by the terminal deoxynucleotidyltransferase TdT [9] before the joining of the rearranging V and J genes (for review [1,2]). If there is no nucleotide trimming, P-REGION nucleotides may be observed [10]. These are short inverted-repeat sequences identified at the V-J junction, resulting from the dissymetric opening of the hairpin formed at the ends of the coding regions during V-J recombination [1,2,10,11].

The rich and apparently random mechanisms that regulate the trimming at the 3'V-REGION and 5'J-REGION ends of the V and J genes and the addition of the N-REGION sequence between these two genes lead to a diversity of nucleotide junction sequences, hence of the resulting amino acids and the proteins coded from these. The mechanisms of nucleotide trimming and N-diversity are thus part of the reason for the large diversity of antigen receptors of the human adaptive immune response, 10^{12} antibodies and 10^{12} T cell receptors per individual [1,2].

As far as we know, no systematic standardization of human T cell receptor junction variables, nor any detailed statistical analysis of these variables has previously been undertaken. This in part is due to the difficulty in compiling large, clean data sets until very recently. This has been overcome by the development of the IMGTfi specialised sequence and gene databases IMGT/LIGM-DB and IMGT/GENE-DB [12,13], tools for sequence analysis, IMGT/V-QUEST and IMGT/JunctionAnalysis [7,8], and related Web resources (IMGT Repertoire) [4,5,6]. To answer current needs due to an increase of experimental data, IMGT/V-QUEST [7] was improved in May 2006. The new version with additional functionalities allows analysis by batches of sequences, provides a detailed description of the mutations and integrates the results of IMGT/JunctionAnalysis [8]. This leads to a change of scale, making data available for statistical analysis. Owing to the complexity of the mechanisms involved in the high diversity of the junctional process, the aim of this paper was to characterize the criteria required for a standardized and meaningful statistical analysis of the junctions. We performed a thorough statistical analysis of 212 archived human T

Table 1
Statistical summary of the ten important length variables for human TRAV-TRAJ sequences

Variables	Mean	Var.	Min.	1st Q.	Median	3rd Q.	Max.
JUNCTION length (Junc)	41.8	30.7	27	39	42	45	54
N-REGION length (N)	4.8	12.2	0	2	4	7	15
germline 3'V-REGION length (g·V)	13.8	2.3	10	13	13	14	17
post-trimming 3'V-REGION length (post·V)	10.8	6.9	4	9	11	13	17
germline 5'J-REGION length (g·J)	30.8	13.9	23	29	31	32	38
post-trimming 5'J-REGION length (post·J)	26	25.2	9	23	26	29	38
number of trimmed V nucleotides (#V)	3.1	6.1	0	1	3	5	10
number of trimmed J nucleotides (#J)	4.8	11.8	0	2	4	6	21
P3'V length (P3'V)	0.2	0.3	0	0	0	0	3
P5'J length (P5'J)	0.1	0.2	0	0	0	0	3

The lengths of the ten variables are in number of nucleotides. The analysis was performed on a set of 212 sequences extracted from IMGT/LIGM-DB [12], analysed by IMGT/V-QUEST + JCTA [7,8] and expert verified. Abbreviations used in Table 2 shown in parentheses.

cell receptor rearranged TRAV-TRAJ sequences, which were extracted from IMGT/LIGM-DB [12] and analysed by IMGT/V-QUEST [7] with integrated IMGT/JunctionAnalysis [8] (which we designate from now on IMGT/V-QUEST + JCTA). IMGT/V-QUEST automatically aligns the rearranged sequences with germline V and J genes, and selects the closest V and J germline genes from the IMGT reference directory [7]. Next, IMGT/JunctionAnalysis analyses each nucleotide sequence in the junction of the TRAV and TRAJ genes, and its results indicate which and how many nucleotides have been trimmed from the 3' end of the V-REGION and the 5' end of the J-REGION, or predicts the existence of a short P-REGION sequence at either end. Finally, IMGT/JunctionAnalysis predicts which and how many N-REGION nucleotides have been inserted between the 3'V and the 5'J ends.

The paper is structured as follows. First we introduce a list of important variables related to these TRAV-TRAJ junction sequences which can be represented mathematically, and use standard statistical methods to describe and visualise these variables and the relationships between them. We then take a first step to examine the hypothesis that specific nucleotide sequences have an effect on the quantity of trimming of the 3'V-REGION and 5'J-REGION sequences. Following this, we perform statistical analyses on variables related to the frequencies and proportions of A, C, G and T nucleotides in the TRAV-TRAJ junction sequences. Next, we look for differences between the mathematical properties of the 212 TRAV-TRAJ junction sequences which were identified correctly by IMGT/V-QUEST + JCTA (according to manual verification by an expert) and those of 37 sequences for which IMGT/V-QUEST + JCTA provided potentially incorrect or dubious biological results due to sequence polymorphisms, sequencing errors or other doubts on the junction (again according to manual verification by an expert). Using a statistical learning approach, we have created an algorithm which can predict whether a new (unseen) sequence has been correctly analysed by IMGT/V-QUEST + JCTA. This might be integrated into IMGT/V-QUEST + JCTA as an automatic warning for the user to manually check the automatic IMGT/JunctionAnalysis results. Error estimations are presented.

A CLASSICAL STATISTICAL ANALYSIS

Ten important variables were extracted from XML documents that had been created from the output of the alignment (Denis Chaume, V.G. and M.-P. L., unpublished data) and analysis of the 212 TRAV-TRAJ junction sequences by IMGT/V-QUEST+JCTA. This extraction was performed using a detailed Perl language script (called IMGT-PERLSTAT) written especially for these XML documents (written

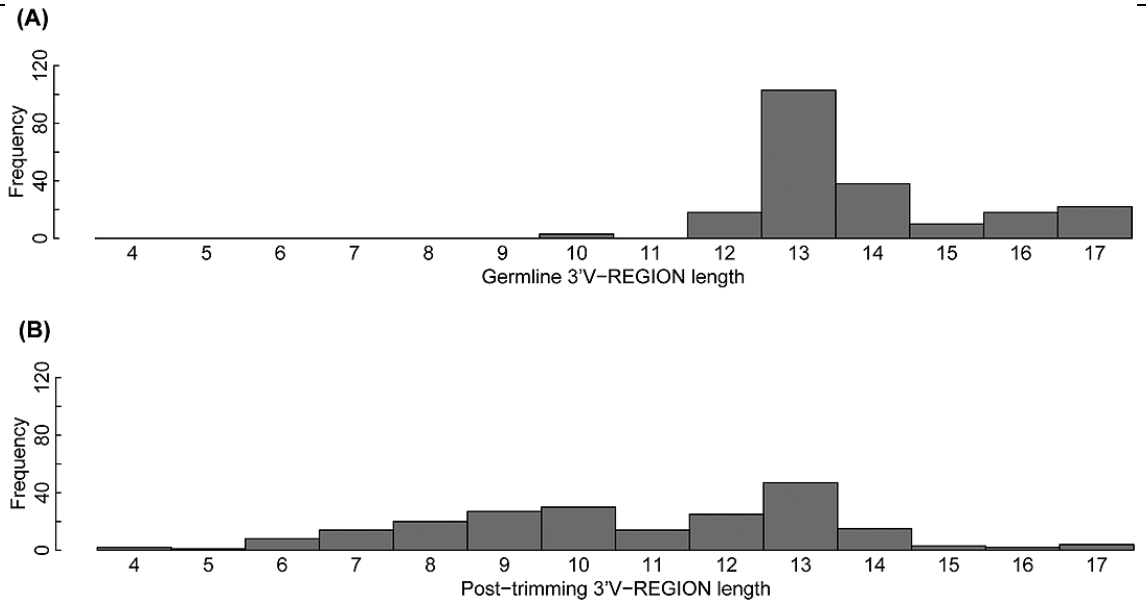

Fig. 3. Histograms of the germline 3'V-REGION length (A) and post-trimming 3'V-REGION length distribution (B) for the set of 212 human rearranged TRAV-TRAJ junction sequences.

by K.B.). Its job was to automatically remove the 10 important variables from the XML and to store it in text files which could be read using standard mathematical software packages, in particular Matlab® and *R*. This collection of variables, shown with some basic statistical results in Table 1, comprises:

- The total length of the JUNCTION sequence (in nucleotides, not in amino acids) from the first nucleotide of the 3'V-REGION to the last nucleotide of the 5'J-REGION ("JUNCTION length", Fig. 2A).
- The length of the N-REGION added between the 3'V-REGION and the 5'J-REGION ("N-REGION length", Fig. 2B).
- The length of the 3'V-REGION before trimming has occurred ("germline 3'V-REGION length", Fig. 3A).
- The length of the 3'V-REGION after trimming has occurred ("post-trimming 3'V-REGION length", Fig. 3B).
- The length of the 5'J-REGION before trimming has occurred ("germline 5'J-REGION length", Fig. 4A).
- The length of the 5'J-REGION after trimming has occurred ("post-trimming 5'J-REGION length", Fig. 4B).
- The number of V nucleotides trimmed off the end of the germline 3'V-REGION ("number of trimmed V nucleotides", Fig. 5A).
- The number of J nucleotides trimmed off the end of the germline 5'J-REGION ("number of trimmed J nucleotides", Fig. 5A).
- The length of the P-REGION downstream of the 3'V-REGION ("P3'V length"), if it exists.
- The length of the P-REGION upstream of the 5'J-REGION ("P5'J length"), if it exists.

Figure 1 gives a simple diagram showing the process of TRAV-TRAJ recombination, with the average length values in parentheses. The histograms in Figs 2, 3, 4 and 5 show all of the pertinent information

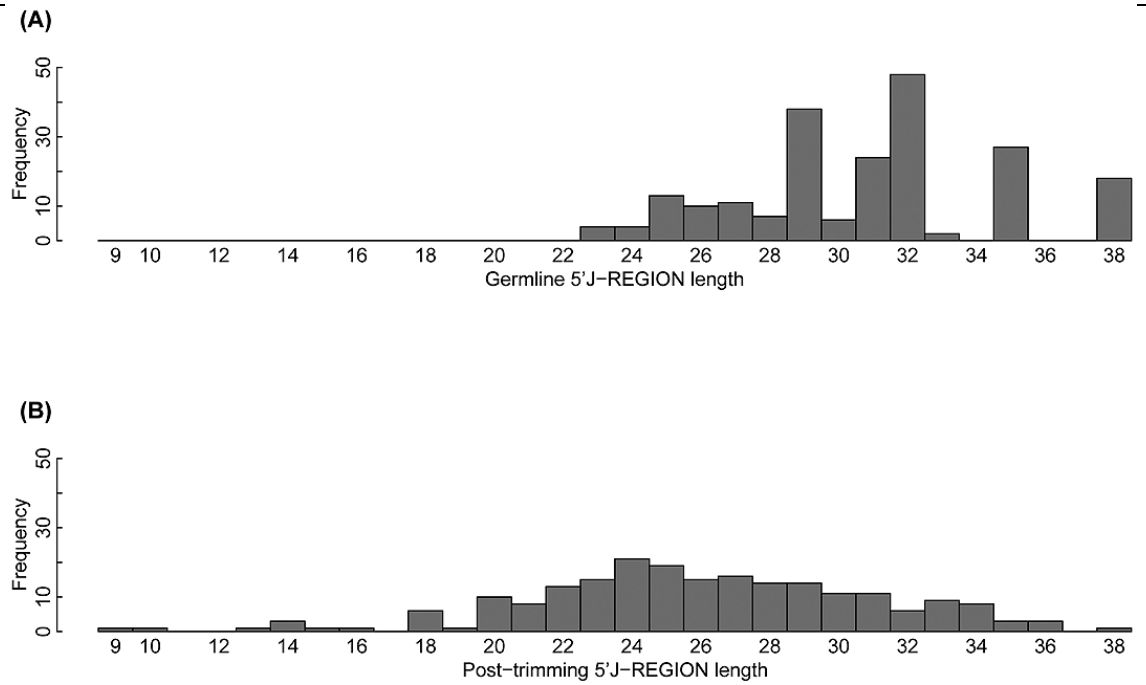

Fig. 4. Histograms of the germline 5'J-REGION length (A) and post-trimming 5'J-REGION length distribution (B) for the set of 212 human rearranged TRAV-TRAJ junction sequences.

for these variables in visual terms (except for the P3'V length and P5'J length variables which were almost always zero, see Table 1).

Analysis of These Variables

Some preliminary observations of Table 1 are in order. We remark that for every variable, the mean is very close to the median, and the 1st and 3rd quartiles are also very close to the mean and median. This means that at least half of the data (related to each variable individually) is grouped closely around the mean and median. However, there is often a significant amount of data far from the mean and median, which can be seen quantitatively in the (usually) large variances.

Individual commentaries should be made about several of the variables:

- The JUNCTION length (Fig. 2A) is typically a multiple of three for this data, reflecting the fact that the great majority of the sequences treated (206 out of 212) are *in frame*, that is the nucleotides by groups of three (codon triplet) are able to code for amino acids. This reflects the bias that sequences submitted by researchers to the databases are more frequently expressed sequences that are in frame. In a random uniform process, we would predict two thirds of naturally occurring sequences to be out of frame. However, the JUNCTION length bias has no obvious effect on the other variables analysed here.
- The N-REGION length data can be divided into two disjoint sets, $N < 10$ and $N \geqslant 10$ (Fig. 2B). Specifically, for $N \geqslant 10$ there is a sharp cutoff in the number of sequences. It is tempting to hypothesize that the N-REGION is formed under a strict time limit. We note also that in 30 examples, there is no N-REGION formed.

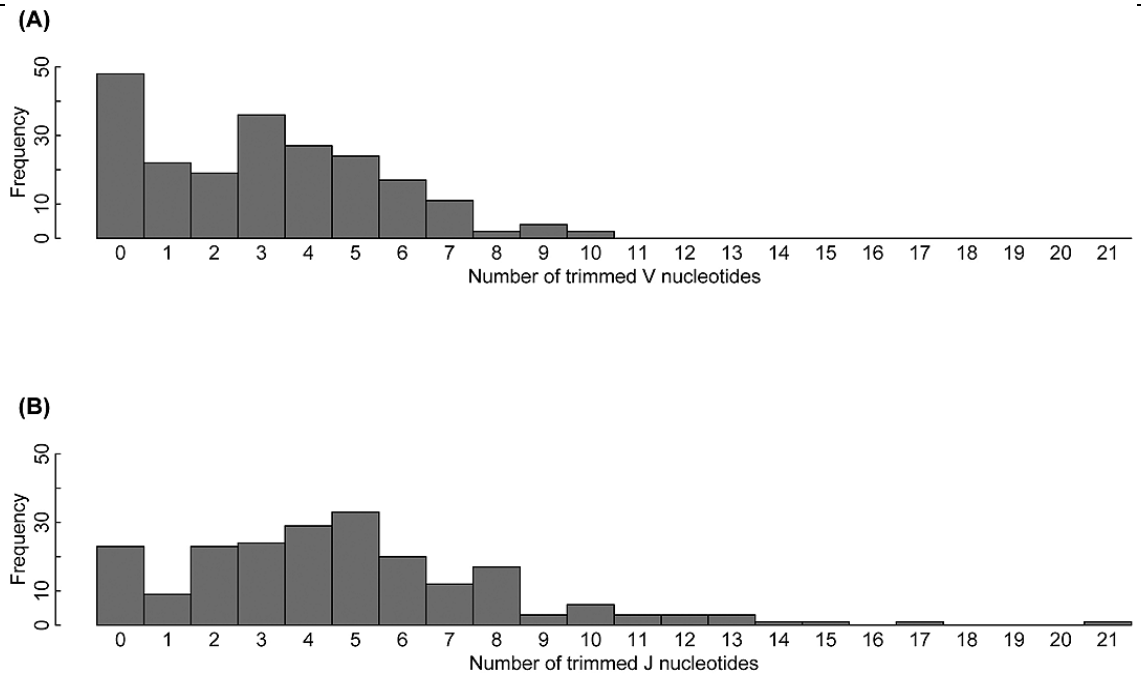

Fig. 5. Histograms of the number of trimmed V nucleotides and number of trimmed J nucleotides for the set of 212 human rearranged TRAV-TRAJ junction sequences.

- The germline 3'V-REGION lengths range between 10 and 17 (Fig. 3A). Nearly half of the germline 3'V-REGION lengths are 13, which is in agreement with the germline human TRAV repertoire as described in IMGT Alignment of Alleles [1,2] and IMGT/GENE-DB [13].
- The germline 5'J-REGION lengths are distributed rather strangely (Fig. 4A). There is a fairly strong increase in the frequencies at intervals of 3: 29, 32, 35, 38, which corresponds to the presence of two nucleotides between the recombination signal and the first complete codon of the J-REGION (thus giving multiples of 3, plus 2). This is in agreement with the heterogeneity of the TRA germline 5'J-REGION lengths in IMGT Alignment of Alleles [1,2] and IMGT/GENE-DB [13].
- The number of trimmed V nucleotides and the number of trimmed J nucleotides are potentially very interesting variables (Fig. 5). One of the future goals of this work is to characterize this trimming mechanism better. Why does it start? Why does it stop? Are the N-REGION nucleotide sequences related to the trimmed sequences? Quantitatively, we see that whilst the largest observed number of trimmed J nucleotides was 21 in this data set, the great majority of germline TRAJ sequences were trimmed by at most 10 nucleotides. In the case of the germline TRAV sequences, the largest observed number of trimmed V nucleotides from such a sequence was 10.
- We remark that the average percentage trimming of the germline 3'V-REGION was 22%, compared with 16% for the germline 5'J-REGION. Nevertheless, the overall average germline 5'J-REGION length is 2.3 times larger than the overall average germline 3'V-REGION length. This means that, even with a smaller average trimming percentage, 1.6 times more J nucleotides are trimmed than V nucleotides, on average. However, when junction sequences are examined one by one, we note that this does not imply that systematically, in each particular junction, *more* J nucleotides are trimmed than V nucleotides. In fact, in more than one third of the junctions, *less* J nucleotides were trimmed off the end of the 5'J-REGION than V nucleotides off the end of the 3'V-REGION.

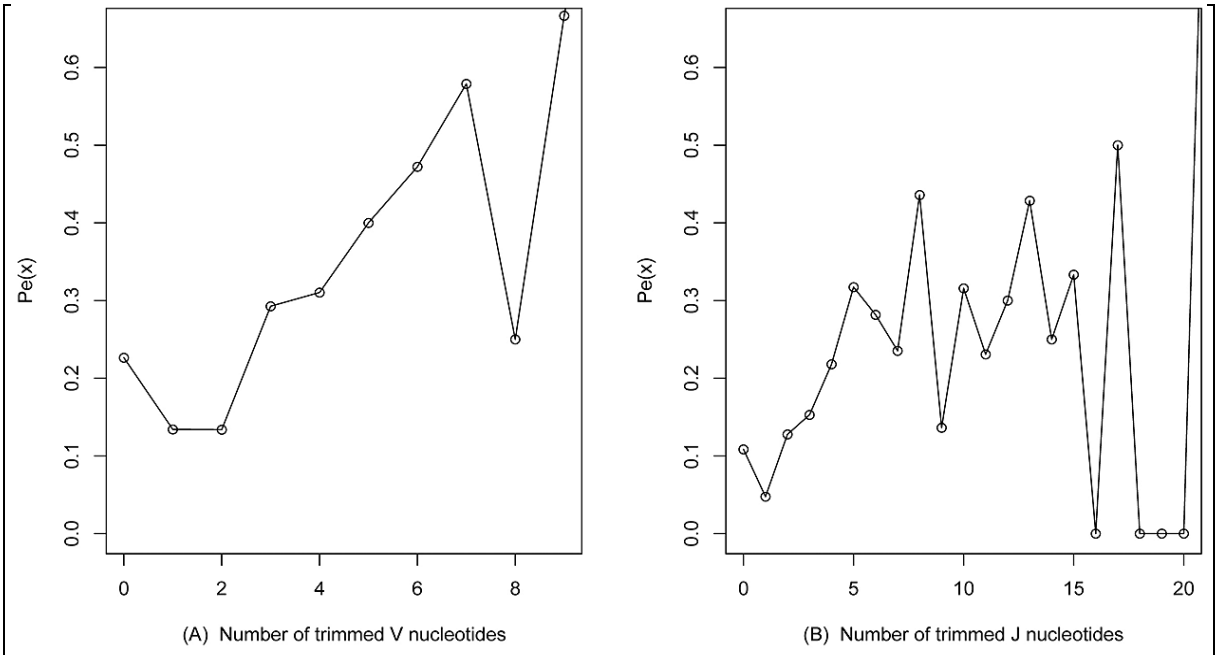

Fig. 6. Empirical probability $P_e(x)$ that trimming stops after x nucleotides.

– The P3'V and P5'J length was almost always zero, with a maximum length in this data set of 3 nucleotides. The P3'V length was equal to 0, 1, 2, 3 exactly 190, 8, 12, 2 times, respectively. The P5'J length was equal to 0, 1, 2, 3 exactly 199, 6, 5, 2 times, respectively. There was no example of the existence of both a P3'V and P5'J sequence in the same junction in this data set.

Probabilities of the number of trimmed 3'V-REGION and 5'J-REGION nucleotides

Whilst there is no known biological mechanism involved, we nevertheless initiated an investigation into whether previously trimmed 3'V-REGION and 5'J-REGION nucleotides had an effect on the continuation of the very same trimming process. More specifically, using the empirical frequencies of trimmed 3'V-REGION and 5'J-REGION nucleotides, we considered the following question: *Given that x nucleotides have been trimmed, what is the* (*empirical*) *probability $P_e(x)$ that the trimming will stop at x and not continue on to $x+1$ or beyond*? This calculation allows us to see quite clearly which trimming lengths tend to be the lengths for which the trimming stops. The results are shown in Fig. 6.

For the 3'V-REGION data (Fig. 6A), we see from 1 to 7 trimmed nucleotides an increasing probability function. Essentially this means that as the number of trimmed V nucleotides increases, the probability of the trimming stopping increases, or equivalently the probability of the trimming process continuing decreases as the number of trimmed V nucleotides increases. However, for the 5'J-REGION data (Fig. 6B), the probability function fluctuates markedly, such as the peak at 8 and the sudden drop at 9. This results directly from the perhaps abnormally large number of sequences for which 8 J nucleotides were trimmed compared with the small number of sequences which had 9 trimmed J nucleotides. This kind of fluctuation might have a biological explanation but equally may just be due to the limited sample size of our data, meaning that occasionally this finite data set approximates poorly the assumed underlying probability distribution.

Table 2
Correlations between the 10 length variables for the IMGT junction reference set of human TRAV-TRAJ junctions

	Junc	N	g·V	post·V	g·J	post·J	#V	#J	P3'V	P5'J
Junc	1	0.37	0.26	0.19	**0.62**	**0.72**	−0.04	−0.39	0.13	0.06
		1	−0.04	−0.24	0.07	−0.14	0.23	0.28	−0.03	−0.13
			1	0.39	0.17	0.11	0.20	0.03	0.02	0.01
post·V				1	−0.13	−0.18	**−0.82**	0.12	0.38	−0.08
g·J					1	**0.73**	0.25	0.02	−0.10	−0.13
post·J						1	0.26	**−0.69**	−0.14	0.12
#V							1	−0.11	−0.40	0.09
#J								1	0.09	−0.33
P3'V									1	−0.07
P5'J										1

The ten variables are defined in the text and reported in Table 1: JUNCTION length (Junc), N-REGION length (N), germline 3'V-REGION length (g·V), post-trimming 3'V-REGION length (post·V), germline 5'J-REGION length (g·J), post-trimming 5'J-REGION length (post·J), number of trimmed V nucleotides (#V), number of trimmed J nucleotides (#J), P3'V length (P3'V) and P5'J length (P5'J).

Correlations between Variables

The next analysis involved calculating the correlations between each of these 10 variables. Firstly, the correlations were estimated using the standard Pearson product-moment correlation coefficient [14]. The results are shown in Table 2, with the most significant correlations shown in bold type. We followed this correlation analysis with a more general analysis of the dependence between each pair of the 10 variables (X, Y), using the standard definition of *mutual information*,

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{f(x)g(y)}$$

where p is the joint probability distribution of X and Y, and f and g are the marginal probability distribution functions of X and Y, respectively [15]. I is able to quantify more generally than the Pearson correlation any functional dependencies between two variables. The larger the value of I, the more *similar* the variables X and Y. In the absence of the actual probability distributions of X and Y, we calculated I empirically.

We show that all of the most significant correlations (shown in bold type in Table 2) do not indicate any particularly new biological insights. The correlations between JUNCTION length and post-trimming 5'J-REGION length (0.72) or germline 5'J-REGION length (0.62) can be explained by the strong dependence of the overall junction length on the length of the (post-trimming or germline) 5'J-REGION. i.e, if the junction is long, it is usually because the 5'J-REGION is long, and *vice versa*. The correlation between the post-trimming 3'V-REGION length and the number of trimmed V nucleotides (-0.82) is expected since the germline 3'V-REGION lengths do not vary very much [1,8] (and these two afore-mentioned variables sum to give the germline 3'V-REGION length). The same is true for the correlation between the post-trimming 5'J-REGION length and the number of trimmed J nucleotides (-0.69). The correlation between the post-trimming 5'J-REGION length and the germline 5'J-REGION length (0.73) is due to the fact that on average only 16% of the germline 5'J-REGION genes are trimmed, and consequently the relationship between these two variables is almost linear, hence the high correlation. Interestingly, there is almost no correlation between the number of trimmed V nucleotides and the germline 3'V-REGION

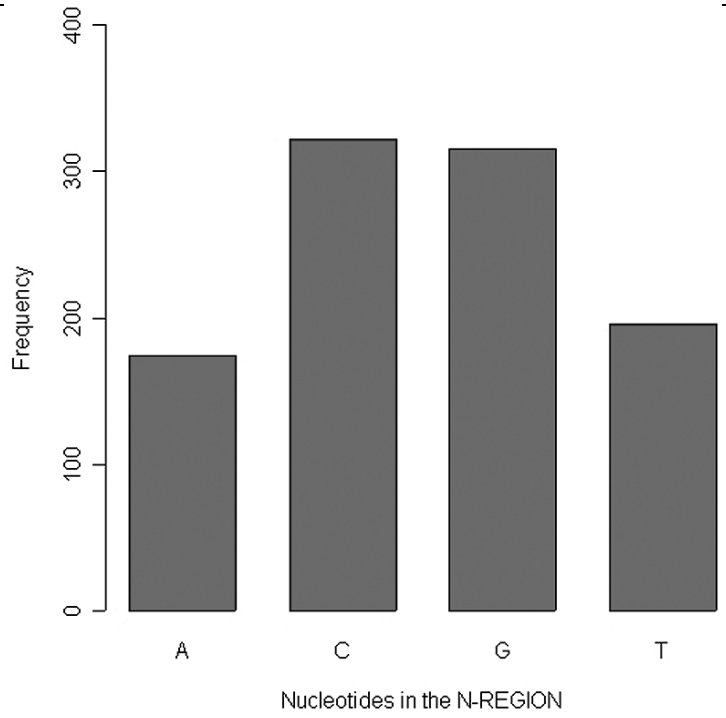

Fig. 7. Total number of each nucleotide type summed over all N-REGION sequences.

length (0.20), nor between the number of trimmed J nucleotides and the germline 5'J-REGION length (0.02). This indicates that the number of V and J nucleotides trimmed at the end of the 3'V-REGION or 5'J-REGION is not linearly related to the original germline 3'V-REGION or 5'J-REGION lengths, respectively.

The mutual information criterion I did not reveal any extremely significant relationships between the 10 variables. However, we note in passing the largest interesting mutual information was found between the N-REGION length and the post-trimming 5'J-REGION length. There is no immediate explanation for this.

Nucleotide-Specific Variables

To finish this section, we briefly looked at variables related to the actual nucleotides in the sequences, rather than just those related to lengths of parts of the junction:

- The overall frequency of each type of nucleotide in the N-REGION.
- The overall frequency of each type of nucleotide in the trimmed 3'V-REGION and 5'J-REGION sequences.

We calculated the frequencies of the A, C, G and T nucleotides, summed over all the N-REGIONs in the data set. This is shown in Fig. 7.

We see that the N-REGION is formed with a clear preference for C/G nucleotides over A/T, an observation often found in the literature. C/G represents 63% of the N-REGION nucleotides, whereas A/T represents 37%. Two possibilities: If the random availability of each of A, C, G and T nucleotides is equal in the cell, then clearly the creation of the N-REGION is not entirely a random process. Indeed,

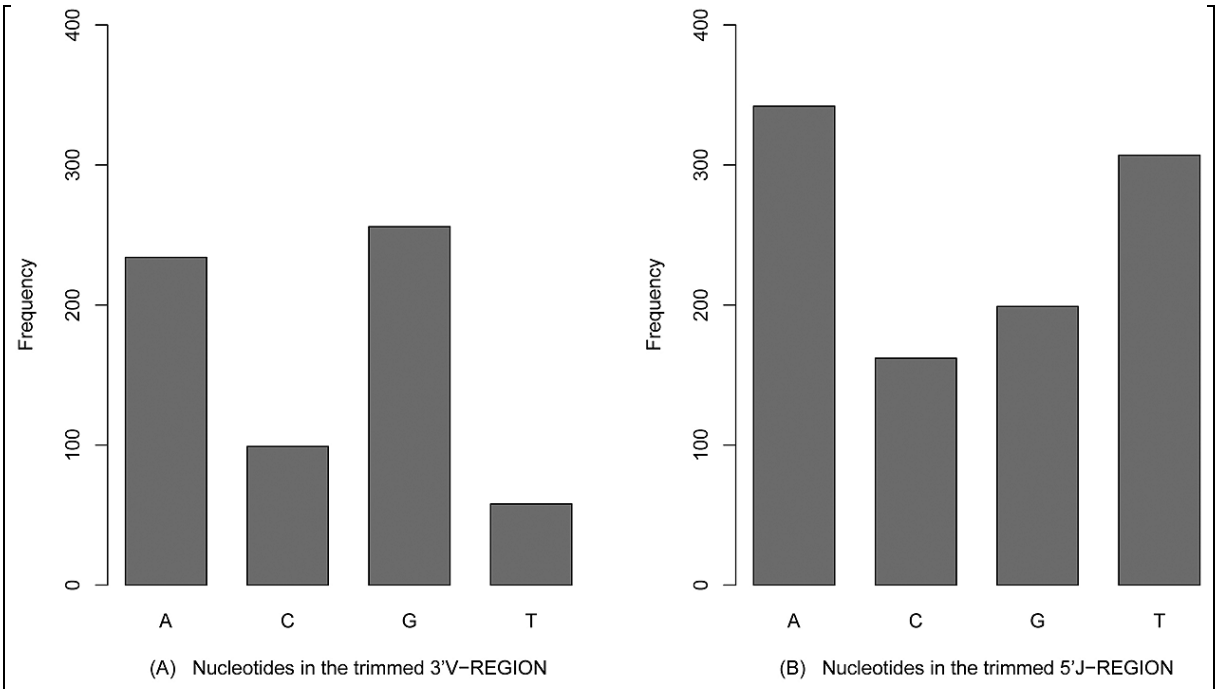

Fig. 8. Total number of each nucleotide type summed over all trimmed 3'V-REGION (A) and 5'J-REGION (B) sequences.

TdT is thought to prefer C/G [9]. The second possibility is that the availability of A/T is inferior to the availability of C/G during the recombination process.

Next, we calculated the overall frequency of A, C, G and T nucleotides in the trimmed 3'V-REGION and 5'J-REGION. This is shown in Fig. 8.

We see in Fig. 8A that A and G nucleotides make up 76% of all trimmed 3'V-REGION nucleotides. This is representative of the fact that the end of the 3'V-REGION is typically rich in A and G nucleotides. Figure 8B shows that A and T nucleotides make up 64% of all trimmed 5'J-REGION nucleotides. This is also representative of the typical richness of A and T nucleotides at the end of the 5'J-REGION.

We briefly looked at the possibility that the trimmed 3'V-REGION and 5'J-REGION nucleotides were recycled into forming the N-REGION. Whilst we are not currently in a position to make any broad conclusions, one thing is sure: it is impossible that C nucleotides from the trimmed 3'V-REGION and 5'J-REGION are systematically recycled to be used as the C nucleotides in the N-REGION, since, upon adding over the 212 sequences, there were only 261 available C nucleotides, yet a total of 322 C nucleotides were found in the N-REGIONs. Whilst this does not mean that C nucleotides cannot be recycled, it does however imply that sometimes C nucleotides must be recruited from elsewhere. Upon looking closer at individual sequences, we ascertained that the same is true for A, G and T nucleotides, since it is common that nucleotides in the N-REGION are not found in the corresponding trimmed 3'V-REGIONs and 5'J-REGIONs.

IMPROVING IMGT/V-QUEST + JCTA USING STATISTICAL LEARNING

In this section, we show how statistical learning could be used to implement a warning signal in IMGT/V-QUEST+JCTA for the user to manually check the automatic alignment and analysis performed

by this software in the case of unusual sequences (potentially new polymorphisms, sequencing errors). In this case we are dealing with what is commonly called *binary classification* [16,17,18,19], since we are aiming to predict a label of $+1$ or -1 for each input sequence, $+1$ if we think the sequence was incorrectly analysed, -1 if we think the sequence was correctly analysed. The general idea is to find a function (also called in this context a *rule* or *classifier*) that mathematically maps each sequence to either the value $+1$ or -1 in such a way that this function does not make many misclassifications. The quality of a classifier is therefore represented by how much of the time it correctly predicts the label.

In order to choose a good function, we must use the data we have already to test which function is the best for us. Unfortunately, we can always find a rule that works 100% correctly on the data we know, but works arbitrarily bad on new data (this is called *overfitting*, see [18], chapter 12). A common strategy to deal with this problem is to choose a rule that classifies correctly most of the known data, but so that the rule is as uncomplicated as possible in a mathematical sense. As a simple example, an uncomplicated rule would be to draw a straight line across a plane and say that any data falling on one side of the line be labelled $+1$, and on the other side -1. A more complicated rule would be to choose a polynomial curve to cut the plane into two halves. Indeed, such a rule may better divide the known data, but as the rule becomes more complicated (e.g., the degree of the polynomial becomes larger) we become less and less sure of the robustness of the rule, that is, the ability of the rule to generalise to new unseen data.

Of course, we only have a finite set of data, and so it is not simple to use this data both to choose a function, and also to predict how well this function will work on new, unseen data. A standard methodology called leave one out [19] involves putting aside one object, then using all of the other data to choose a classification rule, and finally to test this rule on the left-out object. This process is repeated once for each object, which gives us at the end an approximation to the probability that this rule correctly classifies data.

Mathematical Preliminaries

We suppose our data $(X_i, Y_i)_{i=1,\ldots,n}$ are independent and identically distributed random variables, with $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1, +1\}$ representing the label for X_i. Our aim is to design a function $g : \mathbb{R}^d \to \{-1, +1\}$ which predicts Y from X. The performance of this function is measured by the risk:

$$R(g) = \mathbb{P}\{g(X) \neq Y\},$$

that is, the probability that g incorrectly predicts the label of X. In the ideal case, $\mathbf{R}(g)$ should be close to the *Bayes risk* $\mathbf{R}^*$ which is the minimal probability of error

$$\mathbf{R}^* = \inf_g \mathbf{R}(g) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\{|2\eta(X) - 1|\},$$

where the infimum is over all (measurable) functions $g : \mathbb{R}^d \to \{-1, +1\}$, and $\eta(x) = \mathbb{P}\{Y = +1|X = x\}$ is the posterior probability ($\mathbb{P}$ indicates the calculation of a probability and $\mathbb{E}$ the calculation of the expected/average value of the random variable). It is well known (see for example [16,17,18]) that the infimum for $\mathbf{R}^*$ is achieved by the Bayes classifier $g*(x) = 2\,\mathbf{1}_{\{\eta(x)>1/2\}} - 1$.

Unfortunately, we do not know the law of (X, Y), so we cannot calculate $\mathbf{R}(g)$ and must choose our function g another way. We propose to use a strategy called the k-nearest neighbor rule (k-NN) in which we predict the label of a new data object by taking a majority vote over the labels of its k nearest neighbors. To keep this simple, we can consider only odd k. The k-NN rule is a very popular

non-parametric method (see [20], who provides a collection of around 140 important papers). The goal is to find the *best* k_0 in the sense we will now describe. One standard method to find the best k_0 involves simple data-splitting, the n data points we have are split into two sets of l and m points (i.e, $n = l + m$), respectively called the *training set* and *test set* [18]. For each k isn't in document $\{1, 3, \ldots, k_{\max}\}$, we re-predict the labels of the test set via their k nearest neighbors in the training set. We then choose the k_0 which correctly re-predicts the largest fraction of the labels, thus defining our classification rule: $g_n =$ *majority vote over* k_0 *nearest neighbors*. This procedure can be shown to find a rule with an error that converges to the minimum possible error, the Bayes risk $\mathbf{R}^*$ [18].

Implementing k -NN for IMGT/V-QUEST + JCTA

We implemented the described k-NN data-splitting algorithm on the 249 TRAV-TRAJ sequences, of which 212 had the label -1 (correctly analysed) and 37 the label $+1$ (potentially incorrectly analysed), using leave one out to estimate the true error probability of the chosen rule g_n. Each sequence was represented by a vector in $\mathbb{R}^5$, the 5 (out of 10) variables retained being post-trimming 3'V-REGION length, N-REGION length, post-trimming 5'J-REGION length, number of trimmed V nucleotides and number of trimmed J nucleotides, since the other variables are either too rare to help with prediction (P3'V length and P5'J length) or are almost always formed by linear combinations of the 5 retained variables (JUNCTION length, germline 3'V-REGION length and germline 5'J-REGION length). Each of the 249 trials involved removing one sequence, randomly dividing the remaining 248 sequences into two equally sized sets of $l = m = 124$ sequences, then using the test set m on the training set l to automatically select the k_0 that minimises the fraction of incorrectly predicted labels. Then this k_0 was used to re-predict the label of the left-out sequence. The fraction of the 249 trials for which this re-prediction was wrong gave us an estimate of the true error probability of the k-NN rule g_n selected using this algorithm.

Due to potential instability in the data-splitting method, we performed the entire algorithm 500 times in order to not only calculate the average fraction of incorrectly predicted labels, but to show how much the calculated fraction of incorrectly predicted labels varied around this average value. We found the average fraction of incorrectly predicted labels to be 0.1256 with a standard deviation of 0.0097. Whilst at a first glance this seems to be a reasonably good result, we recall that $37/249 = 14.68\%$ of the data has the label $+1$ (i.e, potentially incorrectly analysed). Therefore we could have almost found the same result just by classifying *every* sequence as -1, so we must deepen our analysis to show that we have indeed a useful result. In a way, we are more interested in seeing how good the classification is on the +1 labelled sequences, that is, how good are we at identifying these potentially incorrectly analysed sequences. To examine this, we focus on two important variables, *true positives* (*TP*) and *false positives* (*FP*). A *TP* is when we predict the label $+1$ and the real label is indeed $+1$, whereas a *FP* is when we predict $+1$ but the real label is -1. We simultaneously want the number of *TP*'s, $|TP|$ to be large and the number of *FP*'s, $|FP|$ to be small, though we would like to control the relative importance of these two criteria.

Obviously, by minimising the fraction of incorrectly predicted labels, we have no direct control on making $|TP|$ large and $|FP|$ small exactly in the way we would like. We therefore considered a two-pronged analysis. First we formed empirical estimates of $|TP|$ and $|FP|$ calculated during the previous analysis where we minimised the fraction of incorrectly predicted labels, to see whether nevertheless we find useful results. Indeed, we found for $|TP|$ and $|FP|$ the means 8.4480 and 4.7280, with standard deviations 1.8510 and 1.7286, respectively. This means that on average, we correctly removed 8.4480/37

$= 22.83\%$ of the potentially incorrectly analysed sequences. As well as this, on average $8.4480/(8.4480 + 4.7280) = 64.12\%$ of the sequences we predicted to be potentially incorrectly analysed actually were.

Secondly, we decided to slightly alter the algorithm to, instead of minimising the fraction of incorrectly predicted labels, actually maximise a function of the form

$$f(|TP|, |FP|) = a|TP| - b|FP|,$$

with $a, b > 0$. The choice of a and b will reflect the relative importance we place on *TP* and *FP*. In fact, minimising the fraction of incorrectly predicted labels as we did before is the special case where we set $a = b = 1$. In the present altered algorithm, we chose $a = 1$ and $b = 1/20$ to reflect the fact that we want to especially maximise $|TP|$ and we don't mind too much if $|FP|$ is a little larger than before. This schema is similar to the medical situation where for example in a blood test, we want to maximise the probability of detecting a virus, and simultaneously control probability of incorrectly telling someone they have the virus. We remark also that maximising this function f is equivalent to minimising

$$a \sum_{i=1}^{m} 1_{\{g(X_i) \neq Y_i, Y_i = 1\}} + b \sum_{i=1}^{m} 1_{\{g(X_i) \neq Y_i, Y_i = -1\}},$$

which shows more clearly that maximising f corresponds to minimising a *re-weighting* of the two types of error which make up the incorrectly predicted labels. We implemented this adapted algorithm, and again ran it 500 times to account for potential volatility due to the random data-splitting. We found for $|TP|$ and $|FP|$, the means 14.7020 and 15.1800, with standard deviations 2.0779 and 2.7617, respectively. This means that on average, we correctly removed $14.7020/37 = 39.74\%$ of the incorrectly analysed sequences. As well as this, on average $14.7020/(14.7020 + 15.1800) = 49.2\%$ of the sequences we predicted to be incorrectly analysed actually were. In reality, this means that with this new algorithm we almost double the average percentage of potentially incorrectly analysed sequences we automatically detect from 22.83% to 39.74%, but at the same time, the percentage of the predicted $+1$'s that actually have the real label $+1$ drops from 64.12% to 49.2%. In terms of implementing a warning system in IMGT/V-QUEST + JCTA, it is perhaps preferable to have a larger $|TP|$ as we do in the adapted algorithm, even if it means an increase in $|FP|$, so long as $|FP|$ remains reasonably small. We remark finally that the average fraction of incorrectly predicted labels found for this adapted algorithm was 0.1414, so we have not lost too much in the global quality of the result (the average fraction of incorrectly predicted labels found earlier was 0.1256) when we implement this adapted algorithm.

DISCUSSION

This paper had two important and different goals: the setting up of *Standardization* and the development of an *Automatic Warning System* to detect suspiciously aligned or analysed TRAV-TRAV junction sequences.

For that which concerns standardization, we have defined and given intuitive names to ten variables that we consider important when analysing TRAV-TRAJ junction sequences. As well as this, we have provided detailed statistical results, both graphically and numerically based on a set of 212 TRAV-TRAJ junction sequences which had first been extensively verified, one by one in minute detail by an expert to have been correctly sequenced and analysed by IMGT/V-QUEST + JCTA. This IMGT junction reference set is available on the IMGT® Website (http://imgt.cines.fr) and will serve as a statistical

frame of reference against which repertoires in diverse normal and pathological situations can now be compared, as well as repertoires between species.

Secondly this paper dealt with the creation of an automatic warning system for IMGT/V-QUEST + JCTA, whose goal was to identify suspiciously aligned or analysed sequences. In the current version of IMGT/V-QUEST + JCTA, a sequence entered by the user is aligned and annotated. Current in-built rules allow approximately 85% of entered sequences to be correctly automatically aligned and annotated. For the remaining 15%, problems may occur due to several types of atypical sequences: there may be an unusual N-REGION owing to extensive trimming (upstream of the 2nd-CYS codon and downstream of the J-PHE codon), there may be a new polymorphic site, a J primer may have been used or there was simply a sequencing error. Indeed, it is unrealistic to expect an automatic tool to be able to consider *every* biological possibility. At present, users can simply manually verify the alignment and analysis of their sequence by comparison with IMGT Alignment of Alleles [1,2].

However, the new version of IMGT/V-QUEST will allow treatment of sequences by batches of 50, hence the interest in implementing an automatic warning system that can identify sequences that have been potentially misaligned or misanalysed for whatever reason, rather than expecting users to spend vast amounts of time manually re-checking every sequence's alignment and analysis. This is doubly important if users expect to obtain biologically pertinent results in subsequent analyses they perform using the automatic outputs of IMGT/V-QUEST – otherwise the *noise* due to the incorrectly analysed sequences may have an adverse effect on their subsequent analyses and conclusions.

In this paper we have made a first step in creating such an automatic warning system using a simple k-NN algorithm in $\mathbb{R}^5$. To cast the present result into a practical light, if we suppose someone arrives with 1000 new TRAV-TRAJ junction sequences, we currently expect approximately 150 of them (i.e, 15%) to be potentially incorrectly analysed. With this new k-NN model we can expect to automatically identify about 60 of them, and well as wrongly suspecting around 60 other sequences to be potentially incorrectly analysed. Consequently, we also expect around $150-60 = 90$ potentially incorrectly analysed sequences to be undetected. Whilst this is surely an improvement on having 150 potentially incorrectly analysed sequences remain undetected, we do hope in future work to improve the quality of this warning system, so that we can detect a larger percentage of these 150 suspicious sequences, perhaps by using a larger number of predictive variables or by implementing a more sophisticated algorithm.

We remark next that this work has been especially useful in leading to improvements in IMGT® tools. Indeed, of the 37 sequences which were considered suspicious, 25 would now not be detected as such if the same analysis was done again. In effect, as a result of our analysis, 13 of these sequences allowed us to update 8 entries of the IMGT reference directory of IMGT/V-QUEST and IMGT/JunctionAnalysis, and in addition 11 sequences allowed us to identify without ambiguity 3 new allelic polymorphisms and these new alleles have been added to the IMGT Alignment of Alleles and to the IMGT/V-QUEST and IMGT/JunctionAnalysis tools. For the remaining 13 sequences, 5 correspond to sequencing errors and 8 to potential polymorphisms.

Finally, we note that in future work, it will be of great interest to use the same standardization criteria and to compare these results with non-human data such as mouse data, as well as with T cell receptor TRB, TRG and TRD junctions for both human and non-human data.

ACKNOWLEDGEMENTS

We would like to thank one referee for several pertinent comments and suggestions to improve the manuscript. We are grateful to Xavier Brochet for his contribution to the improvement of the IMGT/V-QUEST software. We thank Gérard Lefranc for fruitful discussion and the IMGT® team for its constant

motivation and expertise. K.B. is the recipient of a doctoral grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR) Université Montpellier II. IMGT® is a registered Centre National de la Recherche Scientifique (CNRS) mark. IMGT® is a National Bioinformatics RIO platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programmes of the European Union and received subventions from Association pour la Recherche sur le Cancer (ARC) and from the Réseau National des Génopoles (RNG), Génopole-Montpellier-Languedoc-Roussillon. IMGT® is currently supported by the CNRS, the MENESR (Université Montpellier II Plan Pluri-Formation, BIOSTIC-LR2004, ACI-IMPBIO IMP82-2004), GIS AGENAE, the Région Languedoc-Roussillon, Agence Nationale de la Recherche (ANR BIOSYS06_135457) and the ImmunoGrid project (IST-2004-028069) of the 6th framework programme of the European Union.

REFERENCES

[1] Lefranc M.-P. and Lefranc G., (2001). The T cell receptor FactsBook, Academic Press, London, UK, 398.
[2] Lefranc M.-P. and Lefranc G., (2001). The Immunoglobulin FactsBook, Academic Press, London, UK, 458.
[3] Giudicelli V. and Lefranc M.-P., (1999). Ontology for Immunogenetics: IMGT-ONTOLOGY, Bioinformatics **15** 1047-1054.
[4] Lefranc, M.-P., Giudicelli V., Ginestoux C., Bosc N., Folch G., Guiraudou D., Jabado-Michaloud J., Magris S., Scaviner D., Thouvenin V., Combres K., Girod D., Jeanjean S., Protat C., Yousfi-Monod M., Duprat E., Kaas Q., Pommié C., Chaume, D. and Lefranc, G. (2003). IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. In Silico Biol. **4**, 0004.
[5] Lefranc, M.-P., Clment, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D. and Lefranc, G. (2004). IMGT-Choreography for Immunogenetics and Immunoinformatics. In Silico Biol. **5**, 0006.
[6] Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clment, O., Chaume, D. and Lefranc, G. (2005). IMGT®, the international ImMunoGeneTics information system. Nucleic Acids Res. **33**, D593-D597.
[7] Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2004). IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. **32**, W435-W440.
[8] 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**, i379-i385.
[9] Landau, N. R., St. John, T. P., Weissman, I. L., Wolf, S. C., Silverstone, A. E. and Baltimore, D. (1984). Cloning of terminal transferase cDNA by antibody screening. Proc. Natl. Acad. Sci. USA **81**, 5836-5840.
[10] Lafaille, J. J., DeCloux, A., Bonneville, M., Takagaki, Y. and Tonegawa, S. (1989). Junctional sequences of T cell receptor $\gamma\delta$ genes: implications for $\gamma\delta$ T cell lineages and for a novel intermediate of V-(D)-J joining. Cell **59**, 859-870.
[11] Lewis, S. M.(1994). P nucleotide insertions and the resolution of hairpin DNA structures in mammalian cells. Proc. Natl. Acad. Sci. USA **91**, 1332-1336.
[12] Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D. and Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. **34**, D781-D784.
[13] Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. **33**, D256-D261.
[14] Chen, P. Y.(2002). Correlation: Parametric and nonparametric measures. Sage Publications, CA, USA.
[15] Cover, T. M. and Thomas, J. A.(1991). Elements of Information Theory. John Wiley & Sons, New York, USA.
[16] Boucheron, S., Bousquet, O. and Lugosi, G.(2005). Theory of classification: A survey of some recent advances. ESAIM: Probability and Statistics **9**, 323-375.
[17] Bousquet, O., Boucheron, S. and Lugosi, G.(2004). Introduction to statistical learning theory. *In*: Advanced Lectures in Machine Learning, Bousquet, O., Luxburg, U. V. and Rätsch, G. (eds), Springer, New York, USA, pp. 169-207.
[18] Devroye, L., Györfi, L. and Lugosi, G.(1996). A Probabilistic Theory of Pattern Recognition. Springer, New York, USA.
[19] Duda, R. O., Hart, P. E. and Stork, D. G.(2000). Pattern Classification. Wiley Interscience, New York, USA.
[20] Dasarathy, B. V.(1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE CS Press, Los Alamitos, USA.