Errors in patent application sequence listings

Robert Jones

Seemingly minor errors in genetic sequence listings can cause costly delays when applying for a patent.

Filing a patent application is the key step in protecting a biotechnology invention. It is a race where being second doesn't count. The owners of key patents gain tremendous value and control over the market for the covered technology. Accordingly, patent attorneys are under great pressure to file on new research as quickly as possible, while ensuring both accuracy and completeness. With so much at stake, an application is the last place you want to make a clerical error. But this is exactly what is happening. In 2002, approximately one in every eight patent applications submitted to the US Patent and Trademark Office (USPTO; Washington, DC, USA) was rejected due to formatting errors in the attached sequence listing. This is a remarkably high failure rate.

These are extremely expensive mistakes to make. Correction incurs additional attorney and filing fees and months of delay, not to mention embarrassment and lost business for the law offices involved. Where do these errors stem from and what can be done to eliminate them?

What is a sequence listing?

The requirement to include sequences in a separate sequence listing, as opposed to inclusion in the text of a patent application, or as a drawing, was introduced in 1990. This ensured that the growing number of sequences was reported in a consistent, machine-readable form. This was replaced in 1998 by a new format which is more amenable to automated processing and which can better represent the diverse types of sequence data that have emerged in recent years. This is now the standard for all patent offices around the world[1].

In its simplest form, each block of sequence is preceded by mandatory information, such as the type of sequence (DNA, RNA or protein), its length and the organism it is derived from, along with optional information, such as sequence features and related publications. Each piece of information is identified by a specific numeric tag contained within angle brackets: for example, the length of each sequence is preceded by <211> and the type is preceded by <212>.

Of particular importance is the 'SEQID' number, preceded by a '<210>' tag, that uniquely identifies each block of sequence. Sequences within a single listing begin with SEQID 1 and are numbered consecutively. These numbers are used in the text of a filing to relate claims and descriptions to specific sequences. It should be stressed that SEQID numbers are the *only* way to relate a specific claim to a specific sequence. Entering the wrong SEQID number in a claim can completely invalidate that claim. To minimize the risk of errors, one should always prepare the sequence listing first and then write specific claims that refer to them.

Applications with hundreds or thousands of sequences are now commonplace. There is no practical way to generate sequence listings by hand with this much data. These sequences are initially assembled using analysis software, and the resulting files are passed to specialized software that converts them into sequence listings. The USPTO makes its software, called PatentIn, available free of charge, and several companies market commercial software for this task. These tools should have all but eliminated errors. The software works well and the vast majority of applicants use it. So why is there still such a problem? To answer that, we need to look for the source of these errors.

Types of errors encountered

Staff members at the USPTO were very helpful in providing information on the types of errors they encounter. In addition, the World Intellectual Property Organization (WIPO; Geneva, Switzerland) makes available a number of the sequence listings submitted to them on their FTP site. The most frequent errors identified at the USPTO include files not being submitted as plain text, incorrect line lengths and line wrapping, nonconsecutive SEQID numbers, missing numeric tags and invalid sequence features, especially those that describe ambiguities in the reported sequence. The latter is a particular problem for applications that include incomplete sequence data, such as single reads from cDNA libraries.

In order to quantify these errors, I surveyed sequence listings from the WIPO FTP site. These represent a subset of the listings submitted during the period August 2001 through October 2002. Out of the 192 listings examined, 53 (28%) had at least one problem as reported by our own validation software. The largest single type of error involved the representation of dates. In 19 of the listings, these were either missing or were given in the wrong format. Ten of the files were Microsoft Word documents rather than the required plain text format. There were a total of 40 other errors, ranging from extraneous text at the end of the listings to missing numeric tags and inconsistencies in the reported number of sequences. Multiple problems were found in 13 of the listings.

Any one of these errors, regardless of type, would be enough for the USPTO to reject the sequence listing. Why these have apparently been accepted by the WIPO is unclear. That office may be more flexible in dealing with minor errors, but some of these are severe and should render the listing completely invalid. The USPTO exhibits no such flexibility. A format error of any kind results in the sequence listing being rejected. According to the USPTO, in the first ten months of 2002, they rejected 2,240 listings out of a total of 17,627 received (12.7%). Prior experience with these rigid acceptance criteria may result in better screening by applicants prior

Robert Jones is at Craic Computing LLC, 911 East Pike Street, Suite 208, Seattle, Washington 98122, USA. e-mail: jones@craic.com

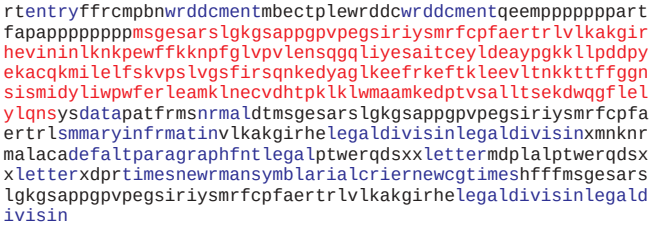

Figure 1 In a protein sequence, the use of color serves to identify the 'real' sequence (red), extraneous text (black) and English words that identify the text as extraneous (blue).

to submission to the USPTO and that, in turn, may explain the lower frequency of errors relative to the WIPO.

Errors in the sequence data

Whereas the vast majority of the errors relate to the format of listings, there is a small but dramatic subset where the errors are reflected in the sequences themselves.

In at least eight issued patents there are English language words embedded within certain protein sequences. These are not the result of a chance match between one or two words and a legitimate sequence. Rather, they appear to be real words from which characters that do not represent valid amino acids, such as the letters 'J', 'O' and 'U' have been removed, presumably by some piece of sequence analysis software. A striking example is shown in **Figure 1**, in which the real sequence is shown in red, surrounded by two blocks of extraneous text shown in black. Within this, strings of characters that appear to represent edited words are shown in blue. In this case we know what the real sequence is as the patent includes it separately, without error, in a figure in the patent. **Table 1** shows some of these words and their translation. Most of them relate to word processing and printing, providing a strong clue as to their origin.

It is important to note that errors of this kind will not, by themselves, result in a listing being rejected by the patent office. They are likely, however, to invalidate any claims that are made about the sequences.

The problem seems to lie in what happens to the sequence listings after they have been created. It is often necessary to modify the contents of a listing prior to submission of the filing to reflect changes in the application date, the inventors or the title. The person making the changes has to understand the specific format required for each data item in the listing. This seems especially critical with dates, the most common single error in the WIPO listings.

For a regular document, one would not hesitate to open the file in Microsoft Word, edit the text and then save the file. Therein lies the problem. Sequence listings are plain text files, not Word documents. Current versions of Word and similar programs contain a range of features such as multiple fonts, line wrapping and automatic correction of spelling mistakes. When these are applied to the plain text of a listing, one can unwittingly make major changes to the content.

One of the problems we have seen in the WIPO and other listings is the presence of extraneous text, often at the bottom of a listing. This can range from a few random characters to short sets of words. These may involve 'header' and 'footer' information such as page numbers and modification dates. Precisely how this text finds its way into a listing is impossible to determine. Possible causes include software bugs, the transfer of documents between different versions of word processing applications, the cutting and pasting of text between applications and the transfer of documents via e-mail.

Table 1 Words found in the sequence in Figure 1, and their original form

Protein sequence	Original text
micrsft	Microsoft
wrddcment	Word document
nrmal	normal
smmaryinfrmatin	summary information
legaldivisin	legal division
defaltparagraphfnt	default paragraph font
timesnewrman	Times New Roman
symbl	symbol
crier	Courier
hplaseretseriesii	HP LaserJet Series II
prtein	protein
recvery	recovery
captin	caption
ftnte	footnote

The implications of submitting a bad sequence listing

USPTO staff members are extremely clear in their response to errors: they check all listings with their version of the Checker validation software. If they find a problem—any problem—then it is your job to fix it. They will send you a Notice of Incomplete Application along with a summary of the errors and advice on how to correct them. You need to fix the problem and resubmit the listing. If everything checks out the second time around, then the filing can move forward. If not, you will be sent another notice and one more chance to fix the problem. Fail to do so and your application will be abandoned—three strikes and you're out.

In most cases, applicants will resubmit a valid listing after their first notice. But bear in mind that this diversion has delayed your application by at least a few months, and has cost you time and money in attorney hours and filing fees. Moreover, patent attorneys who make this mistake face the very real risk of losing business from an upset client.

What you can do

All of these problems can be avoided by following some simple guidelines.

- Always save sequence listings files as plain text, and never as Microsoft Word documents.
- Let the scientists that are providing the data do all the work in preparing the input file of sequences.
- Use software to generate the sequence listing. Do not even think of doing this by hand. If you need to modify the header information in a listing, then make absolutely sure that your changes are compatible with the required format.
- If you need to make any changes to the sequence information, then go back to the original input file. Make your changes there and generate a new sequence listing from that file. Never cut and paste blocks of text between listings.
- Finally, use the validation software available from the patent office to double-check your listing before you file your application.

Conclusions

The reality of preparing a patent application involves research and legal staff in multiple rounds of writing, editing and review. The rigid requirements for sequence data sit uneasily in this process. All sequence data should be prepared prior to working on the text of the application and should be considered as final. If changes are needed, they should be handled very carefully, and all references to the SEQID numbers made within the text should be checked for consistency.

Rejection of a patent application reflects badly on everyone involved. By using the guidelines discussed here you can minimize the chance of this happening to you.

ACKNOWLEDGMENTS

The author gratefully acknowledges the assistance of staff at the US Patent and Trademark Office in preparing this article.

1. See 37 CFR §§ 1.821–825 and, separately, in WIPO Standard ST.25.