



# Assessment Criteria for Select Agent Sequences

Robert Jones

This paper discusses selection criteria for gene and genome sequences from Select Agent pathogens for use in screening orders at DNA synthesis companies. It also describes a set of related criteria for use in assessing what action to take in response to a positive screening match. These are based on realistic scenarios of how pathogens might be engineered and the relative difficulty of these, compared to the direct use of an infectious agent as a weapon.

## Introduction

Among the many pathogens of humans, animals and plants that are found in nature, there are some that pose especially high risks to society, Smallpox, Ebola virus or Anthrax for example. The possession of these Select Agents is controlled by a number of national and international regulations, with a primary focus on live infectious agents or toxins.

The possession of DNA from one of these organisms is a very different matter. Segments of DNA from even the most dangerous pathogen may well pose no immediate threat to human health. But these are of great concern because of how they might be used by someone intent on causing harm, to engineer or even recreate an infectious agent from its sequence alone.

While most of the potential scenarios that use genetic engineering involve far more effort, expertise and uncertainty than the direct use of a pathogen, we have to consider these as potential threats to biosecurity.

The emerging field of synthetic biology and the established technologies of genetic manipulation contribute greatly to medicine, agriculture and society. Commercial DNA synthesis services play a special role in this by providing custom research materials at low cost and fast turnaround time. But this speed and efficiency might be exploited by someone wishing to engineer a pathogen for use as a biological weapon.

By screening synthesis orders against a database of pathogen sequences we have the potential to detect any attempt to do this and to take action that would stop the work. The basic idea behind sequence screening is straightforward, but there are three major issues that must be addressed before approach can be widely deployed.

- Conservation of sequences between select agents and non-pathogens results in false positive matches.

- The selection of pathogen genes and genomes for inclusion in the target database for screening is subjective and requires domain expertise.
- No standard protocol exists for assessing the importance of any positive match and deciding what follow-up action to take.

Merely screening against all sequences from all the organisms in the Select Agent rules is too simplistic an approach. Large portions of the bacterial pathogen genomes are highly conserved in related non-pathogens and result in many false positive matches. With viral pathogens, on the other hand, a realistic threat may only come from someone working with an entire genome. We need to select an appropriate set of sequences from each organism based on credible scenarios in which each might be engineered to cause harm.

The current select agent regulations form the starting point for this effort but we need a way to assess the risk posed by individual genes and genomes among those pathogens. Here I propose some criteria that we might use in this process. These are based on realistic scenarios in which pathogens might be engineered.

In parallel to this, synthesis companies, regulatory and law enforcement agencies need a way to rank the importance of any given match to the pathogen database so as to direct their response. I propose a similar set of criteria that might be used for this purpose.

These are intended as a basis for discussion and the examples I provide of their application deserve review by those with particular expertise in specific pathogens. They should be familiar to anyone who has thought about the problem, but describing them here may be helpful.

## **Criteria for Selecting Pathogen Sequences for the Screening Database**

Scenarios in which pathogens might be engineered using synthetic DNA can be arranged into a few sets:

1. Recreate an entire infectious organism from a synthetic genome
2. Engineer a live pathogen to extend its host range, increase its virulence or to add antibiotic resistance
3. Engineer a non-pathogen to make it pathogenic by adding genes involved in virulence or toxin production
4. Engineer the production of large amounts of a toxin in an organism with the goal of purifying that toxin

The scenario in which a totally novel organism is designed and constructed is not realistic with today's technology, but it should be kept in mind as we consider the future.

The difficulty involved in each scenario varies greatly depending on the organism involved. Our knowledge of the biology of each pathogen varies, there may or may not be efficient ways in which to manipulate them and the technology required to recreate the entire organism may not yet be available. Although in the last of these, we should assume that the technology will be forthcoming.

Here are some criteria that we might use for deciding which sequences should be screened against. I have posed these as a series of questions. They are not intended to be complete; rather they may serve as a starting point for further discussion in the community.

- Can the live pathogen be obtained in the wild without undue effort?
- Is it possible to recreate the infectious agent from synthetic DNA?
- Are there specific genes that play a critical role in pathogenicity? If so, can these be transferred to similar, non-pathogenic species?
- Is the species or genus amenable to genetic manipulation while retaining pathogenicity?
- Are there obvious ways to transfer pathogenicity to a related non-pathogen or would these require substantial research?
- Does the pathogenicity of the organism involve the production of a toxin? Can the genes involved in toxin production be transferred to another organism?
- Are there sequences in the genome of the organism that are used for legitimate genetic manipulation?

The following are examples of how these might guide our selection of sequences in different species. I've broken down the discussion by the type of organism.

## **Viruses**

*Can the live pathogen be obtained in the wild without undue effort?*

Smallpox is unique in that it has been eradicated in the wild and the only known stocks are in high security labs. Therefore it is effectively unavailable leaving synthesis of the entire virus, using publicly available sequence information, as the only route to obtaining live virus.

In contrast, Yellow Fever virus is widespread in several parts of the world. Obtaining this in the wild might be much easier than attempting the synthetic route. Viruses in this class

might be treated differently in the assessment phase of screening and there may be some examples that we might want to drop them from the screening database entirely.

*Is it possible to recreate the infectious agent from synthetic DNA (or RNA)?*

The recreation of infectious particles of Poliovirus and the 1918 strain of Influenza virus from totally synthetic DNA show that at least some viruses are amenable to this. Single stranded (positive strand) RNA viruses, such as poliovirus and the flavivirus group, are able to 'boot up' by mimicking mRNA in the host cell. Some double stranded DNA viruses, such as the Herpesvirus group, utilize host cell polymerases for replication. It may be possible to create infectious particles from genomes in both of these groups.

Other groups of viruses, such as smallpox, do not use host polymerases and these are viewed as posing much greater technical challenges. However we cannot and should not presume that these are insurmountable.

That would direct us to include *all* viral genomes in the screening database.

*Are there specific genes that play a critical role in pathogenicity?*

*Is the species or genus amenable to genetic manipulation while retaining pathogenicity?*

The small genomes of viruses, and the constraints imposed by secondary structure and packaging of those in the viral capsid, may mean that it is very difficult to engineer a benign species into a pathogen. Certainly such an attempt would require considerable research. Human viruses like Influenza and HIV show high sequence variability but relatively little variation in the function of their genes and their genomic organization.

On the other hand, bacteriophages have proven quite amenable to having genes inserted into their genomes and this ability is used widely in genetic manipulation.

If we can assess how viruses might be engineered we might be able to tailor sequence screening algorithms to detect these.

*Are there sequences in the genome of the organism that are used for legitimate genetic manipulation?*

Clearly a lot of legitimate work goes on in the production of vaccines, diagnostics and therapeutics against these viruses. But in addition certain sequences have been used in a much broader range of applications. One example of this would be the use of gene promoters from poxviruses to drive transcription of arbitrary genes in mammalian tissue culture.

Sequences like these will give rise to many false positives. By excluding or somehow 'tagging' these we can better assess any matches to them. It should be straightforward to assemble a list of these.

## **Bacteria and Fungi**

*Can the live pathogen be obtained in the wild without undue effort?*

Most of the bacterial select agents can be found in the field in some parts of the world, although the difficulty of isolating a sample will vary greatly between pathogens. *Bacillus anthracis*, for example, is rare in the developed world but continues to infect ruminants in a number of countries.

Isolates of all the agents are studied in research labs around the world and, although the US and some other countries carefully control access to these stocks, there is the potential for a lab in a country with less regulation to be the source for a live agent.

*Is it possible to recreate the infectious agent from synthetic DNA?*

To date (early 2008) no bacteria has been completely recreated from synthetic DNA but the work of Venter, Smith and colleagues with *Mycobacterium genitalium* may well provide the first successful example of this. Assuming this becomes possible, it will likely remain a substantial undertaking, such that isolation of a bacterial pathogen from the field will continue to be the easiest way to obtain the live agent.

This may direct us to exclude sequences from certain agents where it is felt there is no reasonable scenario in which their genome might be synthesized. This is clearly a trade off between reducing the false positive rate at the cost of reducing the overall coverage of the database.

*Are there specific genes that play a critical role in pathogenicity?*

*Is the species or genus amenable to genetic manipulation while retaining pathogenicity?*

In contrast to complete genome synthesis, transfer of specific genes from a pathogen into a related non-pathogen would be relatively easy. Conventional genetic manipulation relies on techniques for gene transfer among bacteria and bacterial evolution has been driven in large part by lateral gene transfer between species. As a result it is quite easy to think of ways in which non-pathogens might be engineered to express a toxin from Anthrax or Cholera for example.

Despite the relative ease of these manipulations, it is unclear how successful an engineered pathogen might be if dispersed into the population. Conventional wisdom would say that pathogens have evolved into a form that is near optimal for their environment and that any significant change to their genome is likely to be deleterious.

Genes that play a role in virulence pose a less clear risk. Our understanding of virulence is relatively limited and varies between organisms. Pathogenicity in *F. tularensis*, for example, is poorly understood at present. As research uncovers these mechanisms we may be able to identify specific genes that should be screened for. But for now this

ignorance means that an attempt to engineer a bioweapon using virulence genes would require considerable effort and experiment. That makes these unattractive as a target.

Engineering resistance to antibiotics in an otherwise vulnerable pathogen is a scenario of particular concern. Drug resistance genes are known for most antibiotics and the manipulation and transfer of these genes is a central tool in legitimate molecular biology. The ubiquity of these genes makes it effectively impossible to assess their relevance if there were found in a DNA synthesis order. Furthermore, insertion of such genes into a pathogen genome is not in any way restricted to specific regions of that genome. This makes detection of this scenario extremely difficult.

The implications for sequence screening are several. It should be straightforward to identify specific genes for toxins and well-defined mechanisms of virulence and include those in the screening database. Expert knowledge and more research will be required to identify virulence genes in general. Antibiotic resistance genes, although important, are simply used too frequently in legitimate research to make screening against them practical.

*Does the pathogenicity of the organism involve the production of a toxin?  
Can the genes involved in toxin production be transferred to another organism?*

*B. anthracis, C. botulinum, E. coli O157:H7 and V. cholerae* are all examples where a toxic bacterial protein is responsible for cell destruction in the pathogen's host. In many cases, bacterial toxin genes have already been transferred to other bacterial strains for research purposes and it should be assumed that all genes encoding protein toxins can be transferred in this way.

The inference is that all toxin genes from bacterial select agents should be screened against.

*Are there sequences in the genome of the organism that are used for legitimate genetic manipulation?*

There are many of examples of this. Cholera toxin has been used to target specific cell types for killing by attaching it to other proteins that bind specifically to those cells. These are clear examples of the 'dual use' of a gene. While sequence screening can detect any instance of these in a synthesis order, it cannot infer the intent of the person placing that order.

These will certainly give rise to false positives but that may not be a bad thing as legitimate constructs containing these genes can pose direct biosafety risks to the staff at gene synthesis companies that work with them. Detecting these through screening may be a useful safeguard for companies.

## Toxins

The bacterial toxins have already been discussed, but the non-bacterial toxins require special consideration from the perspective of DNA synthesis. Some of these are proteins or peptides, such as ricin, abrin and the conotoxins, whereas others are complex organic molecules, such as Tetrodotoxin or Saxitoxin.

With these, only the toxins themselves are select agents, *not* the organisms that produce them. Scenarios for their use involve the production and direct dispersal of the toxin, as opposed to an infectious agent that can replicate in a population.

The protein and peptide toxins are encoded directly by genes and these might be engineered or transferred to other organisms in order to make large amounts of the material. These specific genes can be easily included in a screening database, but there is no reason to include other genes from these species.

The lower molecular weight organic toxins are a different matter. No single gene encodes the toxin, rather a biosynthetic pathway, with enzymes encoded by multiple genes, is responsible for their production. But these pathways most likely share steps with the other pathways for innocuous molecules. Identifying toxin-specific steps might be possible but this would require expertise in the specific field.

*Can a toxin be obtained in the field without great effort?  
Can it be obtained by direct chemical/peptide synthesis?*

Ricin, for example, appears to be easily produced in relatively large quantities from Castor beans. There would seem to be little or no reason for anyone to engineer its synthesis in another organism.

The low molecular weight toxins appear to be amenable to direct chemical synthesis, albeit requiring significant expertise. Whether or not this route would be easier to follow than engineering production in an organism is unclear.

*How complex, and well understood, is the natural synthetic pathway for low molecular weight toxins?*

Attempting to recreate and engineer a complex pathway in another organism is not impossible. The current attempts to produce an anti-malarial compound in *E. coli* by the Keasling group in Berkeley is a striking example of this. But the effort required to do this might effectively make this scenario impractical for producing a bioweapon. With that in mind, we might not want to consider screening genes involved in these pathways.

*Are there other genes from non-select agents that we should be concerned about?*

Thus far we have used the Select Agent lists as our starting point. But these are entirely based on a ranking of infectious agents and toxins. There may be specific genes from other species that could pose a risk if engineered or transferred. Although no scenarios like this are obvious to me, we should keep an open mind about these as new technologies continue to develop.

## **Criteria for Ranking the Importance of Positive Matches from Screening**

There are two components to sequence screening. The first is the comparison of a query against the database of select agent sequences and the determination of significant matches, where *significance* involves statistics and an understanding of sequence conservation.

The second is the assessment of significant, or important, positive matches in terms of the action that the synthesis vendor should take in response. Here *significance* involves the potential risk associated with the sequence, the likelihood that it is being used for legitimate research and operational factors such as the effort required from company and regulatory staff in order to follow up on the match.

The false positive rate with screening is a problem in the first component that currently has a large impact in the second. That is outside the scope of this paper and will be discussed elsewhere. Here I want to assume that the false positive issue is manageable and instead focus on *operational* criteria for deciding the action to be taken when a positive match is encountered.

Clearly not all matches warrant the same response. An attempt to synthesize the smallpox genome is far more serious than an order for a single gene from a plant pathogen. It warrants a very different response, perhaps involving different agencies, and given limited resources in this area, the second case might not warrant any response at all.

We need to formulate guidelines that can determine the appropriate response to specific matches. These will vary according to the pathogen and to the company or agency that is reviewing the match. The FBI, for example, might want to be informed of all matches to certain species but might only act on a small subset of those.

I want to propose some basic criteria that might help in this debate. Again, I'll present them as series of questions and then illustrate how they might be applied.

- Is the match to a human pathogen?
- Is the agent a match to an agricultural pathogen?

- How immediate is the threat posed by this match? Are some scenarios for engineering pathogens faster to implement than others?
- Does a given scenario require extensive research and testing or does it use established technology?
- Is this match an example of legitimate direct research on the pathogen? Are there legitimate uses for sequences from this pathogen, outside of direct research?
- Can we establish 'white lists' of labs that work with specific pathogens and which are deemed to pose no threat?

*Is the match to a human pathogen?*

*Is the agent a match to an agricultural pathogen?*

In any ranking of biological threats, those to humans will be higher than to animals or plants. The type of threat posed by a given match may trigger different responses from different agencies. A match to a human threat may trigger action from the CDC and FBI, whereas a plant pathogen may only trigger a response from the USDA. Each agency may draw up its own list of agents in terms of response, and different agencies may take a lead role in specific responses.

*How immediate is the threat posed by this match?*

*Does a given scenario require extensive research and testing or does it use established technology?*

Different scenarios will vary greatly in the time and effort they require. Transfer of an antibiotic resistance gene to a bacterial pathogen might be straightforward whereas extending the host range of a pathogen would be a major research project.

It can be easy for our community to envisage complex scenarios where a pathogen might be engineered to cause greater harm. But these typically require substantial expertise, funding and time, and they carry a much higher risk of discovery as more reagents and equipment will be required. With easier ways to cause harm available to someone with that goal, these scenarios are less likely to occur and perhaps they should be given a lower significance.

*Is this match an example of legitimate direct research on the pathogen?*

*Are there legitimate uses for sequences from this pathogen, outside of direct research?*

*Can we establish 'white lists' of labs that work with specific pathogens and which are deemed to pose no threat?*

This is equivalent to the false positive problem of the screening software. Here we have a *true* positive match to a pathogen sequence but it may or may not be intended for a legitimate purpose.

Hopefully, we can assume that the vast majority of positive matches from screening will represent legitimate research. We don't want to burden companies and regulatory agencies with having to follow up with the researchers in every one of these cases, and yet we don't want to miss a single request that is not legitimate.

One approach is to identify labs that do legitimate research, such as those registered with the CDC under the select agent regulations, and exempt orders from these labs from follow up action. Similarly the Venter Institute/MIT/CSIS study has proposed ways to pre-approve labs using biological safety officers at universities, etc. who would assess and vouch for the legitimacy of these labs.

While these could greatly reduce the amount of follow up work, one has to assume that someone attempting to engineer a pathogen is likely to have received training in exactly this sort of lab. The source of the Anthrax mailings in 2001 is likely to have been someone from, or with access to, a legitimate lab working on that strain on the bacterium.

This is a dilemma for any work involving pathogens and is not unique to synthetic biology.

## **Conclusions**

Clearly the issues involved in selecting sequences for screening and deciding what actions to take when matches are found are not trivial.

As with all aspects of security, one has to strive to minimize a threat and yet recognize that operational factors, such as budget, expertise and available time, will conspire to produce an imperfect solution. Deciding what level of security is 'good enough' is a difficult challenge that requires much debate.

In our domain, the use of sequence screening is generally viewed as an approach that can be widely deployed with a relatively low direct cost. But the problem of assessing false positive matches from screening and 'true positives' from legitimate researchers needs to be solved.

Progress is being made in the development of screening algorithms. Hopefully this paper will contribute to the discussion regarding the sequences to screen against and the action to be taken on finding a positive match.

I would welcome comments from readers of this paper.

Robert Jones ([jones@craic.com](mailto:jones@craic.com))  
Craic Computing LLC, 911 East Pike St #231, Seattle, WA 98122

This report is published under the terms of the Creative Commons Attribution-No Derivative Works 3.0 United States License (<http://creativecommons.org/licenses/by-nd/3.0/us/>)