SAIVGeM: spreadsheet analysis of immunoglobulin V_H gene mutations

Bradley T. Messmer

Institute for Medical Research, Manhasset, NY, USA

BioTechniques 39:353-358 (September 2005)

The analysis of mutations in immunoglobulin heavy chain variable (IGHV) region genes is a tedious process when performed by hand on multiple sequences. This report describes a set of linked Microsoft® Excel® files that perform several common analyses on large numbers of IGHV sequences. The spreadsheet analysis of immunoglobulin V_H gene mutations (SAIVGeM) package determines the distribution of mutations among each nucleotide, the nature of the mutation at both the nucleotide and amino acid level, the frequency of mutation in the A/G G C/T A/T (RGYW) hotspot motifs of both strand polarity, and the distribution of replacement and silent mutations among the complementarity determining regions (CDRs) and the framework regions (FRs) of the immunoglobulin gene as defined by either the Kabat or IMGT conventions. These parameters are summarized and graphically presented where appropriate. In addition, the SAIVGeM package analyzes those mutations that occur in third positions of redundant codons. Because any nucleotide change in these positions is inherently silent, these positions can be used to study the mutational spectra without biases from the selection of protein structure.

INTRODUCTION

Antibodies bind to antigens via the variable domains of the immunoglobulin heavy and light chains. These variable domains are composed of V, D, and J genes that are recombined from a larger repertoire during the development of a mature B lymphocyte (1). During an immune response, the variable region genes of antibodies with affinity for the antigen acquire somatic mutations. B lymphocytes expressing higher affinity antibodies arising from these mutant sequences are selected and expanded in a process termed affinity maturation (2).

The mutational process during affinity maturation is not random (3). In addition to the elevated rate of mutation, there are known biases in targeting, particularly to certain hotspots such as the A/G G C/T A/T (RGYW) motif (4), and the nucleotide changes show a bias toward transitions over transversions (5) that is somewhat sequence context-dependent (6). There is also selection for mutations that will enhance the antigen binding affinity while maintaining the structural integrity of the variable domain. Thus,

the regions of the antibody molecule that control the interaction with the antigen, the complementarity determining regions (CDRs), are enriched in mutations that lead to amino acid replacements while the more structural regions, termed framework regions (FRs), are enriched for silent mutations (7).

The inherent susceptibility of the various regions of the immunoglobulin molecule for mutation and the mechanistic basis for the observed mutational spectra is not completely defined, so empiric comparisons between groups of sequences are often used to establish if they are a product of the same set of processes (8). This is becoming a particularly important issue in many B cell-derived malignancies. For example, chronic lymphocytic leukemia (CLL) patients have a better clinical outcome if the leukemic B cell has an antibody gene with mutations (9,10), but the pathophysiologically basis of this phenomenon is unknown.

Studies of this and other B cell malignancies as well as studies of normal and autoimmune B cell populations generate large numbers of immunoglobulin gene sequences.

There are several web-based tools for the analysis of single V_H sequences [VBASE, vbase.mrc-cpe.cam.ac.uk; IgBLAST (Immunoglobulin Basic Local Alignment Search Tool), www.ncbi.nlm.nih.gov/igblast; IMGT/V-QUEST, imgt.cines.fr/textes/vquest (11); VBASE2, www.dnaplot.de/vbase2 (12)], but there are currently no publicly available tools for the simultaneous analysis of many sequences. This report describes a series of linked Microsoft® Excel® spreadsheets that calculate several commonly ascertained parameters of the mutations in the V_H gene segment of the immunoglobulin heavy chain variable (IGHV) region. The spreadsheets do not use macros or Visual® Basic® and require no alterations to default settings.

MATERIALS AND METHODS

The SAIVGeM package consists of six linked Excel spreadsheets entitled *sequencedata.xls*, *ntmutation.xls*, *aamutation.xls*, *RGYW.xls*, *RS.xls*, and *summary.xls*. These files should all be located in the same folder and the entire folder copied for each analysis performed. The sequences to be analyzed are copied into the *sequencedata.xls* file, and each of the other files are sequentially opened and updated. The *summary.xls* file contains several summaries and graphical representations of the data. Each module is described below. The spreadsheets were created in Excel 2002 for the Microsoft Windows™ 2000 operating system but should be compatible with other versions (not tested). The files may be found at cbl.ucsd.edu/messmer/messmer.htm.

Data Format

The sequences are entered into the sequences worksheet of the *sequencedata.xls* workbook in a vertical format, one sequence per column, where each cell is a single nucleotide position (example in Figure 1). Each column for sequence entry requires the name of the nearest germline gene to be in row 1, in IMGT allele format (13) (IGHVx-y*z, where x-y is the numerical name of the V_H germline gene and z is the

allele number in a two-digit format, such as IGVH4-34*02). The sequence name or identifier is in row 2, followed by the sequence in rows 3–305. The sequences must be vertically aligned such that the first nucleotide of the first codon (for glutamic acid or glutamate) is in row 3. Leader sequences are not analyzed and should be trimmed. If the 5' end of a V_H sequence is incomplete, the sequence should be entered starting at the appropriate position (e.g., if the first 10 nucleotides of the sequence are missing, then the sequence should be entered beginning in row 13). The TGT codon encoding the cysteine at the end of FR3 represents the last nucleotides that should be included.

Groups of sequences can be appropriately formatted from multiple alignments through a series of simple steps. First, an alignment of all V_H sequences of interest must be created such that the relative starting positions are appropriately spaced. If all sequences begin with the first nucleotide of the V_H coding sequence, there is no need to space them further. The sequence identifiers should be on the left. Ideally, the alignment should contain a space, or some other delimiting character, in between each nucleotide (otherwise, it can be added with a text editor later), and the sequences should be broken every 175 nucleotides. This is because Excel will only allow the creation of 256 columns of data, and the sequences must initially be entered in a horizontal series of cells that will be subsequently transposed. The alignment should be saved as a text file (.txt) and opened from within Excel. The text import wizard should launch. On the first page of the wizard, the original data type box for “delimited” should be checked and the first line with a sequence should be indicated in the “Start import at row:” dialog box. After selecting “Next,” the “Space” box should be checked under “delimiters,” and the “Treat consecutive delimiters as one” box checked, followed by the “Finish” button. The sequences should now be imported into Excel in a horizontal alignment. The sequences are cut and pasted with the “transpose” option selected into a new worksheet to create the vertical alignment for copying into the sequence worksheet of the sequencedata.xls file.

sequencedata.xls

This workbook contains three sheets: sequences, germline match, and germline. The sequences worksheet is where the sequence data are input as described above. In addition, the total number of each nucleotide in each input sequence is calculated (rows 308–312) for later use. This version of SAIVGeM can analyze 100 sequences, but more can be handled by copying the columns in all analysis sheets to as many new columns as is required. However, this approach may cause difficulty as the memory requirements become substantial, and Excel may not function properly. Alternately, separate SAIVGeM folders can be made for each 100 sequence group and the data pooled after analysis.

The germline match worksheet calls the nucleotide sequence of the most similar germline gene (indicated on row 1 of sequences) at each position for which there are sequence data. It should be noted that the germline gene and identifier information in rows 1 and 2 are linked on this and all other worksheets to the entries in rows 1 and 2 of the sequence worksheet and do not need to be separately entered. The germline worksheet is the complete IMGT allele database (13) represented vertically and is the source of data for the germline match sheet. In addition, sequences containing deletion or insertions will not be properly aligned. In that case, the remaining mutations can be analyzed by filling in the sequence in the case of a deletion or excising the inserted sequence.

ntmutation.xls

This workbook contains four sheets: nt match, silent redundant, germ sr, and germline sr database. The nt match worksheet determines where there are differences between the sequence and the most similar germline gene. The formula in each cell of the data array (B3:CZ305) first checks if there are sequence data for the given position in the sequences sheet of the sequencedata.xls file. If there is no sequence, the formula leaves the cell blank. If sequence is present, it then checks the germline match sheet to determine if

the nucleotide at that position is the same as the germline. If so, it returns an “x.” If not, it concatenates the germline nucleotide and the sequence nucleotide to indicate the nature of the mutation; that is, GC indicates that the germline G was replaced with a C. Rows 308–319 sum each of the 12 possible types of mutation, and row 320 indicates the total number of mutations in each sequence. When a new project is begun, the nt match sheet should be carefully inspected for any obvious errors of alignment or germline gene assignment. Row 321 represents the mutation as a percentage of the nucleotides in the sequence. Row 322 indicates the percent similarity to the germline sequence. The total number of each type of mutation and the total mutations for all sequences are calculated in cells B326–B338.

The germline sr database sheet is a reference sheet that indicates, for each germline gene and allele, which nucleotides are inherently redundant because they are the third position of synonymous codons. Any mutation at these positions will be silent. As such, these positions may be used to discern features of the mutational mechanism without a concern that selection for amino acid sequence is biasing the results. The germ sr sheet identifies which positions in the analyzed sequences are these inherently silent redundant positions. If so, the formula returns the germline value of the position; otherwise, it returns a blank cell. Rows 308–311 sum the number of

	A	B	C	D	E	F	G	H
1	IGHV1-2*01	IGHV3-11*01	IGHV3-11*01	IGHV3-30*01	IGHV5-51*01	IGHV5-51*01	IGHV5-10*01	IGHV5-10*01
2	position	1	2	3	4	5	6	7
3		A	G	G	C	G	G	C
4		A	A	A	A	A	A	A
5		G	G	G	G	G	G	G
6		G	G	G	G	G	G	G
7		T	T	T	T	T	T	T
8		C	C	C	C	C	C	C
9		C	C	C	C	C	C	C
10		A	A	A	A	A	A	A
11		G	G	G	G	G	G	G
12		C	C	C	C	C	C	C
13		T	T	T	T	T	T	T
14		G	G	G	G	G	G	G
15		G	G	G	G	G	G	G
16		T	T	T	T	T	T	T
17		G	G	G	G	G	G	G
18		G	C	C	C	C	C	C
19		A	A	A	A	A	A	A
20		G	G	G	G	G	G	G
21		T	T	T	T	T	T	T
22		C	C	C	C	C	C	C
23		T	T	T	T	T	T	T
24		G	G	G	G	G	G	G
25		G	G	G	G	G	G	G
26		G	G	G	G	G	G	G
27		G	G	G	G	G	G	G
28		G	G	G	G	G	G	G
29		A	A	A	A	A	A	A
30		A	A	A	A	A	A	A

Figure 1. Data entry worksheet. The sequences worksheet from the sequencedata.xls file is shown.

germline redundant positions consisting of each nucleotide.

The silent redundant sheet determines which silent mutations, calculated from nt match, occurred in silent redundant positions and returns the concatenation of the germline and the sequence nucleotide, as for nt match. The mutations are summed in rows 308–320 and the page sums calculated in cells B325–B337 as with nt match.

aamutation.xls

This workbook consists of four worksheets: aa, germline aa, aa change, and genetic code. The sheets aa and germline aa translate the sequences and the most similar germline counterparts to single-letter representations of the amino acids using the human genetic codon from the genetic code sheet. The aa change sheet functions much like the nt match sheet, returning a hyphen when the translated sequence amino acid is the same as the germline and concatenating the germline and sequence amino acid codes if they are different. If a silent mutation occurred in a nucleotide of the codon, an “X” is put in the appropriate cell. At the bottom of this sheet, a matrix of the replacement mutations is calculated, and the total number of conservative [as defined by polarity and volume (C), (AGPST), (RKH), (EDQN), (ILVM), and (FWY) (14)] and nonconservative replacements.

RGYW.xls

The workbook RGYW.xls determines which mutations occurred in

RGYW (R = AG, Y = CT, and W = AT) motif locations. There are three sheets: RGYW, germline RGYW, and germline RGYW potential. The sheet germline RGYW potential is a reference database that indicates, for each germline gene, which nucleotides reside in an RGYW motif on the template (+), nontemplate (-), or both strands (+/-). The germline RGYW sheet determines which positions in the sequences being analyzed were located in RGYW motifs. The RGYW sheet annotates the strand as above at every site where a mutation (from the nt match worksheet of the ntmutation.xls file) was recorded in an RGYW motif. At the bottom of the sheet, the total number of mutations occurring in RGYW motif on each or both strands is summed, as well as the total number of RGYW motif present on each or both strands in the germline gene.

RS.xls

This workbook consists of three analysis sheets (RS, RS IMGT pos, and RS positional sums), four graph sheets (R graph, S graph, RS graph, and RS windows graph), and one reference data sheet (IMGT spacing). The RS sheet determines if each mutation from the nt match sheet of ntmutation.xls resulted in an amino acid replacement (R) or was silent (S). The RS IMGT pos sheet adjusted the positions of the mutations from the RS sheet to match the IMGT position convention (15). It does this by referencing the IMGT spacing sheet that contains a code for the number of spaces at the ends of CDR1 and CDR2 in the IMGT format for each germline gene. That code is placed in row 3. The data array checks the code to determine the proper IMGT position of each cell in the RS sheet data array. At the bottom of the sheet, the total number of R and S mutations is calculated for each sequence as well as for the CDRs and FRs defined by either the Kabat or IMGT designations.

Several graphs are created, all using the data in the RS positional sums sheet. The RS positional sums sheet sums the R and S mutations at each IMGT position. In addition, it calculates a local R to S ratio for each position by considering a window of 7, 9, or 11 nucleotides up and downstream. The R graph and S graph sheets contain graphs of the number of R and S mutations at each position, respectively. The RS graph sheet combines the latter two, and the RS windows graph sheet graphs the local R to S ratios from the windows calculated as described above.

summary.xls

The summary.xls file consists of four worksheets that summarize various aspects of the other files in the package. The summary data collects the nucleotide and mutation information from the various files and calculates mutation frequencies as a function of nucleotide, the transition to transversion ratio, the overall percentage of mutations occurring in RGYW motifs in total and on each strand, and the overall percentage of RGYW motifs in the germline sequences used by the data set. The mutation stats sheet collates the overall mutation statistics. The seq stats sheet determines and graphs the distribution of sequences by the IGHV subgroup and gene used as well as the distribution of mutations among the sequences. The nt stats sheet provides a more detailed summary of the mutations as they distribute among the nucleotides.

File Links and Optimization

The links between the various files of the SAIVGeM package are shown in Figure 2. Depending on the power and memory of the computer being used, it may not be reasonably possible to open all of the files in the SAIVGeM package simultaneously. In this case, there are several remedies. First, the files may be opened, updated, and then closed sequentially according to their links. The operation may be faster if all of the linked files are already open. Alternately, the links between files can be broken once updated, thus significantly reducing the memory require-

Figure 2. Linked relationship between files in the SAIVGeM folder. The arrows point in the direction of the information flow. SAIVGeM, spreadsheet analysis of immunoglobulin V_H gene mutations.

ments but eliminating the ability to change the data from upstream links. The data arrays may also be frozen by copying the data and re-pasting it in with only the values, thus eliminating the formulas that calculated the data and significantly reducing file sizes and memory requirements. Finally, unused columns that contain formulas can be deleted.

Sample Data Set

Fifty-one immunoglobulin sequences from persistent polyclonal B lymphocytosis (PPBL) patients (16) were downloaded from GenBank® (accession nos. AYO33238–AYO33288) and formatted for SAIVGeM analysis using the Megalign™ v5.06 sequence alignment software (DNASar, Madison, WI, USA).

RESULTS AND DISCUSSION

The SAIVGeM package was used to analyze a set of 51 immunoglobulin sequences from a published study of patients with PPBL (16). The most similar germline IGHV gene was indicated in the original study and annotated at the top row of the sequences sheet of the sequencedata.xls file. The sequences were appropriately formatted and copied into the sequencedata.xls file of a new copy of the SAIVGeM folder. Each file of the SAIVGeM package was opened and updated. Selected results contained in the summary.xls file are shown in Figure 3. The full files can be found at cbl.ucsd.edu/messmer/messmer.htm.

The SAIVGeM package simplifies and automates the analysis of mutational characteristics in V_H genes, using the widely used Excel program. There are several advantages to this type of batch analysis. Foremost is the accuracy of the calculations. Once the sequence data are entered into the sequencedata.xls file, all other operations are performed within Excel so there is no opportunity for human error. Because the data at each level of analysis are preserved in a spreadsheet, other custom analyses, statistics, or graphing can easily be performed or

the data readily exported to another program.

On a conceptual level, the analysis of the redundant codon positions provides an opportunity to discriminate the biases introduced by the mutational mechanism and the effects of selection. Immunoglobulin sequence analysis often seeks to determine if there is significant evidence for antigen selection. This is usually analyzed for each sequence separately, using models that assume a degree of randomness

in the mutational process (17,18). However, these methods are quite limited because the few mutations in a given sequence lack statistical power to allow a confident determination for most sequences, and the underlying assumptions of randomness may be flawed (19). By considering groups of sequences, the results produced by SAIVGeM should be more amenable to statistically significant conclusions.

Figure 3. Sample data contained in the summary.xls file. (A) The mutation stats sheet summarizes the mutation in the entire data set. The FR and CDR R: ratios are determined according to both the Kabat and IMGT conventions. (B) The seq stats sheets displays the V_H gene and mutation distributions among the data set. (C) The nt stats sheet summarizes the mutational spectra at the nucleotide level for all mutations and for those that occur in the third position of redundant codons. FR, framework region; CDR, complementarity determining region; nt, nucleotide.

ACKNOWLEDGMENTS

The author is a Lymphoma Research Foundation Fellow. The author thanks Nicholas Chiorazzi for significant discussions of the work and Davorka Messmer and Tom Kipps for comments on the manuscript.

COMPETING INTERESTS STATEMENT

The author has no competing interests.

REFERENCES

1. **Lefranc, M.-P. and G. Lefranc.** 2001. The Immunoglobulin Factsbook. Academic Press, London.
2. **Wagner, S.D. and M.S. Neuberger.** 1996. Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.* 14:441-457.
3. **Neuberger, M.S., M.R. Ehrenstein, N. Klix, C.J. Jolly, J. Yelamos, C. Rada, and C. Milstein.** 1998. Monitoring and interpreting the intrinsic features of somatic hypermutation. *Immunol. Rev.* 162:107-116.
4. **Betz, A.G., M.S. Neuberger, and C. Milstein.** 1993. Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes. *Immunol. Today* 14:405-411.
5. **Neuberger, M.S. and C. Milstein.** 1995. Somatic hypermutation. *Curr. Opin. Immunol.* 7:248-254.
6. **Cowell, L.G. and T.B. Kepler.** 2000. The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation. *J. Immunol.* 164:1971-1976.
7. **Dorner, T., H.P. Brezinschek, R.I. Brezinschek, S.J. Foster, R. Domiati-Saad, and P.E. Lipsky.** 1997. Analysis of the frequency and pattern of somatic mutations within nonproductively rearranged human variable heavy chain genes. *J. Immunol.* 158:2779-2789.
8. **Messmer, B.T., E. Albesiano, D. Messmer, and N. Chiorazzi.** 2004. The pattern and distribution of immunoglobulin VH gene mutations in chronic lymphocytic leukemia B cells are consistent with the canonical somatic hypermutation process. *Blood* 103:3490-3495.
9. **Damle, R.N., T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S.L. Allen, A. Buchbinder, D. Budman, et al.** 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94:1840-1847.
10. **Hamblin, T.J., Z. Davis, A. Gardiner, D.G. Oscier, and F.K. Stevenson.** 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 94:1848-1854.
11. **Giudicelli, V., D. Chaume, and M.P. Lefranc.** 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-W440.
12. **Retter, I., H.H. Althaus, R. Munch, and W. Muller.** 2005. VBASE2, an Integrative V gene database. *Nucleic Acids Res.* 33 Database Issue:D671-D674.
13. **Giudicelli, V., D. Chaume, and M.P. Lefranc.** 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33 Database Issue:D256-D261.
14. **Zhang, J.** 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* 50:56-68.
15. **Lefranc, M.P., C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc.** 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27:55-77.
16. **Salcedo, I., A. Campos-Caro, A. Sampalo, E. Reales, and J.A. Brieva.** 2002. Persistent polyclonal B lymphocytosis: an expansion of cells showing IgVH gene mutations and phenotypic features of normal lymphocytes from the CD27+ marginal zone B-cell compartment. *Br. J. Haematol.* 116:662-666.
17. **Chang, B. and P. Casali.** 1994. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunol. Today* 15:367-373.
18. **Lossos, I.S., R. Tibshirani, B. Narasimhan, and R. Levy.** 2000. The inference of antigen selection on Ig genes. *J. Immunol.* 165:5122-5126.
19. **Dunn-Walters, D.K. and J. Spencer.** 1998. Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology* 95:339-345.

Received 9 February 2005; accepted 10 March 2005.

Address correspondence to Bradley T. Messmer, University of California, San Diego Cancer Center, 3855 Health Science Drive #0820, La Jolla, CA 92093-0820, USA. e-mail: bmessmer@ucsd.edu

To purchase reprints
of this article, contact
apfeffer@BioTechniques.com