

Information resources for biotechnologists

Part 1: Sequences

Different sequence databases contain sequences from different sources, and may list very different sequences, even from the same source. So what are the implications, and how should we search for sequences? Robert Austin and Damon Ridley write that the whole area is a minefield for the unwary...

In recent years chemists have become familiar with electronic information sources, in which search options range from bibliographic and key word options to graphical structure and reaction options. As many chemists in academia and industry are becoming involved in biotechnology areas, they are wondering about search opportunities in their new field.

Of course, the biotechnology industry also has a range of information sources, and of search options. Again they include bibliographic and key word options, but in many cases scientists now need to search sequences rather than structures. In this article we discuss some issues in the searching of sequences; in a subsequent article we shall discuss searches in non-sequence databases.

GenBank

The flagship sequence database for biotechnologists is GenBank, which was set up in 1982. Currently the National Center for Biotechnology Information (NCBI) is responsible for GenBank. Details of the database, and of various search

opportunities, may be found through the NCBI site (www.ncbi.nlm.nih.gov). In summary, it contains mainly nucleotide sequences either directly input by scientists to the NCBI, or sequences entered through the organisation's European Molecular Biology Laboratory (EMBL) (www.embl-heidelberg.de/) and the Japanese DDJB (<http://ftp2.ddbj.nig.ac.jp:8000/getstart-e.html>).¹ These (directly input) sequences make up the majority of the database, but it also contains some sequences reported in journals and patents.

The primary advantages in searching GenBank through the NCBI are that it is comprehensive and searches may, in most instances, be conducted free of charge – very attractive indeed! But is it the world's *most comprehensive* sequence database? Does it cover all areas required by those interested in sequence searching? And are there other issues that the sequence searcher should be aware of?

The answer to the first two questions is a definite 'no' (which begs the question as to what

sequences it does *not* contain), and the answer to the last question is a definite 'yes'. Related to this final aspect is the way in which sequences are entered (e.g. different authors may submit identical sequences) or may be subsequently modified (e.g. because new research may have found errors in the original assignments, or the original assignments may have been incomplete).

In order to understand GenBank more fully, it is helpful to go back to the original sources – the sequences discovered by scientists. Scientists publish sequences primarily in journals and patents, but they also have the option to directly enter sequences in GenBank (in fact around 85% of the sequences in GenBank do not have journal or patent references).

The questions then are:

- How comprehensively do scientists enter sequences directly?
- What is the situation for sequences known to industry and usually withheld because of intellectual property considerations?

- When sequences appear in *journals*, how do they become part of GenBank (or indeed of other sequence databases)?
- What is the situation with sequences in *patents*?

The answer to the first question is unknown. However, just as with research in chemistry there is a large amount of knowledge that has not been (and which never may be) published. Second, we know that many companies have their own resources, but there is little that we can do about this other than to hope that eventually the scientific community can share the information. Third, GenBank does cover a range of journals but we have been unable to determine the complete list of titles. However, from the nucleotide search we present below, it is apparent that coverage is not complete both with respect to titles and sequences. Finally, we also address some issues relating to sequences in patents below.

Other sequence databases

Acronyms abound! There is SWISS-PROT, PIR, EST, GSS, HTG, SNP... the list goes on. The result is different databases – with different entries. Many organisations are involved in the area, but again questions of comprehensiveness and accuracy of the data arise. Biotechnologists would have been sold, or would be aware of, certain options, but a critical point is how all of these sources compare. We also have to keep in mind the differences in nucleic acid and protein sequence databases, and options for the translations (of the nucleotide sequences into amino acid sequences, and vice versa).

We do not attempt here to make cross-the-board comparisons, but instead we discuss issues for two of the world's major databases that cover sequences in patents.²

Table 1. Comparison of some numbers of records in GenBank, GENESEQ and Registry, May 2002

	GenBank*	GENESEQ*	CAS Registry*
Nucleic acid (total)	16.5	1.7	19.0
Nucleic acid (from patents)	0.6	1.7	1.8
Protein (total)	0.6 [†]	0.7	1.4
Protein (from patents)	0.02 [†]	0.7	0.75

* Numbers are millions.

[†] Numbers refer to translated proteins in GenPept.

Table 2. A comparison of nucleotide sequences in the CAS Registry and GENESEQ from the original Australian Patent AU724493

SQL	CAS RN	GENESEQ AAQ (R)	Differences
Nucleic acid			
1621	336886-13-16	66423	
1620	336903-96-9		Does not have extra A at 1323
1132	158007-25-1	66424	
1132	336903-97-0		Has two more C, two less G
838	336196-32-8	66421	
651	336903-95-8		
650		66425	Does not have T at 600
496	158007-29-5	66422	
414	158007-27-3		
411		66420	Does not have final TAA
357	336196-33-9	66419	
21	336208-27-6	66426	
15	336208-28-7	66427	
13	336208-29-8	66428	
9	336208-30-1	66429	
Amino acid			
137	158007-26-2	53810	
119	150475-55-1	53809	

One of these sources is the GENESEQ database from Derwent (www.isinet.com/emea/geneseq/). It contains over 1.7 million nucleic acid sequences from patents, and the GENESEQ website states: '[the database] has been shown to consistently contain a massive 50% unique sequence data compared to public access databases'. Possibly,

some evidence for this statement comes from the document 'A comparison of Derwent's GENESEQ with SWISS-PROT and EMBL' (www.isinet.com/emea/geneseq/images/comparison.pdf), but this comparison was based on 1998 data and in any case there are many 'public access databases' other than SWISS-PROT and EMBL!

In fact the world's largest single source of sequence information is in the Chemical Abstracts Service (CAS) Registry database, and here the latest number of sequences it contains may be obtained through www.cas.org/cgi-bin/regreport.pl. In addition to over 18 million chemical substances, the CAS Registry database contains *the complete GenBank database*, plus over two million nucleic acid sequences from patents and additional sequences from journal articles.

Table 1 summarises the numbers of sequences in GenBank, GENESEQ and the CAS Registry. Clearly the CAS Registry has 2.5 million *additional* nucleic acid sequences to GenBank, but how do the 1.7 million nucleic acid sequences in GENESEQ fit in?³ We may partly answer this by comparing actual sequences, and we present here two randomly chosen examples.

Comparison of a patent (protein sequences)

Patent WO9700886 is titled 'Obesity protein intermediates and their preparation and use'. The document analysts from Derwent posted 24 sequences from this patent in GENESEQ, while CAS analysts posted 10 sequences.⁴ However, as usual, numbers alone do not tell the full story and further analyses are nearly always needed. In fact the GENESEQ record contains two nucleic acid sequences and six oligopeptide sequences, and the key 'obesity protein intermediates' are thus listed in 16 sequences. All have 146 amino acids, but in fact there are only eight different sequences.⁵

So, the real comparison to be made is between these eight different sequences and the 10 sequences in the CAS Registry. Since the analysts obtained these sequences from the same original patent, you would think the eight GENESEQ sequences would be identical to at least

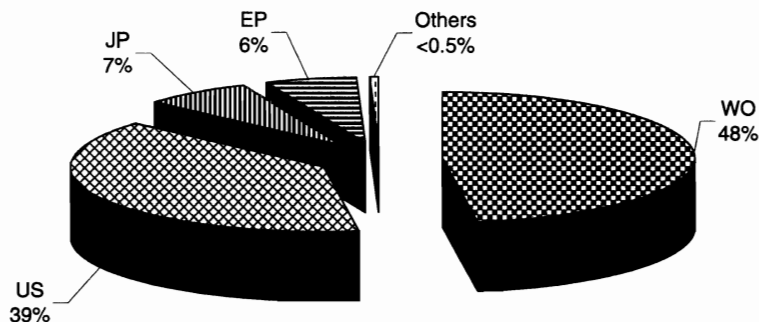


Figure 1. Nucleic acid sequences in GenBank by patent source (April 2002). The number of sequences from Australian patents was less than 0.05% of the total. There were about 150 000 peptide sequences from patents in the GenBank Peptide database (GenPept), taken almost exclusively from US patents. GenPept also included a further 16 000 additional protein sequences translated from the coding regions of patent nucleic acid sequences in GenBank.

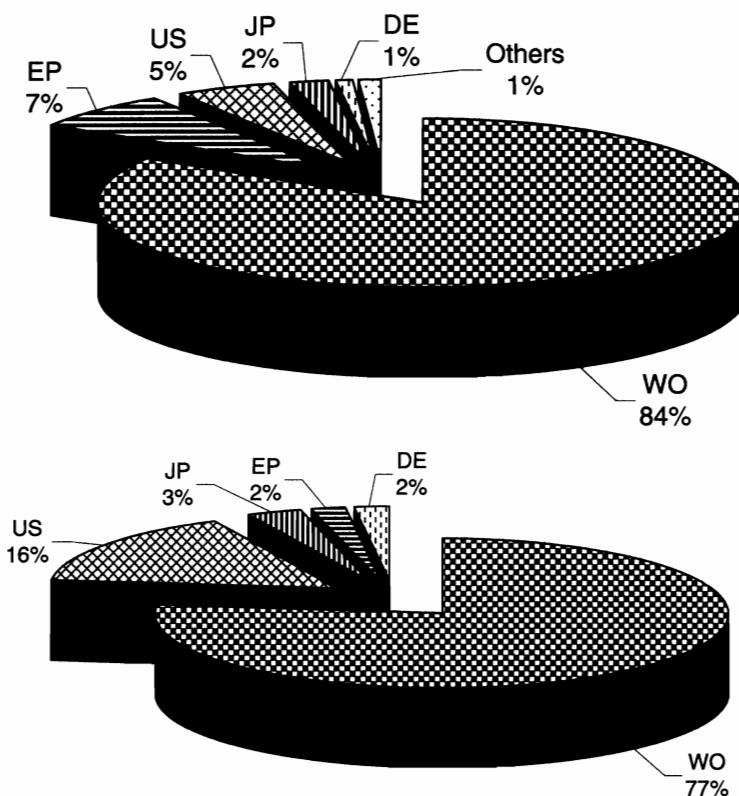


Figure 2. Nucleic acid sequences in GENESEQ (upper) and CAS Registry (lower) by patent source. (Note: Data for CAS Registry was compared only for the five major patent sources listed.)

eight of the 10 sequences in the CAS Registry. Wrong – all eight sequences in GENESEQ (which as already noted contain 146 amino acids) are *different* from all the 10 sequences in the CAS Registry (which all contain 148 amino acids)! Actually, the differences are in fact 'small' since the CAS Registry sequences all contain two additional amino acids at the start of the sequence; that is, one of the CAS Registry sequences is:

```
1 MRVPIQKVQD DTKTLIKTIV
TRINDISHTQ SVSSKQKVTG
LDFIPGLHPI
51 LTLSKMDQTL AVYQQILTSM
PSRNVIQISN DLENLRDLLH
VLAFSKSCHL
101 PAASGLETLD
SLGGVLEASG YSTEVVLSR
LQGSQDMLW QLDLSPGC
```

and two of the GENESEQ sequences contain the same amino acids *except for the MR at the beginning.* (continued p10)

Table 3. Answers from subsequence search on GGCTGTGTTTGGCTCTCTCC, May 2002

	GenBank	GENESEQ	CAS Registry
Sequences	110	37	145
Sequences from patents	1	37	29
Sequences from journals	26	0	35
Sequences with no references	83	0	83
Patents retrieved	14	23	26
Journal articles retrieved	13	0	14

* Results when answers (i.e. the sequences) from the CAS Registry were searched in the CAS bibliographic database, CAPLUS.

Comparison of a patent (nucleotide sequences)

The Australian patent AU724493, titled 'Developmental regulation in anther tissue plants', is indexed with 15 sequences in the CAS Registry and 13 in GENESEQ. Fifteen versus 13 – but how do these compare?

Table 2 gives the answer, and the good news is that about half of the sequences are identical. However there are subtle differences. For example, the CAS Registry has a sequence of 1621 nucleotides and this appears to correlate with a sequence of 1621 nucleotides in GENESEQ, but the CAS Registry also has a sequence of 1620 nucleotides in which the adenoside at 1323 is missing.

So how do GenBank, GENESEQ and CAS Registry compare (patents)?

Given the (surprising?) differences in individual cases perhaps we should return to the big picture. We found there were almost 600 000 sequences from patents in GenBank and the breakdown of patent sequences from the different patent organisations is shown in Figure 1.⁶ The corresponding breakdown for sequences from patents in GENESEQ and the CAS Registry is shown in Figure 2.

What this shows is that patent coverage overall is different, and we have already seen that

coverage for individual patents is also different. This is the material that sequence searchers have at their disposal, and clearly their search strategies need to make allowance for the differences!

Searching for sequences

There are two basic ways to search for sequences. The first method involves matching of sequence codes, and typical options include exact or subsequence matches. In these categories there are options to perform 'family' searches which allow certain substitutions of units. For example, only aspartic acid units in answers will be retrieved in an *exact* sequence match when the queries contain aspartic acids, but in an *exact family* search, aspartic acid and glutamic acid units will be retrieved when the queries contain aspartic acids, since aspartic and glutamic acids are defined in the same 'family' (here amino acids with acidic side chains).

Subsequence searches may incorporate many variables, including undefined sequences between specific sequence components, and may be devised to allow for other factors (like concatenation).⁷ However, in view of the differences in the databases (e.g. the differences apparent in Table 2), it is clear that there are many challenges

for those who search databases by sequence code matches.

The second method for searching sequences is called homology (or similarity) searching, and the most common methods involve BLAST,⁸ FASTA⁹ or Smith-Waterman algorithms. Essentially the searcher enters a sequence and the different algorithms retrieve answers arranged by a ranking order. A 100% similarity more or less means an exact match, while a 90% similarity effectively means that the retrieved sequence matches the query sequence closely, although there are differences (which may be seen by aligning the query with the answer). An issue here is that the DNA of each individual is different, so often biotechnologists want similarities rather than exact matches. Besides, the key interest may not be in the sequence but in the function – and now we go into other stages of the science called bioinformatics, which starts from sequences but then moves into function, prediction of function, and ultimately into prediction, prevention or cure of disease.

Example of a nucleotide subsequence search (patents)

As an example of the outcome of a sequence search, consider a subsequence search for the nucleotide sequence GGCTGTGTTTGGCTCTCTCC.¹⁰ A summary of the results is shown in Table 3, and scientists may quickly assess the implications. For example, if you only search GenBank you miss out on an additional 35 sequences in the CAS Registry; if you only search GENESEQ you are not retrieving sequences in journals, or sequences that are 'published' only in GenBank/CAS Registry through direct entry of sequences by scientists.

However, comparison of GENESEQ and Registry is not so straightforward. Table 4 gives a

Table 4. Comparison of sequences obtained in subsequence search for nucleotides in GENESEQ (37 sequences) and in the CAS Registry (29 sequences)*

GENESEQ	CAS Registry	GENESEQ	CAS Registry
20	20		1476
22	22	1477, 1477	1477
200		1557	1557
461	461	3383, 3383, 3383, 3383	3383, 3383
792			
	1053	5674	5674
1056, 1056	1056, 1056		9141, 9141
1059, 1059	1059, 1059		
1071	1071		143005
1225, 1225, 1225	125, 1225, 1225		143008
1344	1344	143068, 143068, 143068	143068, 143068
1376, 1376	1376		152680
1414, 1414, 1414, 1414	1414	152740	
1442			
	1444		

* Data here relates only to lengths of sequences retrieved (e.g. GENESEQ retrieved five sequences with 1414 nucleotides; CAS Registry retrieved one sequence of this length; CAS Registry gave a sequence of length 152 680 nucleotides while the most similar sequence from GENESEQ had 152 740 nucleotides).

comparison between the 37 sequences in GENESEQ and the 29 (from patents) in the CAS Registry. Again there are similarities, but there are also significant differences.

And how do the 23 patents from GENESEQ compare with the 26 patents from which the CAS Registry answers derive? In fact a direct comparison of the patents indicates that 19 are common to both systems. So does this mean that there are four unique patents in GENESEQ, and seven unique patents through the CAS Registry? No – for example the four patents in question in GENESEQ do have records in the CAS Registry databases; the reason the patents were retrieved only in GENESEQ (and not in the CAS databases) is because of the sequence search terms entered.¹¹ Again this reflects different indexing of sequences, and the search strategies employed by the searcher.

So what does all of this mean?

There are many conclusions from all of this. *Inter alia*, first, GenBank does not contain all reported sequences so those who rely solely on this database, and indeed on other databases that are available free of charge on the web, may be missing out on answers. Second, a critical point for the sequence searcher to consider is whether there is a need to search for sequences in patents. If the answer is yes (and essentially it virtually always is!), then other databases need to be searched. Third, we have shown that there are surprising differences in the ways in which GENESEQ and the CAS Registry enter sequences from patents. Fourth, the implications of these differences need to be considered and, as a general rule, it probably is better to search by homology; if sequence match searching is performed then variations (often quite unpredictable) may need to be entered in the query. Fifth, it is advisable to search a variety of databases.¹²

So if you thought a ‘sequence is a sequence’, and if you thought that a search on some free sequence databases on the web covered everything and was all you need, then think again! This whole area is a minefield for the unwary. The solutions come first through being aware of the issues, and second through ‘application of scientific method’ to your searches. You should perform searches as you perform experiments in the laboratory. That is, you should know your science (in this case databases) before you start. It helps if you are aware of the potential outcomes of your searches, and if you critically evaluate your results and revise searches to improve the ‘experiment’.

If you don’t get it right, then the painstaking research in which your company is investing may all come to nothing, or the integrity of your academic research may be in jeopardy. Be warned!

However, the final point to make is that if you invest time in studying your options, you will not only learn much about your areas of interest but you also will become aware of new research opportunities.

Notes

- 1 Details of this 'International Nucleotide Sequence Database Collaboration' may be found at <http://www.ncbi.nlm.nih.gov/collab/>.
- 2 Sequences in patents are very important in industry and academia. Usually they are not published elsewhere, so searchers really need to know patent coverage for the databases they use. This may be done by contacting the database producer. For example, we were advised by SWISS-PROT that they do not systematically include sequences from patent applications, partly because these sequences often concern artificially generated sequences outside the scope of the database. However, such policies may change with time and we emphasise the need to obtain up-to-date information directly from the producer.
- 3 The numbers quoted in this article were obtained at various times between March and May 2000. Every effort was made to correlate the data, but small differences in numbers may reflect time differences when data was collected. However, numbers change rapidly, e.g., about 3500 sequences are entered into GenBank daily, and as we shall see a key factor is how many sequences are recorded in the different databases for a single article.
- 4 We were unable to retrieve a reference to WO9700886 in GenBank.
- 5 Each of the eight different sequences appears in two records; in turn each pair of records differs by descriptions in the 'Features Tables', which are another issue for sequence searchers to know about!
- 6 The GenPept database had approximately 150 000 protein sequences which appeared to come only from US patents.
- 7 Details of sequence search options are beyond the scope of this article. However the point is that there are many alternatives with which those who search biosequences should be familiar.
- 8 See for example Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. *Nucleic Acids Res* 1997, **25**, 3389.
- 9 See for example Pearson W.R., Lipman D.J. *Proc. Natl Acad. Sci.* 1988, **85**, 2444.
- 10 The results from this search, conducted in May 2002, are reported here.
- 11 In a similar way, records for the seven unique patents in the CAS Registry databases are in GENESEQ.
- 12 One of us (DDR) is involved in training courses in Australia, and may be contacted for further information.

Robert Austin, FIZ Karlsruhe, Germany and **Damon D. Ridley FRACI**, School of Chemistry, University of NSW

Email robert.austin12@verizon.net, d.ridley@chem.usyd.edu.au