Extracting human antibody sequences from public databases for antibody humanization: high frequency of species assignment errors

(Short Communication)

Andrew C.R. Martin[1] and Anthony R. Rees[2]

[1]Institute of Structural and Molecular Biology, Division of Biosciences, Darwin Building, University College, London, Gower Street, London WC1E 6BT, United Kingdom

[2]Rees Consulting AB, Näs Focksta 19, 75591 Uppsala, Sweden

Contact: rees@reesconsultingab.com –or– andrew@bioinf.org.uk

Running title: Errors in antibody species annotation

April 25, 2016

Abstract

In antibody humanization, CDRs from a 'donor' antibody are often grafted onto a human framework selected by high sequence identity with the donor. In our own humanization experiments, we have found that species information is often incorrect. Here we take three mouse antibodies and perform BLAST searches against sequences annotated as being human. We find that the first genuine human hits for the six chains appear at positions 30, 4, 11, 24, 18 and 29 in the hit lists. This illustrates both the need for caution in performing humanization and for improvements in annotation.

Keywords

Sequence databases; Annotation errors; Database searches

Since the 1980s, the humanization of antibody sequences has become a necessary prerequisite to the development of therapeutic antibodies derived from monoclonals or antibody libraries of rodent (mouse or rat) or other species (e.g. rabbit). In general, the approach to humanization is to graft the 'complementarity determining regions' (CDRs) from a donor antibody (the non-human antibody having the required specificity) onto a human acceptor framework. The earliest humanization attempts made use of one of the few existing human antibodies (Fab fragments or Bence-Jones light-chain dimers from myeloma patients) whose structures were known. Initially, limited framework substitutions in the human acceptor framework were necessary to restore binding to the humanized rodent variable region. The low level of substitution required (only one framework change in the heavy chain during humanization of CAMPATH (Riechmann *et al.*, 1988) and a single heavy chain residue contact identified, but not implemented, in an anti-lysozyme humanization (Verhoeyen *et al.*, 1988)) was likely due to a stroke of good luck since subsequently, other rodent antibodies have required a significant number of framework changes to restore complete binding (Queen *et al.*, 1989, for example).

Since the early humanization experiments by the Winter group, several academic and biotechnology company groups have published methods that range from targeting framework residues within a certain distance of a CDR residue to essentially 'look and see' methods that try to make best guesses at potentially important CDR/framework interactions by inspection of the x-ray structure, but which are often limited by lack of structural data (Tsurushita *et al.*, 2005). Numerous methods have been proposed involving subtle nuances on existing protocols as well as more radical approaches. These include molecular modelling to identify necessary framework changes (Queen *et al.*, 1989; Carter *et al.*, 1992), guided selection using phage display (Jespers *et al.*, 1994), variable region 'resurfacing' (Roguska *et al.*, 1994), 'superhumanization' (Tan *et al.*, 2002; Mader and Kunert, 2010), germline humanization (Pelat *et al.*, 2008; Bernett *et al.*, 2010), methods based on comparing rodent and human sequence and structural data (Covaceuszach *et al.*, 2012), consensus frameworks (Couto *et al.*, 1995, for example), use of consensus framework positions shown to be critical for

CDR conformation and for which 'back to mouse' data were publicly available (Haidar *et al.*, 2012), co-optimization of CDRs and framework (Wu *et al.*, 1999), redesign of the CDRs rather than making framework changes (Hanf *et al.*, 2014), and others. In addition, web-based tools to aid humanization strategies are available such as those of Martin (`www.abysis.org`) and Olimpieri *et al.* (2015).

A requirement for most, if not all, the humanization methods described in the literature is that human heavy and light chain sequences that are potential acceptors for the donor CDRs are actually 'human'. Antibody sequences are submitted to a number of sequence databases and are supposedly checked and correctly annotated. Until July 2000, the most accurate source of such sequence annotation was the Kabat 'Sequences of Immunological Interest' – an online update of the classic book (Kabat *et al.*, 1991). Currently, antibody sequences are contained in IMGT, Genbank/EMBL-ENA/DDBJ (and their protein translations, Genpept and UniProtKB/trEMBL) and the Protein Databank. With the exception of IMGT, these are general resources containing DNA and protein sequences and structures. Resources such as SACS (Allcorn and Martin, 2002), AbDb (Ferdous & Martin, in preparation) and SabDab (Dunbar *et al.*, 2014) extract antibody information from the PDB, while EMBLIG (Couch, Porter, Swindells and Martin, in preparation) does the same for EMBL-ENA data. It is worth noting that UniProtKB/SwissProt, probably the best annotated protein sequence database, does not contain antibody sequences beyond a few examples.

Examples of 'header' information retrieved from such databases are shown in Figure 1. From the 'organism' assignment, the reader can determine whether or not the sequence is human, rodent or from another species. Or can they? An equally important question is whether these data can be extracted automatically and reliably by a computer program.

In our own humanization studies (which use a combination of structure, overall sequence identity, unusual framework sequence, unusual residues at particular positions and CDR/framework contact

analysis) when performing BLAST searches against supposedly human sequences from KABAT, EMBL-ENA and the PDB, we have observed systematic errors in species assignment. As with many approaches, we start by identifying a human acceptor sequence with high sequence identity to the (often mouse) donor sequence (Queen *et al.*, 1989). Thus a BLAST search is performed, using the donor sequence of interest, against a database of human sequences. Inevitably, if the donor sequence is mouse and there are any mouse sequences mis-annotated as being human, these will be extracted at the top of the BLAST hit list.

Here we present three examples where we have taken well-known publicly-available mouse antibody sequences – Gloop2 (Darsley and Rees, 1985), HyHel-5 (Smith-Gill *et al.*, 1982) and 4D5 (Carter *et al.*, 1992) – and searched them against a database of antibody sequences annotated as being human. A local BLAST database was created containing sequences annotated as human from Kabat, IMGT, EMBLIG (derived from EMBL-ENA) and the PDB. UniProtKB/SwissProt could not be used because of its lack of antibody sequences. Counts of sequences from the different sources are provided in Supplementary File `DatabaseContent.pdf` Extraction of appropriate sequences from these resources relied on a protein chain being annotated as being human, but excluding cases of chains that have species information clearly indicating that they are chimeric or were a synthetic construct (e.g. PDB entries 1BBJ, 1BVL, 4XTR). BLASTP was then used to search the mouse antibody sequences against this database.

The hits obtained from such BLAST searches were cross-checked by a further online BLAST search of each hit against the IMGT Domain Display reference sequences (`www.imgt.org/BlastSearch`) and by calculation of 'humanness' scores (Abhinandan and Martin, 2007). Further manual investigation of the original source papers was then carried out.

Analysis of the BLAST results

The BLAST search results, showing the first two pages of hits for the three antibodies, are shown in Supplementary File `BlastHits.pdf`.

For the Gloop2 heavy chain, the top two hits are from PDB entry 3DGG, a mouse monoclonal antibody produced in a human embryonic kidney cell line (HEK 293T). The third hit, PDB entry 3D85, is a chimeric antibody with mouse V-regions and human constant regions (Beyer *et al.*, 2008) as is the 20th hit, CAT05563.1 from patent WO2006126069. The fourth–eighth hits are humanized mouse antibodies from patents WO2010061360 and WO2008047242. The ninth and 10th hits are from PDB entry 1AXS, a mouse catalytic monoclonal antibody (Ulrich *et al.*, 1997), while the 11th and 12th hits are from PDB entry 3DIF, another mouse antibody expressed in HEK 293T cells. Overall, the first 27 entries are identified as mouse when searching against the IMGT reference sequences. The 28th and 29th entries, CAI54212.1 and K049980, are ambiguous when searched against the IMGT reference sequences. For CAI54212.1, the top hit in the IMGT reference set is mouse, the next two are human and the fourth is from *Macaca mulatta*. For K049980, the top four hits in the IMGT reference set are all mouse, but the fifth is human. Checking the original literature (CAI54212.1: Patent EP1491632A2; K049980: Kashmiri *et al.* (1995)), it turns out that both are humanized antibodies. The first true human hit is 30th in the list from PDB entry 3HC0.

For the Gloop2 light chain, the top two entries are both mouse catalytic monoclonal antibodies. The third entry is a mouse monoclonal antibody against digoxin (Lemeulle *et al.*, 1998). The first *bona fide* human light chain is the fourth hit (JX027410).

For the HyHEL-5 heavy chain, the top two hits are from PDB entry 3DIF, a mouse monoclonal. The next sequence, 4UOM is reported in the literature to be the human antibody F5, but we believe it may be a mouse sequence (see below). It is described in the same paper as humanized antibody 4B4C-4 (hit number 46, chain A from PDB entry 4UOK). The next seven hits are all mouse

monoclonals. The first true human antibody is the 11th hit, 4QCI which is derived from the HuCal Gold human library, but this is followed by 3SQO, a humanized mouse antibody (J16).

For the HyHEL-5 light chain, the first nine hits are mouse monoclonal antibodies. The 10th hit, 4UOK, is the humanized light chain partner of the heavy chain discussed above. The next hit (Chain L from PDB entry 4IDJ) is a human sequence (Foletti *et al.*, 2013), but this seems to be a rather unusual sequence since the next 12 hits all appear to be mouse. Using the IMGT reference set, the source of the next hit, CBM42819.1 which is 20th in the list, is ambiguous. The top IMGT reference set hit is mouse, the second is human and the next six are mouse. The sequence comes from patent WO2010052556 and is, in fact, a humanized antibody. The next (21st) hit (chain L from PDB entry 1MIM) is a chimeric antibody with mouse variable domains while the 22nd (CAT05561.1) is mouse and the 23rd is 4UOM which, as above, is described in the literature as human, but which we believe may be mouse. The next hit, chain X from PDB entry 4K7P, appears to be the first genuine human hit appearing 24th in the list.

For the 4D5 heavy chain, none of the 17 top entries is from a *bona fide* human antibody. The first two hits (2RCS and 1AJ7 from the PDB) are mouse catalytic antibodies (to establish this requires following up second order references). In fact the sequence annotation for both 1AJ7 and 2RCS indicates (incorrectly) that the heavy chain is human and the light chain is mouse. Of the remaining hits in the top 17, 1FVC, 1FVD and 1FVE are humanized versions of mouse antibody 4D5; 4HKZ, 4UB0 are mutational studies on existing antibodies cetuximab (a chimeric antibody having a mouse variable region) and trastuzumab (a humanized mouse antibody); 3BE1, 4IOI and 4HJG are also the humanized version of mouse antibody 4D5, trastuzumab; and 1HKL and 1GAF are mouse catalytic antibodies. The 16th hit, CS490801.1, is a humanized mouse antibody. For the 17th hit, CAM57950.1, the assignment was ambiguous based on a search against the IMGT reference sequences: the first hit was a mouse sequence while the next three were all human. Like the other

ambiguous hits described above, it also turns out to be a humanized mouse antibody. The first true human hit is chain A from PDB entry 3B9V coming 18th in the list.

Finally, for the 4D5 light chain, the top two hits from mouse PDB entry 3DIF appear to be another case which has been mis-annotated as human because it was expressed in HEK 293T cells. This is followed by 10 sequences which are humanized mouse sequences, three PDB entries (1FVC, 1FVD and 1FVE) from structural studies of humanized 4D5 sequences (trastuzumab) and PDB entry 4UB0 (chimeric cetuximab). The next hits (17th and 18th in the list), AX375917 and CAD26809, are from a patent that describes both chimeric and humanized sequences. The 19th to 22nd hits (1FVE and 1FVD from the PDB) are also humanized. This is followed by six mouse, humanized and chimeric sequences before the first human hit, chain L from PDB entry 3GRW, which is ranked 29th in the list.

What can be seen from the typical database headers shown in Figure 1 and the examples discussed above is that the unsuspecting BLASTer will receive a long list of sequences for each antibody chain frequently misleadingly designated as originating from *Homo sapiens*. Selection of any of these sequences on the assumption they are human could have serious consequences if the antibody is on a therapeutic track and if the provenance is not rigorously pursued. Potentially, this could lead to wasted time and money unless the error is identified quickly. It would be a potential '*horror hamatoxicus*', to misquote Ehrlich, if such a mouse-based sequence made it as far as the clinic. Establishing provenance however is not a trivial task. Many sequences are derived from patents where details of the human framework used are not always provided and may be further obscured by the presence of back mutations that are also not always described.

There appear to be two major sources of errors that should be easy to fix given sufficient cooperation between those who deposit sequences and the teams that annotate them. First, if a sequence is a chimeric antibody that contains non-human (typically mouse) variable domains and

human constant regions, it tends to be annotated as *Homo sapiens*. Similarly humanized antibodies (with non-human CDRs and human variable domain frameworks and constant domains) are also generally annotated as human. This is clearly incorrect: such sequences should be annotated as coming from the two species, or be indicated as artificial constructs. Second, there are a number of instances in the PDB where mouse antibodies expressed in a human cell line (typically HEK 293T cells) are annotated as human. It is not clear where the error comes from in these instances, but it is possible that the species information has not been provided by the authors and some automated algorithm recognizes the word 'human'.

It should be noted that the new formats for PDB files (the XML PDBML format and mmCIF files), recognize the fact that meta-data were not clearly or fully described in the 'legacy' PDB file format and allow more precise information to be presented. However, the problems described above are still present in the mmCIF files for all entries discussed (see Supplementary File `mmCIF.pdf`).

As mentioned above, some more complex difficulties in establishing provenance are exemplified by entries 4UOM and 4UOK in the PDB (see the heavy and light chain searches for HyHEL-5 in Supplementary File `BlastHits.pdf`). The 4UOM structure is described as the F5 Fab bound to the alphavirus Venezuelan equine encephalitis virus (VEEV) and is described in Porta *et al.* (2014). The earlier production of antibody F5 from a human bone marrow library is described in Hunt *et al.* (2010). This paper also describes mouse monoclonal antibodies, one of which, 3B4C-4, was humanized and was also the subject of the structural study by Porta *et al.* with PDB code 4UOK. The heavy chain (chain A) of PDB file 4UOK, (the humanized mouse antibody 3B4C-4 bound to VEEV) has a reasonable humanness score (Abhinandan and Martin, 2007) of –0.7 and a search against IMGT suggests it is most similar to human sequences. 4UOK was originally annotated as mouse, but this was changed to human in October 2014 as is currently generally the case for humanized sequences in the PDB.

Strangely however, the protein sequence of the heavy chain of 4UOM (described as human antibody F5) has a lower humanness score of –0.953 and a search against the IMGT Domain Display reference set suggests it is a mouse heavy chain. What is more, the CDRs of the heavy chains in 4UOK and 4UOM are identical (see Figure 2), while the frameworks have a different sequence as do the light chains. Thus, in attempting to pull together these various threads of information, this leads us to believe that, contrary to the database information, PDB entry 4UOM is actually the mouse sequence of 3B4C-4 and 4UOK is the humanized heavy chain with an undetermined light chain. However, we cannot be sure and have not been able unambiguously to establish their provenance. To be clear, we in no way question the validity of the published work, or of the databases themselves, but merely wish to point out that it is well known that databases contain errors.

Establishing the provenance of antibody sequences requires not just the normal scientist's tools, but also the skills of a detective and a historian. Even then, certain sequences may remain ill-defined. This situation needs to be resolved, particularly if people wish to generate automated algorithms for antibody humanization. As a solution to this problem, we would like to suggest the following modifications to the way in which antibody chain sequences are submitted to, and catalogued by, sequence databases:

- **First**, authors must be required to provide clear and unambiguous species information. For new submissions, provision of such information must be a requirement of all successful submissions.

- **Second**, where antibodies are chimeric or humanized, a standard way of indicating this information must be provided by the databases. Some PDB entries, such as 1GAF and 1AXS, have species information that states they are human, but do provide additional information to state that it is in fact a chimeric (see Figure 3a and b.) However, this is free

text information and the format is different between entries so cannot be parsed automatically. The same issues are present in the mmCIF versions of the files. Other entries contain organism information for both species (e.g. PDB files 1BBJ, 4KAQ, 4MA3) and were easily excluded from the searches performed here (see Figure 3c).

- **Third**, there needs to be a thorough review of current database entries to correct species information. Where possible, confirmation should be sort from authors for entries that have had automatic species annotation and, in other cases, database annotators need to check original literature. At a minimum, there needs to be a record of where the species annotation has come from (author, annotator, text mining, or elsewhere).

Conclusions

We hope this analysis will be helpful to researchers humanizing antibodies using standard antibody sequence databases. Of course, the concept of database errors is not new, but in the antibody field, the consequences of erroneous species annotation can be annoying, costly and potentially dangerous. We hope this analysis will allow database managers and annotators to take a serious look at methods of sequence description. Automated classification by software will only resolve the issues we have identified if the software is sufficiently 'intelligent', and the fact that many sequences are artificial constructs and do not have a single species origin must be acknowledged in a standard way in the annotations. While species annotation is a particular problem for antibodies, similar problems are almost certainly present in database entries for other proteins. The authors are happy to be contacted for further suggestions on how to improve this particular problem for antibody sequences.

References

Abhinandan, K. R. and Martin, A. C. R. (2007). *J Mol Biol*, **369**,852-862.

Allcorn, L. C. and Martin, A. C. R. (2002). *Bioinformatics*, **18**,175-181.

Bernett, M. J., Karki, S., Moore, G. L., Leung, I. W. L., Chen, H., Pong, E., Nguyen, D.-H. T., Jacinto, J., Zalevsky, J., Muchhal, U. S., Desjarlais, J. R. and Lazar, G. A. (2010). *J Mol Biol*, **396**,1474-1490.

Beyer, B. M., Ingram, R., Ramanathan, L., Reichert, P., Le, H. V., Madison, V. and Orth, P. (2008). *J Mol Biol*, **382**,942-955.

Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B., Henner, D., Wong, W. L., Rowland, A. M., Kotts, C., Carver, M. E. and Shepard, H. M. (1992). *Proc Natl Acad Sci U S A*, **89**,4285-4289.

Couto, J. R., Christian, R. B., Peterson, J. A. and Ceriani, R. L. (1995). *Cancer Res*, **55**,5973s-5977s.

Covaceuszach, S., Marinelli, S., Krastanova, I., Ugolini, G., Pavone, F., Lamba, D. and Cattaneo, A. (2012). *PLoS One*, **7**,e32212-e32212.

Darsley, M. J. and Rees, A. R. (1985). *EMBO J*, **4**,393-398.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J. and Deane, C. M. (2014). *Nucleic Acids Res*, **42**,D1140-D1146.

Foletti, D., Strop, P., Shaughnessy, L., Hasa-Moreno, A., Casas, M. G., Russell, M., Bee, C., Wu, S., Pham, A., Zeng, Z., Pons, J., Rajpal, A. and Shelton, D. (2013). *J Mol Biol*, **425**,1641-1654.

Haidar, J. N., Yuan, Q.-A., Zeng, L., Snavely, M., Luna, X., Zhang, H., Zhu, W., Ludwig, D. L. and Zhu, Z. (2012). *Proteins*, **80**,896-912.

Hanf, K. J. M., Arndt, J. W., Chen, L. L., Jarpe, M., Boriack-Sjodin, P. A., Li, Y., van Vlijmen, H. W. T., Pepinsky, R. B., Simon, K. J. and Lugovskoy, A. (2014). *Methods*, **65**,68-76.

Hunt, A. R., Frederickson, S., Maruyama, T., Roehrig, J. T. and Blair, C. D. (2010). *PLoS Negl Trop Dis*, **4**,e739-e739.

Jespers, L. S., Roberts, A., Mahler, S. M., Winter, G. and Hoogenboom, H. R. (1994). *Biotechnology (N Y)*, **12**,899-903.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. and Foeller, C., (1991). *Sequences of Proteins of Immunological Interest.* U.S. Department of Health and Human Services, Fifth edition.

Kashmiri, S. V., Shu, L., Padlan, E. A., Milenic, D. E., Schlom, J. and Hand, P. H. (1995). *Hybridoma*, **14**,461-473.

Lemeulle, C., Chardès, T., Montavon, C., Chaabihi, H., Mani, J. C., Pugnière, M., Cerutti, M., Devauchelle, G., Pau, B. and Biard-Piechaczyk, M. (1998). *FEBS Lett*, **423**,159-166.

Mader, A. and Kunert, R. (2010). *Protein Eng Des Sel*, **23**,947-954.

Olimpieri, P. P., Marcatili, P. and Tramontano, A. (2015). *Bioinformatics*, **31**,434-435.

Pelat, T., Bedouelle, H., Rees, A. R., Crennell, S. J., Lefranc, M.-P. and Thullier, P. (2008). *J Mol Biol*, **384**,1400-1407.

Porta, J., Jose, J., Roehrig, J. T., Blair, C. D., Kuhn, R. J. and Rossmann, M. G. (2014). *J Virol*, **88**,9616-9623.

Queen, C., Schneider, W. P., Selick, H. E., Payne, P. W., Landolfi, N. F., Duncan, J. F., Avdalovic, N. M., Levitt, M., Junghans, R. P. and Waldmann, T. A. (1989). *Proc Natl Acad Sci U S A*, **86**,10029-10033.

Riechmann, L., Clark, M., Waldmann, H. and Winter, G. (1988). *Nature*, **332**,323-327.

Roguska, M. A., Pedersen, J. T., Keddy, C. A., Henry, A. H., Searle, S. J., Lambert, J. M., Goldmacher, V. S., Blättler, W. A., Rees, A. R. and Guild, B. C. (1994). *Proc Natl Acad Sci U S A*, **91**,969-973.

Smith-Gill, S. J., Wilson, A. C., Potter, M., Prager, E. M., Feldmann, R. J. and Mainhart, C. R. (1982). *J Immunol*, **128**,314-322.

Tan, P., Mitchell, D. A., Buss, T. N., Holmes, M. A., Anasetti, C. and Foote, J. (2002). *J Immunol*, **169**,1119-1125.

Tsurushita, N., Hinton, P. R. and Kumar, S. (2005). *Methods*, **36**,69-83.

Ulrich, H. D., Mundorff, E., Santarsiero, B. D., Driggers, E. M., Stevens, R. C. and Schultz, P. G. (1997). *Nature*, **389**,271-275.

Verhoeyen, M., Milstein, C. and Winter, G. (1988). *Science*, **239**,1534-1536.

Wu, H., Nie, Y., Huse, W. D. and Watkins, J. D. (1999). *J Mol Biol*, **294**,151-162.

Figure Legends

Figure 1 Relevant extracts from databank headers for example entries labelled as *Homo sapiens* but which are actually mouse antibody chains. a) 3DGG (PDB) Actual: Mouse monoclonal antibody expressed in human cell line; b) 3D85 (PDB) Actual: Mouse monoclonal V-regions as part of a chimeric antibody; c) 1AXS (PDB) Actual: Mouse monoclonal antibody; d) CAT05563.1 (EMBL-ENA) Actual: Chimeric antibody with mouse V-regions and human constant regions; e) 041427/1C10 (Kabat) Actual: Mouse monoclonal against digoxin.

Figure 2 Sequence alignment of the heavy chains of the antibodies in PDB files 4UOK and 4UOM numbered according to the Chothia numbering scheme and indicating the CDRs.

Figure 3 Relevant species ('SOURCE') information from PDB files a) 1GAF and b) 1AXS containing chimeric antibodies. While the key annotation lines state that the sequence is human, additional information shows that it is chimeric. c) 1BBJ has more informative key species annotation lines that show clearly that the chain is chimeric.

Figure 1

a)

```
HEADER    IMMUNE SYSTEM                           13-JUN-08   3DGG
TITLE     CRYSTAL STRUCTURE OF FABOX108
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: FABOX108 LIGHT CHAIN FRAGMENT;
COMPND   6 MOL_ID: 2;
COMPND   7 MOLECULE: FABOX108 HEAVY CHAIN FRAGMENT;
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE  10 MOL_ID: 2;
SOURCE  11 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
```

b)

```
HEADER    IMMUNE SYSTEM/CYTOKINE                  22-MAY-08   3D85
TITLE     CRYSTAL STRUCTURE OF IL-23 IN COMPLEX WITH NEUTRALIZING FAB
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: FAB OF ANTIBODY 7G10, LIGHT CHAIN;
COMPND   5 MOL_ID: 2;
COMPND   6 MOLECULE: FAB OF ANTIBODY 7G10, HEAVY CHAIN;
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   5 MOL_ID: 2;
SOURCE   6 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
```

c)

```
HEADER    CATALYTIC ANTIBODY                      20-OCT-97   1AXS
TITLE     MATURE OXY-COPE CATALYTIC ANTIBODY WITH HAPTEN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: OXY-COPE CATALYTIC ANTIBODY;
COMPND   7 MOL_ID: 2;
COMPND   8 MOLECULE: OXY-COPE CATALYTIC ANTIBODY;
SOURCE    MOL_ID: 1;
SOURCE   3 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE  21 MOL_ID: 2;
SOURCE  23 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
```

d)

```
LOCUS       FB985353               1413 bp    DNA     linear   PAT 14-DEC-2008
DEFINITION  Sequence 3 from Patent WO2006126069.
ACCESSION   FB985353
VERSION     FB985353.1  GI:218044635
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

e)

```
KADBID      041427
CREATD      11/16/98
DEFINI      IG KAPPA LIGHT CHAIN VARIABLE REGION
SPECIE      human
AANAME      1C10'CL
NNNAME      1C10
```

Figure 2

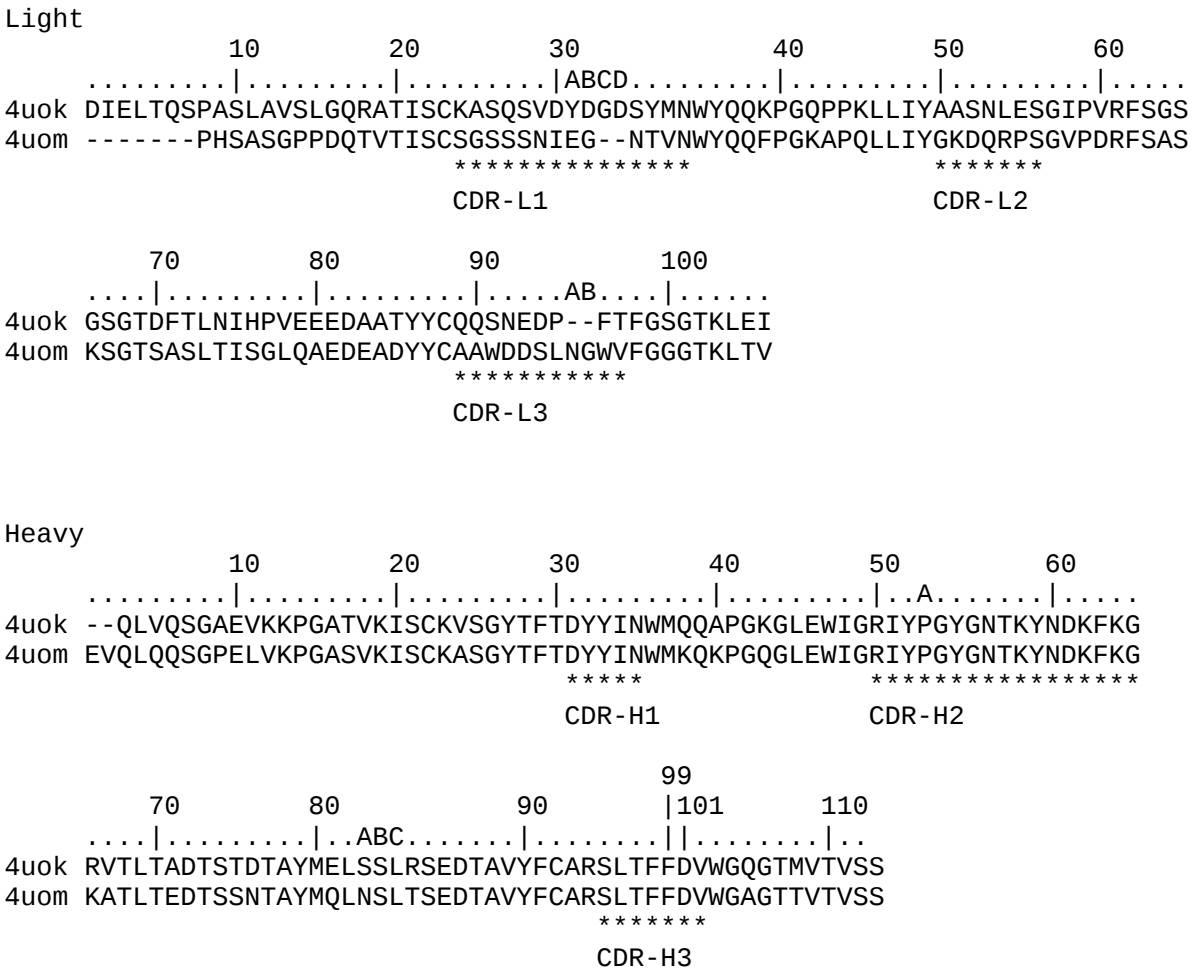
Light
10 20 30 40 50 60
.....|.....|.....|ABCD.....|.....|.....|.....
4uok DIELTQSPASLAVSLGQRATISCKASQSVDYDGDSYMNWYQQKPGQPPKLLIYAASNLESGIPVRFSGS
4uom -------PHSASGPPDQTVTISCSGSSSNIEG--NTVNWYQQFPGKAPQLLIYGKDQRPSGVPDRFSAS
********** *******
CDR-L1 CDR-L2
70 80 90 100
....|.....|.....|.....AB....|.....
4uok GSGTDFTLNIHPVEEEDAATYYCQQSNEDP--FTFGSGTKLEI
4uom KSGTSASLTISGLQAEDEADYYCAAWDDSLNGWVFGGGTKLTV
****
CDR-L3
Heavy
10 20 30 40 50 60
.....|.....|.....|.....|.....|..A.....|.....
4uok --QLVQSGAEVKKPGATVKISCKVSGYTFTDYYINWMQQAPGKGLEWIGRIYPGYGNTKYNDKFKG
4uom EVQLQQSGPELVKPGASVKISCKASGYTFTDYYINWMKQKPGQGLEWIGRIYPGYGNTKYNDKFKG
***** **********
CDR-H1 CDR-H2
99
70 80 90 |101 110
....|.....|..ABC.....|.....||.....|..
4uok RVTLTADTSTDTAYMELSSLRSEDTAVYFCARSLTFFDVWGQGTMVTVSS
4uom KATLTEDTSSNTAYMQLNSLTSEDTAVYFCARSLTFFDVWGAGTTVTVSS
****
CDR-H3

Figure 3

a) 1GAF

```
SOURCE    MOL_ID: 1;
SOURCE   2 FRAGMENT: CONSTANT DOMAINS OF LIGHT AND HEAVY CHAINS;
SOURCE   3 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   4 ORGANISM_COMMON: HUMAN;
SOURCE   5 ORGANISM_TAXID: 9606;
SOURCE   9 OTHER_DETAILS: EACH CHAIN IS A FUSION POLYPEPTIDE WHICH IS
SOURCE  10 PART HUMAN AND PART MOUSE;
```

b) 1AXS

```
SOURCE    MOL_ID: 1;
SOURCE   2 FRAGMENT: CHAIN L, A, 108 - 211, CHAIN H, B, 114 - 214;
SOURCE   3 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   4 ORGANISM_COMMON: HUMAN;
SOURCE   5 ORGANISM_TAXID: 9606;
SOURCE  17 OTHER_DETAILS: THE PROTEIN WAS PRODUCED AS CHIMERIC FAB
SOURCE  18 FRAGMENT. THE CONSTANT DOMAINS ARE HUMAN, THE VARIABLE
SOURCE  19 DOMAINS ARE MURINE.
```

c) 1BBJ

```
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: MUS MUSCULUS, HOMO SAPIENS;
SOURCE   3 ORGANISM_TAXID: 10090, 9606;
```