Providing an automatically derived high quality immunoglobulin V gene sequence database

Ida Retter[1], Werner Müller[2]

Keywords: sequence alignment, secondary database, automatic generation, immunoglobulins

1 Introduction.

With the exponential growth of the primary nucleotide sequence databases GenBank, DDBJ and EMBL [1] the requirement of automatically generated and annotated secondary sequence databases arises. Our aim is to generate an immunoglobulin nucleotide sequence database derived from the EMBL database in an automatic approach, representing the immunological sequence spectrum present in the germ-line. Within an immunoglobulin gene locus there are different sequence configurations possible: The germ-line configuration, in which multiple gene elements, mainly the so-called V genes, constitute the potential diversity of the antibody molecule, and the rearranged configuration, in which one V gene has been recombined with one or two other gene segments to create a functional antibody coding sequence (for review, see [2]). These rearrangements may or may not include somatic point mutations and deletions. To separate the germ-line encoded from somatically mutated immunoglobulin sequences we use two strategies: On the one hand, we compare V gene sequences from the EMBL database with genomic BAC sequences and regard a 100% match as a germ-line evidence. On the other hand, all rearrangements from the EMBL database are aligned and V genes that are found in at least two independent rearrangements are regarded as germ-line sequences. To maximize data reliability our database does not include information from the annotation part of the EMBL nucleotide entries but provides accurate sequence evaluation by sequence comparison. The program was developed with sequences from the murine immunoglobulin heavy chain locus. However, it can also be applied to the light chain loci, other types of sequences (e.g., T cell receptor genes) and other species.

2 Methods.

Sequence alignment with BLAST. In order to identify all immunoglobulin sequences within the EMBL database we use the BLAST algorithm [3]. Beside the standard subset the High Throughput Genomic Sequence (HTG) and the WGS (Whole Genome Shotgun) databases are included in the search. A number of known immunoglobulin sequences are used as initial query sequences [4]. The BLAST result is subsequently filtered for minimum sequence identity and minimum alignment length.

Sequence alignment with DNAPLOT. DNAPLOT is a sequence alignment program tailored for immunoglobulin nucleotide and protein sequences [5]. The fast alignment algorithm allows sorting the sequences within a multiple alignment and comparing multiple alignments among each other. Furthermore, the DNAPLOT motif recognition functions are used for the automatic V gene annotation.

[1] Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: ida.retter@gbf.de

[2] Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: wmueller@gbf.de

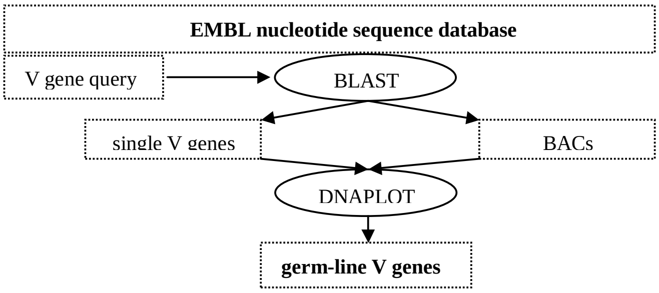

Figure 1: Extraction of V gene sequences from the EMBL database by the BLAST program and germ-line V gene selection by the DNAPLOT program.

3 Results and discussion.

In our database we can classify the V genes into three different quality groups. The first group consists of sequences found in a rearranged form as well as in a non-rearranged configuration. These are germ-line sequences of actively used V gene segments. V gene sequences of the second group are only found as germ-line genes but not recovered in V gene rearrangements. Such V gene segments may represent pseudogenes and the reason for the non-functionality of such sequences might be determined

by a subsequent analysis. The third group of V genes is only found in rearranged sequence list but not in the list of non-rearranged sequences. These V genes represent most likely germ-line genes. However, a little bit of uncertainty remains until in a future generation of the database, a non rearranged counterpart can be recovered from the EMBL nucleotide database.

Due to the automatic generation our sequence data set can be updated any time. It is comprehensive as it takes all published immunological sequences into account. The addition of new entries into the EMBL database will continuously improve the resulting V gene database. In turn, our database provides an important tool for the annotation of genomic sequences of the mouse. The method can be easily adapted to other variable loci, thereby providing the opportunity to analyse species with poor sequence data availability.

4 References and Websites.

[3] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

[2] Honjo, T. and Alt, F.[ed.] 1995. *Immunoglobulin Genes.* London: Academic Press.

[5] http://www.dnaplot.org

[1] Kulikova T. et al. 2004. The EMBL nucleotide sequence database. *Nucleic Acids Research* 32:D27-30.

[4] Lefranc, M.P., Giudicelli,V., Ginestoux,C., Bodmer,J., Muller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbie,V. and Chaume,D. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* 27:209-212.