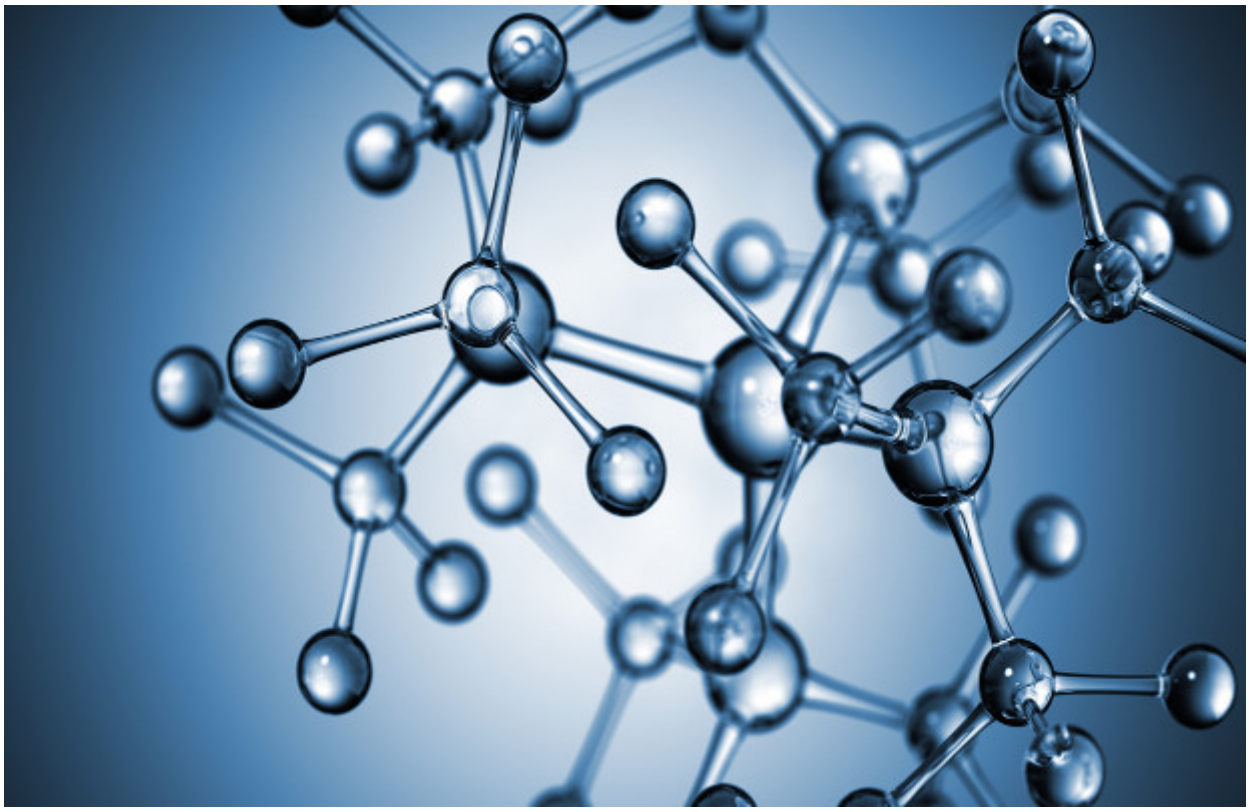
LSIPR
LIFE SCIENCES INTELLECTUAL PROPERTY REVIEW

Variant sequences: the new frontier in sequence IP

12-04-2018

Ellen Sherin



blackjack3d / iStockphoto.com

Variable sequence IP is fertile ground for potential validity challenges based on prior art or lack of enablement, as Ellen Sherin of GQ Life Sciences reports.

A variant sequence is typically a naturally-occurring molecule, most frequently protein, which has been modified in order to impart desired characteristics such as improved stability under anticipated use conditions, or increased activity. Optimising sequences and, of course, the associated IP rights, span most areas of biotechnology.

Defining the scope of one's IP in these areas is analogous to homesteading in the US during the 19th century: just as settlers raced to claim land in the 1800s because it was available, in the 21st century, companies compete to claim a broad variant IP landscape. Possession and ownership of large tracts of land was easy to prove; in the case of IP law, possession and ownership of large numbers of sequences is not quite so easy.

Patents are written that contain many millions of variations. "Best mode" is buried in millions of other options, and enablement, reduction to practice, and possession may all be called into question.

A case in point is *Novozymes v DuPont Nutrition Biosciences*, from 2013. Patent US 7713723, filed by Novozymes, claims priority to US 60/249,104, filed in 2000, reciting variations to seven different parent alpha-amylases, 33 positions, and multiple potential variations at each position. Although initially Novozymes received an award of $18 million for infringement, upon appeal this judgment was reversed and the patent held to be invalid under §112 for failure to satisfy the written description requirement.

The US Court of Appeals for the Federal Circuit found that the specification of the priority filing covered "a potentially enormous number of alpha-amylase variants". Although the '723 patent claimed specific variations, the 2000 priority filing did not disclose those specific variants, and that is at the heart of the verdict.

Searching

Let's switch to searching for IP of this sort, whether for freedom to operate (FTO), patentability, patent examination, or invalidity based on prior art. If a company would like to commercialise a specific variant, or group of related variants, one of the elements in the decision to move forward is an IP white space search, along with patentability and FTO.

Ordinarily a sequence search is relatively straightforward to the experienced searcher: prepare sequences, select proper algorithms, perform search, screen results. Because prior art sequences will most likely be indexed in at least one commercial sequence database, they will be found with a properly executed search. It's also appropriate to text search, but most likely if a sequence exists that is similar to the queried sequence, it will be found during the sequence portion of the search.

When working with variant sequence IP, this is not the case. Variants are frequently recited in the specification and claims as multiple textual lists of differing combinations of variations, sometimes taking coordinates from multiple backbones, sometimes not.

Even when the text descriptions appear standardised on the surface, a quick dive into a dozen related patents will show subtle to significant differences in the notation, which challenge even the most experienced searcher.

Formulae complicate these still further: who knows whether a given sequence will be described explicitly, as a Markush, as a formula, or with free text? A few words can describe hundreds, thousands, or

millions of sequences, almost none of which are in the sequence listings or in commercial databases. Inventors try to claim every possibility, but then are challenged to show enablement.

The following example illustrates these points. N76D + S87R + G118R + S128L + P129Q + S130A, a single subtilase variant expression, which, if expanded translates to 729 possible sequences. This is only one of over 100 of these strings in a single patent. These patents routinely comprise tens, hundreds, or even thousands of sequences of this nature, so by sheer mathematics it's clear that the majority of such sequences will not be present in the sequence listing or in commercial sequence databases, and thus won't be found by a sequence search. At best, the wild type precursor or backbone may be found, along with a handful of exemplary variants.

Formulae

Formulae are a further complication. In addition to varying the amino acids, they may also recite sequence dependencies, indels, or even regions of variable length, including complete absence. Here are two examples:

Claim 1 from CA2756249 "The amino acid motif: .PHI.4-X-.PSI.-L-[T/A]-.PSI.2, wherein .PHI. and .PSI. are in each instance an independently selected hydrophobic non-aromatic amino acid, and X is any amino acid" which is then further narrowed in claim 6: ".PHI. is in each instance independently selected from the group of amino acids consisting of alanine, methionine, isoleucine, leucine, and valine, preferably methionine, isoleucine, leucine, and valine."

Claim 1 from US 20180087081 "X1-X2-X3-X4-X5-X6-X7-X8-X9 (SEQ ID NO:47), wherein: X1 is isoleucine (I), leucine (L), valine (V), alanine (A), or methionine (M); X2 and X3 are each independently any amino acid with the proviso that one or both are K or R; X4 is any amino acid or X4 may be absent when X1 through X3 are present and X5 through X9 are present; X5 is tyrosine (Y), tryptophan (W), or phenylalanine (F); X6 and X7 are each independently any amino acid with the proviso that one or both are lysine (K) or arginine (R); or either one of X6 and X7 may be absent when the other is K or R and when X1 through X5 are present and X8 and X9 are present; and X8 and X9 are any amino acid with the proviso that one or both are leucine (L) or alanine (A); or one of X8 and X9 may be absent when the other is L or A and when X1 through X7 are present."

"Variants are frequently recited in the specification and claims as multiple textual lists of differing combinations of variations."

Variable-length regions are especially troublesome for sequence searching. In the sequence listing, these are written with one explicit sequence length, and the varying length only annotated as a feature, which is apparent on reading the sequence listing, but is not present as all sequences encompassed by these options. In addition, the length definitions may only be in the text, but not in the sequence listing at all. Formulae like these may encompass a large number of potential sequences, the majority of which won't be found by a sequence search, or even by a text search because of the complete absence of uniformity of notation.

All variations

Just as it's unrealistic to index all these sequences explicitly in sequence listings, it's unrealistic to write all combinations of query sequences as explicit queries. Fortunately, products such as GenomeQuest and STN have algorithms that allow these Markush sequences to be written a single time, but which include variations. After removing wild type hits, the remaining answers will comprise all possible variations within these constraints.

It can be further expanded to allow for any variation at positions of interest by simply using X in place of the specific variations, and removing hits containing no specified variations. This methodology is sound, and if a specific variant sequence (or one containing Xaa in place of a specific variant) has been indexed in the sequence listing, it will be found. But, what if the sequence of interest is claimed, but is buried in a long list of variations that was not indexed?

For this reason, it's essential to include text searches in any variant sequence IP analysis. That will be helpful in finding some of the sequences described above, with the combination of letters and numbers, such as N76D, but there are so many different ways to write these combinations that there is still no assurance the majority of legitimate hits will be found.

Finally, the multi-disciplinary skill required to perform these searches shouldn't be underestimated. These searches are best performed by specially-trained, highly experienced searchers, who understand the limitations of current technology and the lack of uniformity in notation. They are extremely time-consuming and labour-intensive if done properly.

The implications of the difficulty in searching this type of art are sobering and exciting—sobering for companies complacently sitting on large amounts of sequence IP; exciting for newcomers to the field who want an opportunity to invalidate others' IP.

Variable sequence IP is fertile ground for validity challenges based on prior art or lack of enablement. If methodology were to be developed to properly index even a percentage of these sequences, there will be a huge body of previously undiscovered art available for mining, litigating, and invalidating. Patent professionals with expertise in this area of sequence IP are well-positioned for future work.

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