

Optimize Your Results: How to Achieve Comprehensive, Faster, Efficient IP Sequence Searching

If you're looking to protect your own sequences or want to make sure you're not infringing on anyone else's IP, you need to review what's already out there. Anyone who's done an IP sequence search before can tell you it's not an easy thing to get right. You want to make sure that you cover all the relevant data, that the data is searched in the right way, and that you have an efficient way to handle the results. You also don't want to accidentally disclose information that should stay confidential. Failure to bring these things together in the right way can impact search results, conclusions and, ultimately, lead to flawed business decisions. So, what are some of the common pitfalls in IP sequence searching and how can you avoid them?

A free IP sequence search usually involves the following steps: search the sequence database on something like the Patent Lens 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, an approach which is inefficient and certainly not comprehensive.

Search Everything You Can

The biggest challenge in IP sequence searching is finding a reliable, complete, and up-to-date source of patent information.

Commercial databases typically contain 100s of millions more sequences than the free options out there. Why is that?

A free option, like the Lens, consists mostly of the low-hanging fruit; sequences filed in ST.25 format and easy authorities. It misses out the more difficult-to-obtain countries and often foregoes sequences in tables and figures. In contrast, commercial databases are typically continuously updated with data streams from patent offices all over the world, including the US, Europe, China, Brazil, India, and the WIPO/PCT offices.

If you need an exhaustive search, for example for a freedom-to-operate or patentability search, you will also need to search public sequence databases like GenBank and RefSeq. This means going to multiple search sites and merging the results somehow. The better commercial solutions let you search and analyse everything at the same time, in the same way.

Free options are often an incomplete source of information, making it less than ideal for answering business-critical questions. It would be very easy to miss something if you are only searching free stuff.

“It goes without saying that **objective and repeatable search results** are important in IP. What does a percentage identity on a claimed sequence really mean if anyone can tweak it using the parameters of a sequence search algorithm like BLAST?”

Search the Right Way

BLAST, the most popular sequence search algorithm, was created with biology in mind. It answers questions like: “Is there an equivalent sequence in another species?” It’s less than perfect for IP-related questions that are typically phrased as: “Find all sequences in the database that are 70% or more identical to my query sequence.”

Note that the 70% identity must be computed over the length of the entire query sequence. All the nucleotides or amino acids in the query are equally important and all of them must be taken into account. BLAST simply isn’t designed to do this. It prefers to use only a piece of the query sequence if that means it can report a high percentage identity over that piece. Restricting the alignment length to a couple of residues will almost always produce 100% identity between any two sequences, but it doesn’t automatically mean anything for your IP position.

BLAST also doesn’t necessarily report all the alignments it finds. It is using a complicated statistical model that decides whether a match is significant or not. Among many other parameters, this model uses the length of the alignment and the database size in its decisions. When the database grows, things that have been found in the past can disappear, something that’s especially apparent when databases are large and alignments are small. It goes without saying that objective and repeatable search results are important in IP.

To overcome these issues, searchers should use “percentage identity” algorithms. These kind of algorithms align the entire query sequence – all of it – while minimizing the number of mismatches, insertions and deletions. They also do not use a statistical model or any algorithm shortcuts. In this way they always produce an objective and complete list of best possible results, regardless of database or alignment size.

If all you are using is BLAST, or some variation of BLAST, you are likely missing out on a lot of search results, especially if you search with smaller sequences like probes, primers, small RNAs, or CDR sequences.

Hits to Answers Immediately

Many sequence search applications present the outcome of a search as a long, static list of alignments. From that list there's no easy way to filter the relevant hits and retrieve information about the related IP documents. A common solution is to print out everything, go through the hits one by one, and look up related patent information on the web. Findings are scribbled on to the printout or put into a spreadsheet. This approach is labour intensive, error prone and inflexible. When the question changes, the entire procedure must be repeated from scratch.

Much preferable is a solution that presents the outcome of a search as an interactive list containing information about the alignments, the sequences, and the related IP documents. This should include important dates, patent title, abstract, claims, assignee, classification, and the legal status of a document. Such a system will allow things like find all hits with 70% identity or more over at least 500 nucleotides, where the sequence is claimed in a granted patent with the filing date before January 2010, group the remaining alignments by patent number and family.

Result analyses like these should be able to be adjusted and expanded at will, greatly increasing efficiency, with answers in minutes instead of days or weeks, allowing you to do more searches and do them earlier in the product development cycle. This also frees up search specialists and patent attorneys for other tasks, such as figuring out how the results impact the IP strategy of your company.

Keep Things Confidential

Everyone understands the confidential nature of IP searches, still there are lots of people who will submit confidential information to public web sites and not even blink an eye. All confidential information should be handled and stored on a secure private network. The communication between your browser and the servers should be fully encrypted. The data you submit should not be used for other purposes, and not be shared without your consent. When you pay for a commercial service this is a given.

Conclusions

Searching through sequence-related IP is a precise task. To do it right requires a comprehensive and up-to-date database covering applications and patents from as many sources as possible, using the right algorithms and parameters for the job. It also requires an efficient way to go from a list of search results to precise answers to the questions being asked. Any shortcut or incomplete solution is very likely to influence the outcome of a search, easily shifting the conclusions and your company's IP strategy.

For more information or help with an upcoming search, contact us at **info@aptean.com** or visit **www.aptean.com**.

About Aptean

Aptean is one of the world's leading providers of purpose-built, industry-specific software that helps manufacturers and distributors effectively run and grow their businesses. With both cloud and on-premise deployment options, Aptean's products, services and unmatched expertise help businesses of all sizes to be Ready for What's Next, Now®. Aptean is headquartered in Alpharetta, Georgia and has offices in North America, Europe and Asia-Pacific.

To learn more about Aptean and the markets we serve, visit **www.aptean.com**.