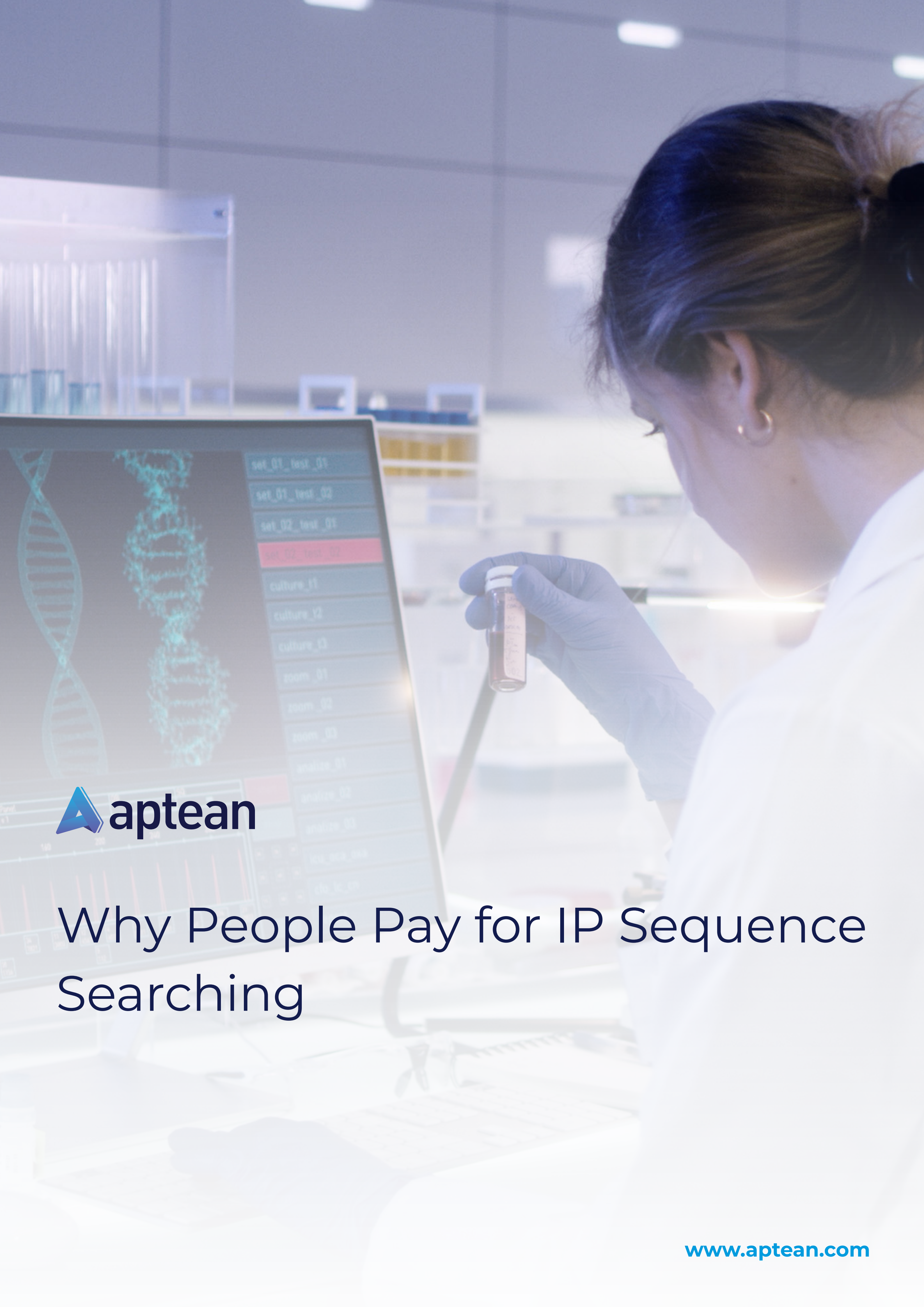

aptean

Why People Pay for IP Sequence Searching

www.aptean.com

If you're looking to protect your own sequences or want to make sure you're not infringing on anyone else's IP, then you need to review what's already out there. Anyone who's done an IP sequence search before can tell you that it's not an easy thing to get right. You want to make sure that you cover all the relevant data, that this 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. Let's look at some of the common pitfalls in IP sequence searching and how to overcome them.

A free IP sequence search usually involves the following steps: search the sequence database on the Patent Lens website, go through the alignments one by one, and lookup 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.

Search Everything You Can

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

As of April 2020, the Lens contains about 352 million sequences. As a comparison, GQ-Pat, Aptean GenomeQuest's IP sequence database, contains a 452 million sequences. That is an additional 100 million sequences.

The Lens consists mostly of the low-hanging fruit, sequences filed in ST25 format and easy authorities. It misses out on more difficult-to-obtain countries and often foregoes sequences in tables and figures. In contrast, our database (GQ-Pat) is continuously updated with data streams from patent offices all over the world, including the US, Europe, China, Brazil, India, and the WIPO/PCT offices. Of course, GQ-Pat also includes the usable parts of databases like GenBank, EMBL, and DDBJ.

In comparison to GQ-Pat, the Lens is a very incomplete source of information, making it less than ideal for answering business-critical questions.

Search the Right Way

It's important to understand that 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 is 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 has to 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 have to be taken into account. BLAST is simply not designed to do this. It prefers to only use a piece of the query sequence if that means it can report a high percentage identity over that piece. To illustrate this point: restricting the alignment length to a couple of residues will almost always produce 100% identity between any two sequences, but it doesn't necessarily answer your query.

BLAST does not necessarily report all the alignments it finds. It has 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. This is 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, Aptean GenomeQuest developed and published the GenePast "percentage identity" algorithm. It aligns the entire query sequence – all of it – while minimizing the number of mismatches, insertions, and deletions. Because it doesn't use a statistical model or algorithm shortcuts, GenePast always produces an objective and complete list of best possible results. It does this regardless of database or alignment size.

Over the two decades, the GenePast algorithm has seen an enormous uptake by a long list of life science companies and patent offices to complement, or even replace, BLAST in IP sequence searching.

Go From 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 is 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 labor intensive, error prone, and very inflexible. When the question changes, the entire procedure has to be repeated from scratch.

GenomeQuest presents the outcome of a search as an interactive list that contains information about the alignments, the sequences, and the related IP documents. This includes important dates, patent title, abstract, claims, assignee, classification, and the legal status of a document. Because of this you can, with a couple of clicks, 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. If desired, you can then group all alignments by patent number and patent family to see all related results together. Result analyses like these can be adjusted and expanded at will. This greatly increases efficiency, allowing you to get answers in minutes instead of days or weeks. It allows you to do more searches and do them earlier in the product development cycle. It also frees up search specialists and patent attorneys for other tasks, like figuring out how the results impact the IP strategy of your company.

GenomeQuest has many features to make this work easier, including customization of the way search results are displayed, the ability to share results with colleagues, Word and Excel report generation, and an alerting service for monitoring new results over time.

Keep Things Confidential

At Aptean GenomeQuest, we understand the confidential nature of IP searches. All submitted data is handled and stored on a secure private network. The communication between your browser and our servers is fully encrypted, and your user account is protected by a password. This means that you are the only one able to see your data, unless you explicitly share it with someone else in your organization. The same cannot always be said for a public service.

Conclusions

Searching through sequence-related IP is a precise task. To do it right requires a comprehensive and updated database covering applications and patents from as many sources as possible. It requires 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 and can easily shift the conclusions and your company's IP strategy.

Over the last two decades, GenomeQuest has been used by many of the largest pharmaceutical, biotech, and agricultural companies in the world, by specialized law firms, and even by patent offices themselves. We would love to help you protect your IP as well.

To learn more or to schedule a demo, email **info@aptean.com**.