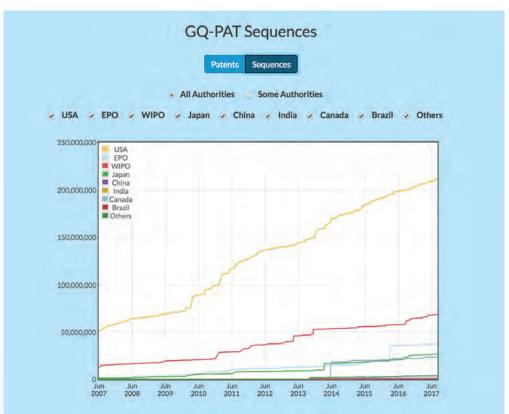# GQ Life Sciences

TEN TRAPS FOR ATTORNEYS TO AVOID IN IP SEQUENCE SEARCH AND ANALYSIS

The growth of sequence IP is nothing short of amazing! In 2007, we had about 50 million sequences – ten years later, we are fast approaching 400 million! Just statistically, the sheer volume of data may contain common sequences or random matches that are similar to many sequence IP filings—but how do you find those genetic needles in a haystack and understand their significance without getting stuck? It takes skill, experience, and the right tool – GenomeQuest!

A free IP sequence search usually involves the following steps: search the GenBank patent divisions on the NCBI BLAST website, go through the alignments one by one, and look up related patent information on the web. Findings are tracked on a printout of the BLAST results or in a spreadsheet, which is very inefficient, and as explained below, certainly not comprehensive.

Trap 1: Not fully understanding the different percent identities – not only the term "percent identity" but the difference between query, subject and alignment % identity.

Subject % identity is very useful in detecting hits that are of potential concern for a claim with "comprising" language;

alignment % identity (often in combination with query and/or subject coordinates) is required to detect subsequence claims, and query % identity corrects alignment % identity for how much of your query sequence is covered.

But, how do you screen by a combination of % identities? Or add in coordinates? GenomeQuest makes this possible!

Which leads to:

Trap 2:   Being unaware of which algorithm and parameters are used to define % identity

There can be enough of a difference between the different algorithms, or even an algorithm like BLAST if the parameters are changed, to affect whether your query sequence falls in or out of the scope of any given % identity element of a claim.

This doesn't mean you have to use all algorithms to search. If the content of a patent document is such that your sequence would be claimed if it fell within the recited % identity, then read the definition of % identity very, very carefully. If you are anywhere close (either in or out of scope) align your sequence vs the claimed sequence using the same algorithm and parameters used to define % identity in the patent document.

This could also be beneficial; you may find that your sequence is outside the scope of a claim once you do this!

Trap 3: Using the wrong algorithm for your search

Blast is not the ONLY sequence search algorithm that should be used in searches, in spite of its wide use and acceptance. Our proprietary algorithm, GenePAST, is especially useful for very short sequence searches, where Blast might return too many, unrelated hits or miss these matches entirely due to their high e-values.

Conversely, Blast is very useful for finding subsequences of interest in longer sequences or even multiple subsequence matches—so sometimes a combination of both is the best solution.

Consider degenerate or variable sequences, where a given residue can actually be more than one choice. Or even more complex example; a sequence with multiple variable positions, where you want to retrieve hits with at least one specific variation, or any combination of variations. How in the world would you do that with Blast? You can't! Not without a lot

of tedious, manual labor! This is where our MOTIF algorithm comes in – it allows specification of this type of sequence, and can be used to find exactly this type of results.

Trap 4: Wasting time with inefficient systems

Often result sets are sliced, diced, refocused, and refined many different ways. First analysts may try a % identity, then a combination, then a different combination, narrow by subject length perhaps, and drill down to specific jurisdictions or legal statuses.   Sometimes the starting parameters bring back too few or too many hits, and they have to be modified; the results evaluated by some method (number of families? Number of hits? Number of hits for a given jurisdiction?) and then the filters adjusted again and again until a satisfactory result set is obtained. All of this takes a LOT of time or can't be done in many systems! But it CAN be done quickly and efficiently in GenomeQuest, the result easily viewed and refined!

	Patent Number
EP1362925-0005 1-1 of 239 [View all 239 Results]	
	EP1362925
EP1362925-0006 1-1 of 49 [View all 49 Results]	
	EP1362925
EP1362925-0001 1-1 of 9,604 [View all 9,604 Results]	
	EP1362925
EP1362925-0002 1-1 of 9,307 [View all 9,307 Results]	
	EP1362925
EP1362925-0003 1-1 of 9,392 [View all 9,392 Results]	
	EP1362925
EP1362925-0004 1-1 of 9,307 [View all 9,307 Results]	
	EP1362925
EP1362925-0007 1-1 of 217 [View all 217 Results]	
	EP1362925

Trap 5: or wasting time with difficult to analyze search reports

Do you like to see alignments in search reports? Grouping results by patent family? A summary of which hits contain multiple query sequences (either by hit sequence or by hit patent)? How about seeing which sequences are identical across a patent family? Or a direct link to patent results? Having everything at your fingertips, including full Claims text and a link to a PDF? All of these are readily available in GenomeQuest!
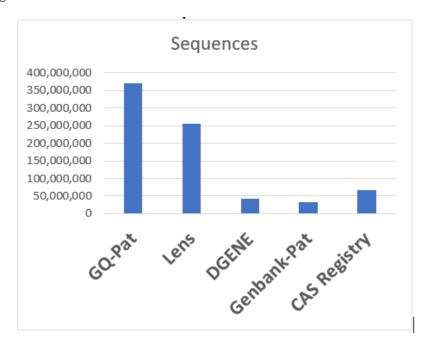
Trap 6: ⚠ **Assuming all databases are "close enough" that it doesn't matter which one you use – Nope!**

Every database is different! Each are updated at different intervals and have different coverages and degrees of data validation. **See here** for a detailed comparison. But some quick and scary facts:

•   Genbank omits sequences from US applications and JP patent proteins;

•   DDBJ, and databases taking feed from it, has applications originating at the JPO with inconsistent SEQ ID Nos compared to what is in the actual patent document. These errors are present in all public databases, as well as in Patent Lens

•   Many databases have long delays between sequence publication and adding them into the databases;

•   Most databases are incomplete; many only have sequences from sequence listings but omit the ones that have to be obtained through tedious mannual curation.

Trap 7: BELIEVING everything you see in a search report

Remember that a search report was prepared at a moment in time. How old is that report? And when was the search actually done? Often it takes time to prepare the report! Always look for the audit trail of a search result to see the search date. Legal statuses change regularly! That reference you didn't want to find may have published the day after the search was done; your competitor's application may have granted, or another application may have been abandoned. Always, always, for any document of interest, determine legal status via the corresponding patent authority, no matter what database was used! Set watches on applications of interest, set alerts on searched sequences.

Trap 8: Disregarding alignments in favor of numbers

True, you can't filter by alignment – but where content is of interest, sometimes the % identities don't tell the whole story. Review the alignment, which is easily accessible in GenomeQuest either through a hyperlink or in various report formats.

Trap 9: Using ANY sequence search system to replace a full text search

Sequence search systems are for searching sequence, and are not as richly annotated as text search systems. Sure, you can do a quick screen and review prioritization by using text to refine sequence search results, but ultimately, if a result passes the % identity screen and has been excluded in a sequence system by text searching, it should still be checked, just to be sure it isn't overlooked.

And finally, Trap 10: NOT using GenomeQuest

GenomeQuest guides you through all these different traps:

• Searching and analyzing by any combination of % identities;
• The largest number of sequences of any patent sequence database, public or commercial;
• Fast and efficient to use;
• Generates easy to analyze reports, with easy access to alignments;
• Availability of four different algorithms, and ability to use a different algorithm on a given search with a few clicks;
• Easy result view and grouping by query, patent, family, subject – so you can easily apply various filter parameters and see what works best;
• Awareness and correction of data integrity issues;
• User training, consulting, and technical support by experienced sequence search and patent professionals.

IP sequence searching is much more complicated than it seems on the surface; however, with sharp eyes, a good GPS, and GenomeQuest for your roadmap there's no reason you can't avoid the dangers lurking just beneath the surface of that nice, clean search report!

CONTACT

To contact us or for more information on how GQ Life Sciences can help you with patent searches, visit our website at: www.gqlifesciences.com.

GQ Life Sciences
4325 Alexander Drive Suite 100
Alpharetta, GA 30022