Ten Mistakes to Avoid When Searching Gene Sequences

Introduction

Over the past 25 years, the GenBank genetic sequence database has doubled in size every 18 months. While this statistic in itself is both impressive and challenging, it tells only part of the story. Public data resources are often thought of as being complete, or mostly complete, but actually hold only a fraction of all sequence data relevant to any serious biological research inquiry or intellectual property (IP) portfolio evaluation. Additional sequence data are piling up rapidly in proprietary databases, in-house databases, desktop hard drives, graphic images and illustrations, and print document collections. In particular, patent sequence information is substantially underrepresented in public sequence data resources, which themselves are often missing published sequence data and include many poorly annotated, incomplete records. Each year, pharmaceutical companies, biotech organizations, academic institutions, and law firms spend millions of dollars on service fees and in-house database development efforts to collect, organize, and maintain sequence data. To the extent that the resulting resources are incomplete, out-of-date, or unusable, additional costs are incurred in the forms of misdirected laboratory effort, annulled biological research discoveries, poorly informed IP portfolio decisions, weakened IP claims, broken patents, and delayed drug development project termination. Avoiding sequence search mistakes is imperative. Here are ten of the most costly errors:

1) Overlooking Patent Sequence Data

Any serious search of sequence information requires a specific and organized effort to include patent sequence data. It is not enough to search Genbank. Genbank and the other commonly accessed sequence databases contain only a partial representation of available patent sequence data, and the data they do contain is often incomplete, poorly annotated, and out-of-date. Many Genbank sequence records contain almost no annotations and are therefore missed by keyword-based search queries. The Genbank Patent Division currently holds approximately 4 million sequence records. By comparison, the GenomeQuest GQ-PAT patent sequence database has over 66 million unique patent number-sequence pairs and over the past 12 months has nearly doubled in size. Relying on a "quick Genbank search" to make scientific or legal decisions is a mistake to be avoided. Due to application processing delays at the US Patent Trademark Office (USPTO), and vigorous genome industry efforts to find, annotate, and market patent sequence information, it is unlikely that Genbank will catch up to the specialized commercial patent sequence collections any time soon.

2) Searching Old Data

Typical genome sequence, annotation, and patent information databases are out-of-date the moment they are ready for use. Conventional genome data resources rely on RDBMS (relational database management system) software, which must rebuild the database in its entirety whenever changes to the database, such as the addition of new information, are made. Due to this limitation, conventional database management systems are incompatible with rapidly expanding information, such as genome sequence data. In an attempt to address this problem, some providers of genome information have broken down the data into a dozens of smaller databases, each with different update and rebuild cycles. While smaller databases do, in fact, rebuild more rapidly, each one is always out-of-date, to a greater or lesser extent, based on the date of its last "build" and the worldwide rate of new genome sequence, annotation, and patent information production. When ready search capability against rapidly expanding information is a functional product requirement, newer data mining and text analytics methods for managing "unstructured" data are used instead of relational database management. The vast majority of scientific and commercial information, including genome information, consists of "unstructured" data. A well-known example of an unstructured data system is the Google search engine.

3) Under-utilizing Annotations Information

Sequence annotations are valuable for defining the range of a sequence query, either before or after the actual search. Unfortunately, the sequence annotation records in many genome information databases are incomplete or disorganized. When sequence annotation selection criteria are applied prior to a sequence search, the lack of complete, accurate annotations will result in the unintentional omission of potentially relevant sequences from search query results. Similarly, if annotation information is used to sift and sort the results of a sequence similarity search, missing or inaccurate annotations can cause potentially relevant results to be discarded or miscategorized. Ascertaining the legal or biological importance of the similarity between any two sequences requires a clean, curated database with organized annotation fields and content. Additional fields, such as bibliographic references, date of earliest publication, and date of sequence disclosure add analytical speed and precision when used with a rapid search result filtering function.

4) Forgetting the "Dark Genome."

Public BLAST portals search only the most readily-accessible elements of the entire universe of genome data. The remaining information is sometimes referred to as the "dark genome." Poorly annotated data in a readily accessible database may be considered part of the dark genome, in that is "hiding in plain sight." Additional data with low search accessibility includes the information held in proprietary databases, desktop hard drives, graphic images and illustrations, and print document collections. Searching the "dark genome" requires access to proprietary data and full-time, multiple-media genome information searching and database curation.

5) Taking Too Much Time

Taking too much time to do a scientific or patent-related sequence search is a root cause of research project and intellectual property decision delays. Researchers might spend weeks scouring the internet for new or "dark genome" data related to a query sequence, or developing lists of databases holding separate or overlapping sets of genomic information. Unintuitive search software user interfaces cause "learning curve" delays, and sequence search outsourcing can cause vendor transaction and project scheduling delays of up to 30 days. Inability to access or utilize sequence record annotations can turn a 20-minute sequence result-filtering project into a days-long process of manually sorting and sifting through sequence similarity query result printouts.

6) Hoping for the Best

Moving forward with a research project without first searching for and evaluating related sequences is a mistake that produces adverse effects on scientific research and IP portfolio management. Failing to get complete sequence search information prior to designing or carrying out research experiments can result in misdirected laboratory efforts and annulled research discoveries. An incomplete evaluation of the sequence patent landscape early in the research cycle, when effective strategies for licensing or working around competitive IP could have been designed and implemented, imparts great costs when completed projects are found to have yielded unusable results.

7) Making Decisions Based on Yesterday's Search Results

Genome sequence information is extremely dynamic. In addition to the steady addition of recorded primary sequence data, scientific and patent information about both new and previously existing sequences also grows and changes on a daily basis. A sequence data query affecting important scientific research and business decisions might not yield the same answer one week from now. The more sequences involved in the decision, the greater the risk. Research groups and businesses without access to an automated and continuous search-and-report system are particularly vulnerable.

8) Using the Wrong Algorithm

Even experienced sequence analyzers often make mistakes in choosing the right algorithm for a sequence search. For example, searching with short sequences using BLAST is a mistake because the BLAST algorithm's heuristic approach misses many approximate hits. BLAST will exclude good hits from search results if they do not produce a perfect match over a set "word size." BLAST will even exclude exact matches when their alignments fail to meet the algorithm's requirement for statistical significance. When search parameters are revised to make heuristics more lenient, BLAST returns too many hits, many of them irrelevant.

9) Too Many Gatekeepers

Restricted access rights to proprietary databases, cumbersome search software user interfaces, and outdated business practices often prohibit direct utilization of sequence data search systems by the person asking the question, who must instead work through one or more gatekeepers. A well-defined project submission process can prevent intended queries from getting "lost in translation," but when sequence searches are outsourced, queries are often composed broadly in order to prevent potentially relevant results from being excluded from the search report returned by the service. This results in an oversized report and a long manual search process for the sequence records of real interest. Gatekeeper delays also inhibit creative sequence data exploration, where hunches and hypotheses can be quickly formed and investigated using fast, iterative database queries.

10) Ignoring Workflow Issues

Commercially licensed or in-house bioinformatics solutions often become very popular within organizations as researchers learn to use them to great advantage. But an effort to provide genome search capability to the user base that does not consider workflow issues can result in the installation of an isolated, standalone information "silo" with an unfamiliar interface. The standalone solution is itself likely to be underutilized, and also fails to take advantage of organizational knowledge built up around previously existing bioinformatics applications.

To get more information about GenomeQuest 4.0, visit our home page at www.genomequest.com, or contact us at contact-us@genomequest.com.