LSIPR
LIFE SCIENCES INTELLECTUAL PROPERTY REVIEW

Searching for sequences: a high stakes game

26-10-2017
By Ellen Sherin

Those working in the field of genetic sequence must clearly understand the subtleties of sequence searching, in order to avoid the risks of infringement and invalidity, as Ellen Sherin of GQ Life Sciences finds out.

The intent of patent searching is to ask: "what's out there that is related to my invention?" It could be prior art affecting patentability; it could be claimed IP (granted or not) with its associated risk of patent infringement. The search results, and interpretation by the IP practitioner, are key factors when deciding the fate of the invention or technology being searched.

The potential financial implications of this type of decision can be enormous. According to the AIPLA 2015 *Report of the Economic Survey*, the median costs of patent litigation through trial and appeal ranged from $600,000 for less than $1 million at risk, to above $5 million for more than $25 million at risk. That's just legal costs, without consideration of damages for infringement. Idenix (Merck) and Gilead's litigation of infringement and invalidity on multiple patents for hepatitis C drugs is an example. Gilead's August 2017 Form 10-Q estimates its potential liability for the various legal actions up to $9 billion.

One case (now on appeal) awarded a $2.5 billion judgment to Merck (with potential for future triple damages). But Merck also lost some other litigations: two Merck patents were invalidated in favour of Gilead's patents—one for lack of enablement, the other for prosecutorial misconduct—and the real winners are ultimately the law firms. *American Lawyer's* July 18, 2017 edition reports a $12.5 million legal fees award in just one of these cases.

Genetic patents are not exempt. In March 2017, an appeal confirmed Bayer's $469 million award against Dow for patent infringement regarding use of a gene conferring glufosinate resistance in plants.

These cases exemplify the high stakes in IP law, and the need to "get it right" the first time—whether to determine freedom to operate, patentability, to decide to kill or advance a research project, or to develop an opinion on the validity of a given patent.

Let's take a closer look at how sequences are addressed in patents. On a very basic level, they are treated as text strings, and the descriptor "percent identity" is used to describe how many sequence letters match. Percent identity is often used in claims, and is a key screening parameter when searching.

Not the whole story

Of course, it isn't quite that simple. The complexity comes when matching up these long strings when there are differences. That's where sequence search algorithms come in. Different algorithms handle matching differently, and the percent identity will also differ as a function of algorithm choice and parameters.

A percent identity claim based on the Smith-Waterman algorithm, which is evaluated compared to a BLAST search result, may lead to erroneous conclusions. Furthermore, sequence search algorithms have settings which affect the percent identity, and even the type of sequence they may find (or miss). Alignments created with the same algorithm but different parameters often give different results.

The percent identity for a search result, when calculated according to a patent's definition, may be within a claim scope, but if the search algorithm or parameters are different, the percent identity may appear outside the claim scope, or vice versa. As a result, either an incorrect clearance may be given, or a promising project may be cancelled.

Missed hits

The BLAST algorithm is the most commonly used algorithm for sequence claims and is also frequently used for sequence searching, so depending on the query sequence's characteristics, that is often a good starting point. However, there are types of query sequences which require either BLAST parameter adjustments or different algorithms for the most complete results.

Short sequences such as complementary determining regions, probes, primers or other short sequences are better searched with an algorithm like GenePast, as BLAST misses hits unless parameters are adjusted for short queries.

Genomic sequences often break into pieces called multiple high-scoring sequence pairs (mHSPs) when aligned against non-genomic DNA. Sequences with significant insertions or deletions may also, as will sequences with relatively long regions of mismatch between matched regions. Depending on the searcher's knowledge and the search product used, mHSPs may be detectable, but the searcher and the IP practitioner must understand how to identify and to evaluate these results.

One approach is GenomeQuest's "query % HSP coverage" field, which combines the different pieces into a group with an overall calculated percent identity, to give the practitioner a preliminary idea of the potential relevance of each group of HSPs.

Variant searches, where positions may have multiple substitutions, are particularly difficult. Consider US20150087572A1, claim 1: "said parent protease amino acid sequence being identical to the amino acid sequence of SEQ ID No:1, said variant protease of said parent protease mutations consisting of one of the following sets of mutations versus said parent

protease: (i) N76D+S87R+G118R+S128L+P129Q+S130A; (ii) N76D+S87R+G118R+S128L+P129Q+S130A+S188D+V244R; …"

Sequences of this nature can't be properly searched by BLAST. These queries are written in a special manner (VAGTIAAL[ND]NSIGVLGVAP[SR]AELYAV), which BLAST doesn't understand. They should be searched using a special algorithm, most typically GenomeQuest's MOTIF algorithm, which identifies sequences comprising either wild type or variant in the specified positions. The only way to search these with BLAST is to use an X at each variant position, and then manually review all hits for the presence of specific variations—a daunting task.

As any searcher knows, the true art of searching is in trying to think of the different ways the search concept may have been described, and to create queries that capture the different, unique, and unusual patent language, and so it goes with sequence work. Sequence claims, just like chemical structure claims, may be written as Markush structures (variable sequences, as above), or as substructure claims (comprising language), or even as a combination of the two in addition to the intuitive full length sequence claim.

Here are some examples:

1. **A simple "comprising" claim: Claim 1. An isolated nucleic acid comprising SEQ ID No: 1.**

   A 100% identity hit, identical sequence and identical length is always significant. However, here's a more complex example: Query sequence comprises SEQ ID No: 1 as part of a much longer sequence. Alignment and subject % identity identify this hit as significant; but a long query sequence may give a low query % identity, and in reviewing a large search report, the result could be overlooked—eg, if SEQ ID No: 1 is short, perhaps 25 residues, and is present in multiple long sequences.

   For a hit of this nature, the alignment percent identity may be useful but even more useful is the subject % identity, which identifies a full length hit to SEQ ID No: 1 as significant, regardless of the length of the query sequence.

2. **A position-specific comprising claim: US6867026B2, claim 1**:

   An isolated protein having glucosyltransferase activity comprising an amino acid sequence, which exhibits at least 95% amino acid homology, as determined by BLAST algorithm, with the amino acid sequence 972-1781 of SEQ ID No: 2.

   Again, the significance of this type of hit may be missed. The subsequence is at best, ~55% of the full-length sequence. With a long query aligning only to that region of sequence, but the "dead-on" significance of that long alignment won't be evident. In order to determine whether any result is within the scope of this claim the query/subject pairs should be filtered by subject coordinates. This is especially powerful with a large number of query sequences being searched simultaneously, and essential once hits like this one are found.

3. **Variant claims need special search methods**

   US20100192985A1 is just one example of many with multiple pages of recited variations. Typically the backbone(s) and perhaps a few variants are included in the sequence listing, with many more set forth in text. No sequence database has even a small fraction of these variant sequences indexed. Relying on sequence searching alone for variants is a huge trap and an invitation to disaster. In addition to using the MOTIF search methodology described above, in combination with some smart Boolean work to remove wild-type and out of scope variant hits, excluded sequence search results and text queries must be combined in a full text database in order to perform a thorough search.

Percent identities at the heart

The term "percent identity" is frequently used in patents claiming biological sequences, but it has multiple meanings. In many cases, it is not defined in the specification, leaving the examiner, and potentially the practitioner and litigators, to try to determine the meaning of its use.

The standard percent identity displayed by BLAST is the alignment % identity, which is independent of the alignment length. Claims language actually often means the query % identity (considering the claimed sequence as the query) but not always, and often fails to state which % identity is meant.

Consider the following examples, all with 100% alignment identity, which means that all the residues that align really match, but there can be many more residues that do not match.

> **Example 1:** Query and subject each with 1,000 residues; 100 residues match perfectly.
>
> **Example 2:** Query is 100 residues, subject is 1,000 residues. Subject comprises query at 100% alignment identity.
>
> **Example 3:** Query is 1,000 residues, subject is 100 residues. Query comprises subject at 100% alignment identity.

And now consider these two sample claims from different, unrelated patents:

> **Sample patent 1:** Claim 1. An isolated nucleic acid comprising SEQ ID No: 1 at 90% identity.
>
> **Sample patent 2:** Claim 1. An isolated protein comprising an amino acid sequence, which exhibits at least 95% amino acid homology, as determined by BLAST algorithm, with the amino acid sequence 972-1781 of SEQ ID No: 1.

Now consider a search returning hundreds or thousands of sequences, with query sequences that are either much longer, much shorter, or the same length as SEQ ID No: 1. Without fully understanding these percent identities, how to combine them in a Boolean statement along with other sequence terms, and how to interpret the results—all of which take skill and experience—legitimate hits may be missed. And that doesn't even begin to address algorithm and parameter choice.

Biological sequence IP is an immense field and is ever-growing. There are more than 375 million sequences in patents alone, and many millions more in non-patent databases, just waiting to be used as prior art to reject a patent before grant, or invalidate it after. By sheer data volume, the probability that any inventive sequence will have multiple hits is substantial.

It's essential that practitioners and searchers working in the field of genetic sequence clearly understand the subtleties of sequence searching: the algorithms and their parameters, the different percent identities and their relationship to claim construction and interpretation, the screening capabilities of different commercial products, and the need for clear definitions in writing sequence claims.

Without this knowledge, there is the risk of infringement on one hand, and invalidity on the other—or conversely, unnecessarily killing a promising new product.

Ellen Sherin is senior product manager at GQ Life Sciences. She is a registered US patent agent, and currently serves as the product manager for GenomeQuest and LifeQuest. Before her appointment at GQ Life Sciences, Sherin worked for a Fortune 500 company for over 35 years. She can be contacted at: ellen.sherin@aptean.com.