Sequence Screening

Robert Jones
Craic Computing LLC, Seattle, Washington

Cite as:

Jones R. 2005. Sequence Screening. In: *Working Papers for Synthetic Genomics: Risks and Benefits for Science and Society*, pp. 1-16. Garfinkel MS, Endy D, Epstein GL, Friedman RM, editors. 2007.

The views and opinions expressed are those of the author of the paper.

(this page blank)

Sequence Screening

Robert Jones
Craic Computing LLC, Seattle, WA

Introduction

The use of biological agents in acts of terrorism has received heightened interest since the mailing of anthrax spores in the United States in 2001. Many scenarios have been considered in which bacterial and viral pathogens could be produced and employed as weapons. While some may seem unlikely, the scientific community has a responsibility to assess all threats and to develop ways to monitor, and perhaps counter, any attempts to carry them out. The focus of the current study concerns the use of DNA synthesis and genetic manipulation to create or modify pathogens.

It can be argued that a terrorist group would be much more likely to use a 'conventional' pathogen, such as anthrax, than to design and engineer a modified organism. While this is convincing, there are several strong reasons why someone might wish to employ synthetic DNA. Conventional threats require that the terrorist has access to the pathogen. Some pathogens, such as anthrax, can be isolated in the wild in certain parts of the world. It is clearly possible to culture natural isolates, but the process can be laborious and may yield a strain that is not well suited for use as a biological weapon. In most cases, the easiest sources for pure cultures of pathogens are the laboratories that work on them. In the US such labs are strictly regulated with measures such as background checks on researchers, careful inventory management and high levels of building security. These make it extremely difficult for anyone outside those laboratories to access the pathogens that they contain. Smallpox virus is an example of a pathogen that, having been eradicated in the wild, could only be obtained from a few specific laboratories, all of which operate under tight security.

The alternative approach that concerns us here is that someone could synthesize the entire genome of a dangerous pathogen, such as smallpox, from scratch. This requires no access

to the secure laboratories. Potentially it requires no prior experience in working with the pathogen. Most troubling is the fact that such synthesis could be accomplished in a conventional molecular biology laboratory, without the need for specialized equipment and without attracting attention to the project from others.

The technology required to synthesize the genome of an entire viral pathogen, or genes thereof, is already available. Rapid development in the field of synthetic biology is destined to make this process easier, faster and cheaper.

This evolution in technology brings with it tremendous benefits to biotechnology and medicine but its potential for abuse is a cause for concern. Being able to determine if nefarious activity is underway will become an important requirement for the regulatory authorities.

Here there is some cause for optimism. Currently the vast majority of DNA synthesis is performed by service companies or by in-house central facilities in universities and large companies. The DNA synthesis industry provides researchers with custom DNA at such low cost and with such convenience that almost all synthesis work takes place in a relatively small number of facilities.

A request for DNA synthesis requires that the customer provide the sequence of the molecule. This creates the opportunity to monitor or screen input sequences for matches to a database of pathogen sequences. Finding a positive match at the time the order was received would allow the vendor to alert the relevant authority and to delay shipment of that DNA.

I have written a software package, called BlackWatch, that implements sequence screening. This paper will describe the operation of this system, its current shortcomings and ways that these might be addressed.

1. The Business of Synthetic DNA

At this point it is worth reviewing the state of the synthetic DNA industry as it stands today. Not only does this provide the venue in which to monitor attempts at engineering

pathogens, but its particular constraints and operating procedures have a significant practical impact on the way any sequence screening strategy might be implemented.

The chemical synthesis of oligonucleotides (oligos), short fragments of DNA, became widely available about 20 years ago with the manufacture of desktop DNA synthesizers. Oligos found widespread use in DNA sequencing with an equal, and perhaps greater, application in Polymerase Chain Reaction (PCR) experiments. The high demand for oligos led to the creation of companies that performed contract DNA synthesis on request. The convenience and low cost of using these vendors has driven substantial growth and competition in this industry and today hardly any research laboratories synthesize oligos themselves.

Fierce competition between the synthesis companies has driven prices down to the point where profit margins are minimal. In fact certain companies appear to offer the service as a 'loss leader' in order to attract customers to their other more lucrative products. These companies try to differentiate themselves on the basis of easy ordering via the Web, fast turnaround and value added options, such as chemical modifications of the oligos. The customer can visit the web site of the vendor, create an account, enter in the sequence of the oligo they want synthesized, enter their credit card details and hit submit. A tube containing the DNA will arrive by express delivery the next day or day after. The cost for this service is remarkable. A typical oligo of perhaps 20 to 25 nucleotides in length will cost around $0.30 per nucleotide, a total of less than $10 for a completely custom organic chemical synthesis. As a result, using one of these services is the preferred option for almost all laboratories.

There are probably several hundred companies or university facilities that offer oligo synthesis services around the world. The throughput of the larger companies is impressive. Integrated DNA Technologies (http://www.idtdna.com) of Iowa, states in its press releases that it synthesizes between 15,000 and 25,000 oligos daily and has more than 60,000 customers worldwide.

With improvements in the technologies behind DNA synthesis and gene assembly, it has become feasible to synthesize entire genes from sets of oligos. Several companies provide this service, some of which derive their entire income from gene synthesis. The technology is more involved than oligo synthesis but costs can be kept low, while maintaining accuracy, through the extensive use of laboratory automation. The distinction between companies that synthesize short oligos and those that synthesize entire genes, assembling these from sets of oligos, is important in the context of sequence screening.

The turnaround time for the synthesis of a gene of a few thousand nucleotides is a couple of weeks and the cost can be as low as $1.60 per nucleotide. At this price point it becomes easier to synthesize certain genes than to try to isolate them from their native genomes. There are around 25 companies in the US that offer this service with about the same number in the rest of the world, mostly in Europe. However, it would appear that most of that work is performed in a small subset of these companies.

The next step in the evolution of these technologies is the synthesis of entire genomes. Already the genomes of poliovirus (Wimmer et al. 2002. *Science 297*: 1016-1018) and bacteriophage phiX174 (Smith et al. 2003. *Proc. Natl. Acad. Sci. USA 100*: 15440-15445) have been synthesized from scratch and used to create infectious virus and phage particles, respectively.

Work is underway at the company Synthetic Genomics (http://www.syntheticgenomics.com) and at the J. Craig Venter Institute to identify the minimal set of genes that are necessary to sustain the bacterium *Mycoplasma genitalium*. Once this minimal genome has been defined, the company intends to use it as the foundation for a range of engineered synthetic organisms that possess novel characteristics.

If the history of DNA sequencing, PCR and oligo synthesis serve as a guide, we can expect the synthesis of genes and small genomes to become routine tools for molecular biology over the next decade.

One final aspect of the business of synthetic DNA is of particular importance. Confidentiality and the protection of intellectual property are extremely important to the biotechnology industry. Oligo vendors help ensure confidentiality by not asking customers about the nature of the sequences that they request or the uses to which they will be put. Indeed, most corporate customers would immediately stop using these vendors if they were required to disclose any information about the requested sequences.

This intentional ignorance about the sequences on the part of the vendor could play into the hands of anyone intent on synthesizing or engineering a pathogen. Widespread use of sequence screening software has the potential to remove this vulnerability while still retaining confidentiality for the vast majority of DNA synthesis customers.

.

2. Sequence Screening

The basic idea behind sequence screening is straightforward. Sequences of oligos or entire genes that are to be synthesized are compared against a specific curated database of sequences from known pathogens, the 'Select Agents'. Any request that produces a significant match to a pathogen is tagged as being of interest and the site administrator is alerted.

I have implemented this approach in the BlackWatch software system. This consists of a custom sequence database, the BLAST sequence comparison software from NCBI (Altschul et al., Nucleic Acids Res. 1997, v25, pp3389-3402) and a set of Perl wrapper scripts that manage the user interface, run the BLAST searches and process the results.

The system can be accessed from a web interface, the UNIX command line and from custom interfaces to relational databases. A schematic diagram of the system in shown in Figure 1.

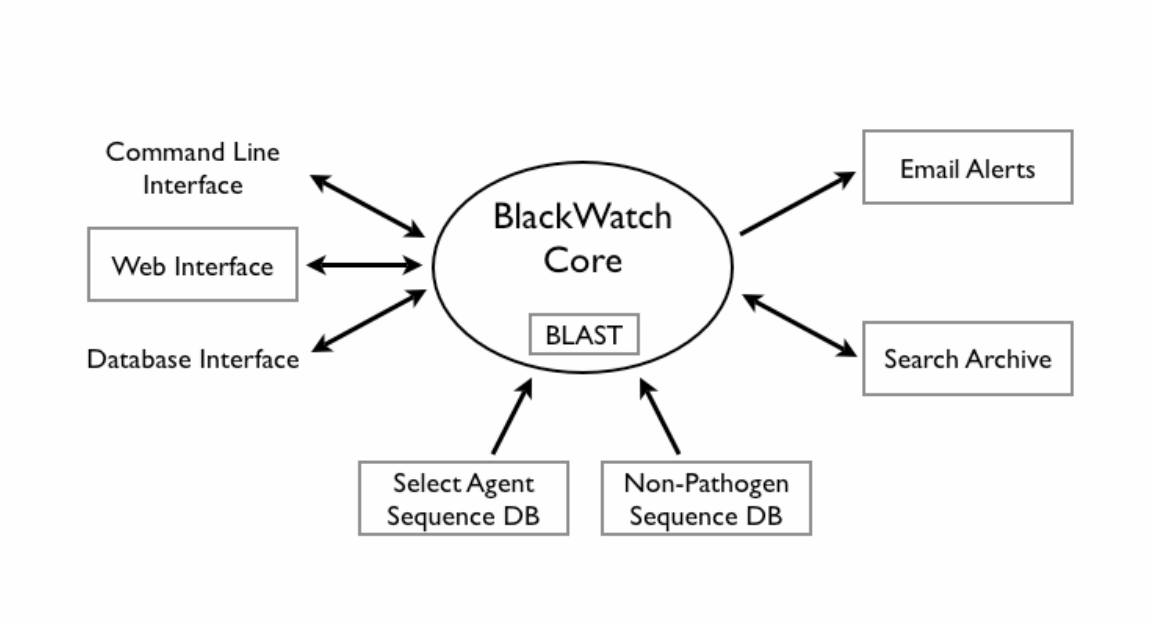

Figure 1: Structure of the BlackWatch Software.

Input sequences are passed to the core scripts from one of the interfaces. An assessment is made on the basis of length as to whether each batch contains short oligos or longer sequences. BLAST searches are initiated against the select agent sequence database. The system currently runs a *blastn* search of input nucleotide sequences against the nucleotide database and a *blastx* search of translated nucleotide sequences against a parallel database of protein sequences from the same pathogens. *tblastx* searches of translated nucleotide sequences against the translated nucleotide database will be introduced in the next version of the software.

BLAST results are processed and matches are assessed based on three criteria – absolute score, statistical significance (E-value), and the coverage of the matching segment. Coverage indicates how much of the query sequence is involved in the match. For an oligonucleotide one would expect the entire query sequence to be included in the alignment, whereas perhaps only part of a larger sequence would be involved. A combination of these criteria is used to select positive matches, with different cutoffs used with oligos relative to long sequences.

Search results for sequences that do not match are discarded, along with the sequences themselves. This is an important component in protecting proprietary information from customers. Positive search results against the select agent database are then searched against the non-pathogen database to see if they also match there. This is to help resolve false positives and is discussed below. Data are archived for each positive result. These include the input sequence, the raw BLAST output and associated information such as the customer identifier, date and time.

The system can be interfaced with relational databases. This will allow it to be driven by production databases at synthesis companies. In the absence of any common architecture for these databases, a custom interface script will have to be written for each company that chooses to set this up. I have successfully integrated the system with an Oracle database during beta testing at a leading oligo synthesis company.

Positive matches can be reported to relevant staff by way of email alerts. These include links to the web interface that will bring up the details of the match.

The search archives can be accessed by customer ID, allowing the history of sequence submissions to be reviewed. This will be important if a customer submits multiple related or overlapping sequences over a period of time. Comments can be added to each match and these are stored in the archive alongside the BLAST output. So one might record why a single match was assessed as a false positive. Later review in the context of other matches might lead you to change that assessment.

Below are some screenshots from the web interface. The first shows the query sequence input screen that will load a FASTA format file or accept sequences that are cut and pasted into the form.

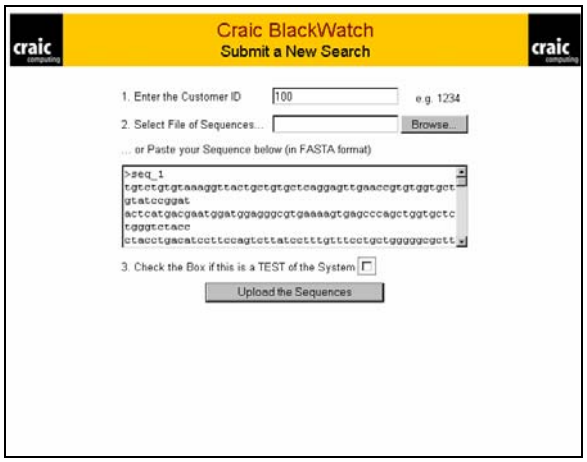

Figure 2: Sequence Input Web Page

Most searches will not produce matches and these are simply acknowledged as having been run. Positive matches are highlighted with links to the GenBank sequence that was hit, the raw BLAST output, the query sequence, etc.

Figure 3: Example of a Positive Match

The BLAST output is available for positive matches, allowing an expert to evaluate the quality of the alignment and thereby assess the likelihood of this being a true or false positive.

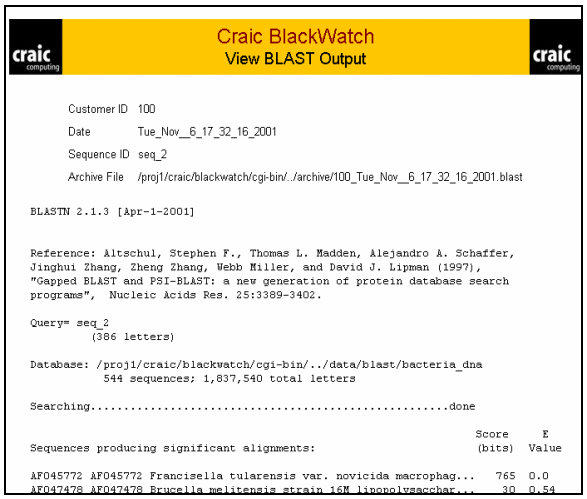

Figure 4: An Example of Detailed BLAST Output for a Positive Match

You can access a demonstration version of BlackWatch at https://biotech.craic.com/blackwatch.

3. The Custom Sequence Database

The database of sequences from select agents is a critical component of the BlackWatch system. Its composition directly influences the numbers of false positive and negative matches, as well as the performance of the search process.

Only sequences from defined select agents are included in the database. A critical issue in sequence screening is the potential disclosure of information about customer sequences. By limiting the database to only select agent sequences, the system minimizes this risk. So oligos related to a human gene would not be expected to match anything in the database. Sequences of bacterial origin, for example, have a much higher risk of matching, especially in light of the approach to false positive control. Understanding the probability of finding such matches will be important in the development and adoption of this system.

There are two approaches to building the sequence database. The first is to limit the sequences to those of genes known to be involved in virulence, toxin synthesis, etc. This highly focused approach would produce a small database with a low probability of false

positive matches. But this approach has several problems. Firstly it requires considerable effort up front in deciding what genes should be included and in extracting only the relevant sequences from GenBank. Secondly it ignores the possibility that genes other than this subset might be employed in the modification of a pathogen.

The alternative approach, which is used in BlackWatch, is to include all sequences that have been assigned to any organism on the select agent list. It is relatively straightforward to extract sequences based on the organism tag in a GenBank record and this selection can be fully automated using simple Perl scripts. Minimal up front effort is required and the data can be made available for searching immediately. It also ensures that all the available data is used in searching, with no preconceptions about how sequences might be used.

The drawbacks of the approach include the potential for redundant data being included in the database, slowing down searches and perhaps creating ambiguity. Some basic checks for redundancy are currently used in the preparation of the database but these could be improved. Perhaps the major problem is that the approach will include sequences of housekeeping genes, such as those for ribosomal proteins, which are highly conserved between diverse species. This raises the probability of false positives significantly.

4. Composition of the Database

All sequences in the database are extracted from the public GenBank database, hosted by the NIH (http://www.ncbi.nlm.nih.gov/Genbank/). This contains sequences for most if not all of the select agents, with complete genomes available for many of the organisms. Anyone attempting to engineer a pathogen using synthetic DNA would be expected to use this same database. No classified or proprietary sequences are included. Not only are these not available to me, but their inclusion would greatly complicate the software and its intended distribution to DNA synthesis companies.

The list of organisms for which all available sequences have been extracted is a composite of those included in the CDC select agent rule (42CFR73), the USDA regulations (7CFR331 and 9CFR121) and the Dept of Commerce Export Administration

"Commerce Control List" (CCL). The composite list specifies a total of 75 organisms and 22 toxins. The breakdown of these is shown in Table 1.

Host	**Human/Animal**	**Animal Only**	**Plant**	**Total**
Pathogen Type				
Viruses	19	12	2	33
Bacteria	15	3	8	26
Fungi	2	0	2	11
Rickettsiae	4	0	9	4
Prions	0	1	0	1
Toxins	22	0	0	22

Table 1: Pathogens in Composite Select Agent List

The list is included as an appendix to this paper and is also available online at: http://biotech.craic.com/blackwatch/regulations/List_of_Select_Agents.pdf

Toxins pose a problem for sequence screening. Protein toxins like abrin, ricin and conotoxin are gene products and so DNA and protein sequences for the toxins themselves are available. In the case of mycotoxins, such as aflatoxin, the molecule is not a protein. In these cases the sequences of genes that encode the biosynthetic pathway may be appropriate targets for sequence screening. This component of the database needs further study.

It might be advisable to include antibiotic resistance genes in the database as an obvious scenario that we need to consider is that of someone introducing antibiotic resistance into an existing bacterial pathogen. Unfortunately the widespread use of these genes in conventional molecular biology would ensure a very large number of false positive matches. This issue should be revisited once progress has been made dealing with the general problem of false positives.

5. Current Implementation of the BlackWatch Software

The software is written in Perl and runs on Linux systems. Porting the scripts to other UNIX variants and Mac OS X would be trivial and a port to the Windows operating system should be straightforward.

The system is in operation on my web server and has been in production use at Blue Heron Biotechnology in Bothell, WA, where it is used to screen requests for entire gene synthesis. It has also been beta tested for a limited period at a leading oligo synthesis company. They chose not to continue using the system for business reasons.

In order for the software to meet the sequence screening needs of the gene and oligo synthesis industry in general, it will require some additional development work. Performance needs to be improved to handle the throughput at large oligo synthesis sites. Integrating the system with existing relational databases that manage orders at these companies needs to be made easier. Most importantly the rate of false positive matches needs to be studied and minimized.

6. False Positives

The primary challenge facing sequence screening is to minimize the number of false positive matches. Every match reported by the system needs to be evaluated at some point by an expert. Those that are deemed to be real may trigger the involvement of the regulatory authorities. Every false positive that passes initial scrutiny will waste considerable time and devalue the importance of the approach in the eyes of those authorities.

Fine tuning the cutoff values for BLAST score, significance and coverage may help reduce false positives in general but will do nothing to address matches to housekeeping genes, etc. The approach that I am experimenting with at the moment is to use a second sequence database of non-pathogens. Any query sequence that hits the pathogen database is then searched against the non-pathogen, or 'reference', database and the corresponding matches, if any are presented to the user alongside the pathogen hits.

Currently the reference database is limited to bacteria and contains the genome sequences for *E.coli* and *B.subtilis*. This screenshot shows the results from a search with a ribosomal protein gene from *S.typhimurium*.

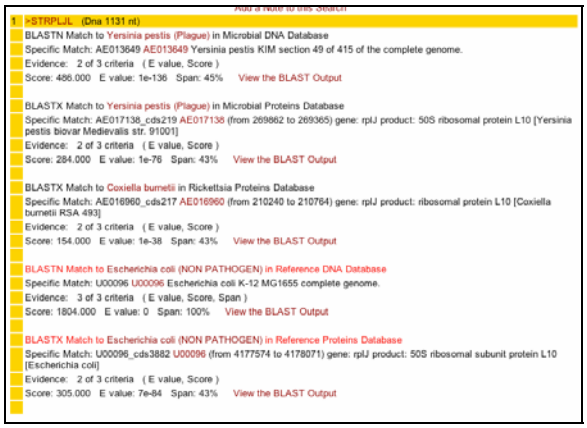

Figure 5: Example of a False Positive Match

This conserved gene has produced a match to the equivalent genes in *Y.pestis* and *Coxiella burnetii* in the pathogen database and also to *E.coli* in the reference database. By comparing the relative scores and significance, a reviewer would judge the query sequence as being more similar to the non-pathogen than to either of the pathogens. Hence this is probably a false positive.

The approach appears quite promising but work needs to be done in creating a comprehensive set of related non-pathogen sequences for viruses, etc., and in automating the process of calling false positives. No approach will catch false positives with 100% accuracy and so an expert reviewer will continue to be required. Perhaps the best that can be achieved is to add weight the scoring of matches according to the biological significance of the matching sequence. A very strong match to a sequence involved in anthrax toxin would be a clear positive match. A match to a less important region of the *B.anthracis* genome would be weighted down. This argues for a sequence database that combines the approach I currently use of capturing all sequences from the pathogens with some degree of expert curation that can define which genes are of particular concern.

False positives are inevitably more likely in the case of oligo sequences because of the sequence length. Here there is the opportunity to do some simulations and real world tests to quantify the problem.

7. Future Developments

There are many scenarios whereby someone who wished to synthesize or modify a pathogen could use the services of synthesis companies and still evade detection by BlackWatch. Minor variation in sequences, such as third position variation in codons, can already be caught by the *blastx* searches against protein sequences. Other scenarios include sending orders for overlapping oligos to different vendors or spreading out orders over a period of time so as to avoid revealing the intent behind a project. One way to address this would be to scan the archived searches across customers, or even across synthesis companies, looking for orders that might be related.

This would require that the results of screening from all vendors be submitted to a central location where these correlations could be made. I return to this idea at the end of the paper. The technical challenges of making these connections are very interesting, but they go hand in hand with a number of important business and confidentiality concerns.

The BLAST sequence comparison software is the obvious choice for comparing relatively large sequences against the database but for oligo comparisons it may be faster to use another approach such as a sequence word lookup table or a suffix tree algorithm. Computational speed could become a problem in high throughput oligo synthesis facilities. The figure of up to 25,000 oligos synthesized per day that Integrated DNA Technologies quotes is sobering. This means that a complete evaluation of each oligo must take place in less than 4 seconds. This can be achieved through a combination of adequate hardware and good software engineering but the system is not currently capable of this throughput.

8. Practical Deployment of Sequence Screening

Beyond the purely technical challenges of the BlackWatch package, its performance and the issue of false positives, there are several broader challenges to its practical deployment in the DNA synthesis industry that need to be overcome.

We need to make it very easy for a synthesis company to obtain, install and operate the software package. The barrier to its procurement can be reduced by making the software available free of charge. Appropriate software engineering can ensure that it is easy to set up and run. External funding from NIH or another agency will be necessary to support the development and deployment of the software. It is unlikely that the synthesis companies would fund the effort themselves.

We need to minimize the cost to the synthesis companies of evaluating the reports that sequence screening will yield. This is the time and effort that staff have to devote to looking at, acting upon, the putative positive matches. Some of these companies, most notably the oligo vendors, operate on very thin profit margins. Any added expense will be most unwelcome, especially if it requires effort on the part of skilled scientists.

But beyond these operational issues there are two major challenges that stand in the way of broad deployment—how to assess the validated, significant matches that do emerge from the screening and what to action to take based on that information. Neither role belongs with the DNA synthesis companies. They require expert knowledge and access to specific staff within the regulatory authorities.

Conclusion

A significant fraction of the synthetic DNA currently being produced today could be monitored by sequence screening at the major oligo and gene synthesis companies. For legitimate customers this process should pose no significant threat to their intellectual property.

For a group wanting to engineer a biological weapon, however, screening could serve as a serious deterrent. They would be faced with the choice of potential discovery by the screening software or having to bring the work in-house and significantly increase the level of effort and expertise needed to accomplish their goal.

Sequence screening has its limitations, as do most technologies that attempt to monitor threats, but I believe it should play an important role in the development of synthetic biology.