Strengths and limitations of the federal guidance on synthetic DNA

To the Editor:
The December issue included a report summarizing the first reactions of the gene synthesis industry to the publication of the US government *Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA*[1]. Some of the questions raised by the federal guidance had already been exposed in your columns[2,3], but none of these previous comments relied on a bioinformatics analysis of the screening protocol proposed by the US government. Here we present the preliminary results of an implementation of this protocol with the hope of documenting the strengths and limitations of the federal guidance.

This document outlines a minimal DNA sequence screening protocol that providers of gene synthesis[4] services are encouraged to follow before fulfilling an order. The objective of the protocol is to identify sequences of concern of any length that are specific to 'select agents or toxins' (SAT) listed on the National Select Agent Registry (http://www.selectagents.gov/). It starts by translating the nucleotide sequence ordered by the customers into each of six possible reading frames. Both the nucleotide and amino acid sequences must then be divided into fragments that are individually aligned against GenBank using a local sequence alignment algorithm. Alignment results are interpreted using the 'best match' criterion, a procedure designed to identify sequences specific to SATs without relying on a curated database of sequences of concern.

Although the federal guidance gives a general method for the automatic identification of potentially dangerous sequences, few instructions are given concerning the exact implementation of the method. Here we describe an interpretation of the method that is amenable to implementation in software (**Fig. 1**). The input DNA sequence to be screened first undergoes a six-frame translation. The resulting six-amino-acid sequences and the two original DNA sequences corresponding

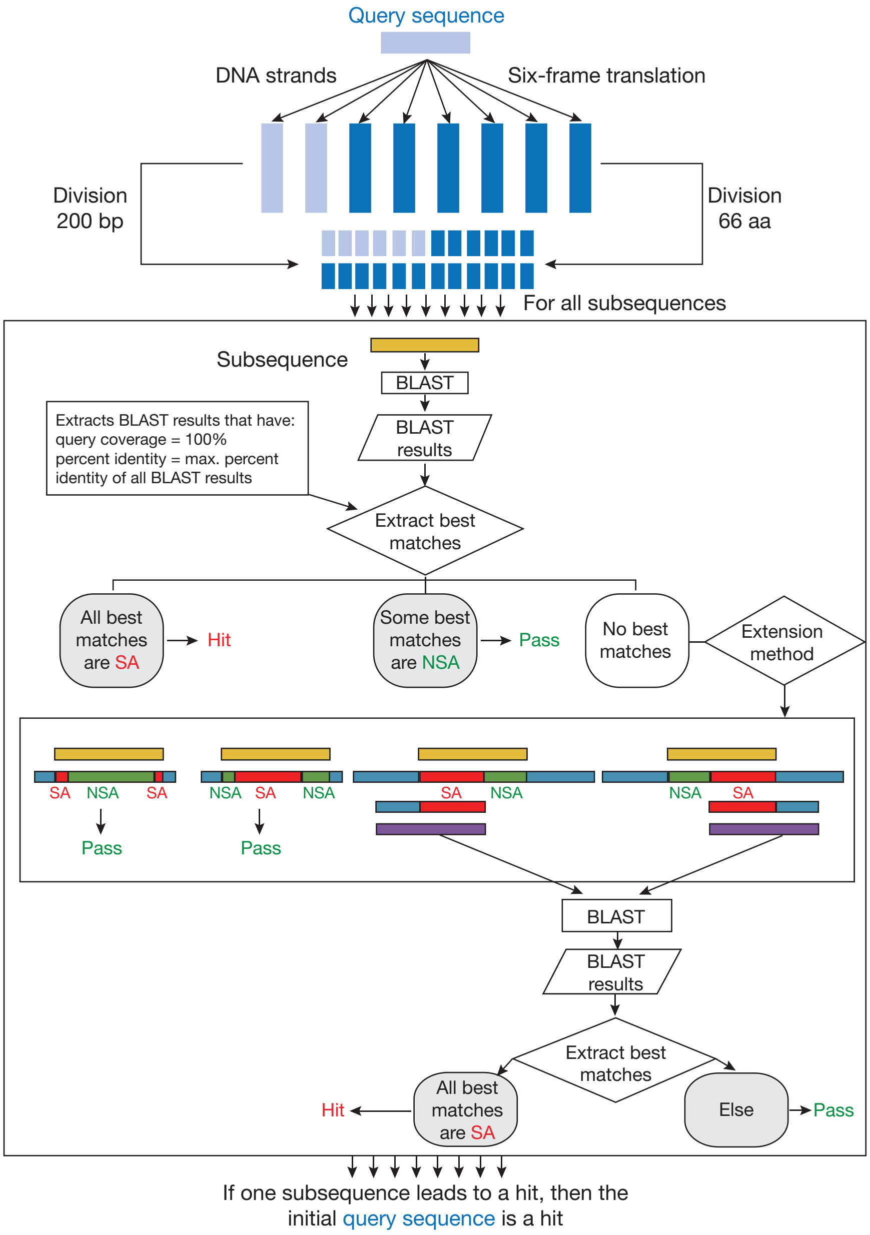

Figure 1 Sequence screening algorithm. The query sequence first undergoes a six-frame translation, then the amino acid sequences and nucleotide sequences are fragmented into the appropriate size. The subsequences are then aligned using BLAST against GenBank and the nature of the best matches is determined. If there is no best match but there are sequences of concern with query coverage >50%, then the alignment extension occurs. The algorithm is repeated on the extended sequences to determine whether original query sequence is a hit to a SAT.

to the two strands of the query sequence are then divided into 66 amino acids (aa) and 200-bp fragments, respectively. When the sequence length is not a multiple of 200 bp or 66 aa, a new subsequence is created using the last 200 bp or 66 aa of the sequence. This subsequence overlaps the last subsequence resulting from the initial fragmentation, but it ensures that the entire sequence is screened.

All of these fragments are then analyzed individually to determine if they should be flagged. They are first aligned against GenBank using BLAST[5]. The best matches are extracted among the BLAST results by selecting the alignments with the highest percent identity over the entire 200-bp fragment (query coverage of 100%). To determine if a best match corresponds to a SAT, the information in the GenBank reference page is cross-referenced with a keyword list. For toxins, keywords include alternative names of the toxin, the names of enzymes that are associated with the production and function of the toxin, and the names of organisms that uniquely produce the toxin. For organisms and viruses, keywords include alternative species names, the names of diseases associated with the entries and any toxins or pathogenic agents uniquely produced by the entry. Two keyword lists were developed. The restricted keyword list has 86 records, whereas the extended keyword list has 340 keywords. If every best match is to a SAT, then the fragment is considered a hit.

A sequence can be fragmented such that a 200-bp region of SAT could unequally straddle two contiguous fragments. To alleviate this issue, the algorithm creates a new 200-bp (of 66 aa) fragment when it detects the presence of an alignment to a SAT longer than 100 bp or 33 aa on either extremity of the subsequence. This new subsequence is composed of the SAT region from the initial fragment and a region from the appropriate adjacent fragment of a length such that the sum of both regions is equal to 200 bp or 66 aa. Every new extended subsequence is compared with GenBank to identify its best matches, as previously described. This thorough analysis is fairly computationally expensive because screening a 1-kb sequence requires at least 40 sequence alignments (two DNA and six protein alignments for each 200-bp fragment). Sequences of several kilobases can be analyzed in a few minutes on a dedicated server or high-end workstation, which should be compatible with the operational constraints of the gene synthesis industry.

The draft guidance published in 2009 focused exclusively on sequences longer than 200 bp, but the final version has removed this exclusion. This decision is unfortunate. Screening short sequences creates all sorts of bioinformatics complications that can affect the quality of the results. The best-match method has been designed to screen long sequences and is not suitable for screening short sequences. Furthermore, by removing the 200-bp limit, the guidance is somewhat inconsistent. Short sequences are more likely to be ordered as oligonucleotides than double-stranded DNA, but screening oligonucleotide orders is outside the scope of the guidance. For all these reasons, we decided to keep the 200-bp restriction in our implementation of the guidance.

To evaluate the performance of this protocol, we developed a test suite of sequences annotated as either SAT (75 sequences) or non-SAT (100 sequences) after manually reviewing alignment results for each sequence. The accuracy of the screen can be estimated by comparing the screen output with the test sequence annotations. Not surprisingly, the performance of the screening protocol depends on the content of the keyword database. The number of false negatives, sequences of concern that are undetected, is minimized when using the extended keyword list (25 false negatives with the limited keyword list versus 1 false negative with the extended keyword list). Because the outcome of the screen is so dependent on the keywords used to analyze alignment results, it would be useful to develop a standardized list of keywords acceptable to all constituencies. Beyond its application in this particular context, these keyword lists are a prerequisite to the development of a sequence-based classification system of SATs[6].

Moreover, we screened the GenoCAD[7,8] parts database. This data set includes 1,258 sequences longer than 200 bp that mimic the order books of gene synthesis companies. The screen returned 32 hits (2.54%). For most hits, the human review did not uncover any significant relation to SATs beyond some local homology between one of many fragments and a SAT sequence. Even so, we found one GenoCAD part closely related to the YopH protein from *Yersinia pestis* (gi|14488772). This protocol is extremely effective at detecting sequences of concern embedded into larger sequences because each 200-bp fragment is analyzed individually. The six-frame translation also ensures that redesigned sequences which take advantage of the degeneracy of the genetic code are easily detected by the protocol. However, it proved difficult to design test sequences by introducing mutations in SAT sequences found in GenBank as there is no simple way to determine if such sequences should be detected or not as the biological activity of these sequences is unknown. It would therefore be useful to develop large and realistic training sets that could be used to assess the performance of software implementations of the guidelines recommended by the government.

Before the publication of the federal guidelines, the International Association–Synthetic Biology (IASB; Heidelberg, Germany) published “Code of conduct for best practices in gene synthesis” and the International Gene Synthesis Consortium (IGSC; San Francisco) released their “Harmonized screening protocol.” Several important differences between the protocols can be confusing to the public and the gene synthesis industry[3]. **Table 1** shows that the industry is advocating a global analysis of the sequence, leaving the responsibility of interpreting the results to a human operator. The federal protocol advocates a more granular approach that requires breaking down sequences into smaller fragments analyzed individually. This high-resolution screen can detect local features of a sequence that may be undetected if the sequence is analyzed globally in one pass. Since it is not practical to manually review the results of all the sequence alignments performed by the federal protocol, the federal document provides objective criteria to identify what should be further investigated. This automatic classification of sequences of concern is both a strength and a weakness. On the one hand, it makes it possible to objectively assess the performance of the screen, something that is not possible

Table 1 Comparison of sequence screening protocols

Recommendation	IASB	IGSC	US
Fragment double-stranded DNA sequence	No	No	200 bp
Screen six-frame translation of DNA sequence	No	Yes	Yes
Screen against curated sequence database	No	Yes	Optional
Defined criteria to identify sequence as a hit	No	No	Best match
Requires human element in screening procedure	Yes	Yes	No

when the results of sequence alignment are interpreted by human operators. On the other hand, the intrinsic limitations of the best-match method may overlook patterns that human operators would detect. Furthermore, determined individuals could take actions before placing an order to ensure that their order does not raise a red flag. In its defense, the government standard has always been described as a bare minimum that does not prevent the use of complementary approaches such as the ones proposed by the industry. In the long term, the security of gene synthesis may not lie as much in standards as in the availability of biosecurity software applications inspired by computer security solutions. Such biosecurity tools would rely on rapidly evolving models of biosecurity threats to provide human operators with the information they need to quickly and efficiently screen all synthetic DNA sequences at the different steps of the design and fabrication process. The wide adoption of such tools would be objective evidence that the community is developing a culture of responsibility, which is unanimously regarded as the best protection against this new biological threat[2,9].

ACKNOWLEDGMENTS
This work was supported by National Science Foundation Award no. 1060776 to Virginia Tech. L.A. was supported by a graduate fellowship from the Science Applications International Corporation (SAIC). M.K., A.S., O.M., G.L. and T.S. were supported by undergraduate research fellowships from the MITRE Corporation.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Laura Adam, Michael Kozar, Gaelle Letort, Olivier Mirat, Arunima Srivastava, Tyler Stewart, Mandy L Wilson & Jean Peccoud

Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA.
e-mail: peccoud@vt.edu

1. Eisenstein, M. *Nat. Biotechnol.* **28**, 1225–1226 (2010).
2. LaVan, D.A. & Marmon, L.M. *Nat. Biotechnol.* **28**, 1010–1012 (2010).
3. Fischer, M. & Maurer, S.M. *Nat. Biotechnol.* **28**, 20–22 (2010).
4. Czar, M.J., Anderson, J.C., Bader, J.S. & Peccoud, J. *Trends Biotechnol.* **27**, 63–72 (2009).
5. Camacho, C. *et al. BMC Bioinformatics* **10**, 421 (2009).
6. Wadman, M. *Nature* **466**, 678 (2010).
7. Czar, M.J., Cai, Y. & Peccoud, J. *Nucleic Acids Res.* **37**, W40–W47 (2009).
8. Cai, Y., Wilson, M.L. & Peccoud, J. *Nucleic Acids Res.* **38**, 2637–2644 (2010).
9. Bennett, G., Gilman, N., Stavrianakis, A. & Rabinow, P. *Nat. Biotechnol.* **27**, 1109–1111 (2009).

Partnering Brazilian biotech with the global pharmaceutical industry

To the Editor:
Previous descriptions of the Brazilian health biotech sector in this journal[1,2] have highlighted several challenges to sustainable development, including inefficient interactions between the public and private sectors[1], a lack of venture financing[1] and a paucity of legal incentives to encourage commercialization of the region's rich biodiversity[2]. Here we would like to emphasize the importance of another issue that prevents Brazilian biotech enterprises from successfully bringing innovative drugs to market—the lack of local partnerships between small and large companies and the poor level of collaboration between Brazilian companies and multinational pharmaceutical companies that can accelerate late-stage clinical development.

nature biotechnology
Targeted gene knockouts in zebrafish
Next-generation sequencing-by-hybridization
Mapping human protease action

One illustration of the behavior of the local health biotech sector is the lack of interaction between the two main industry associations in the country—the National Association of Pharmaceutical Laboratories (ALANAC; http://www.alanac.org.br) and the Brazilian Research-Based Pharmaceutical Manufacturers Association (Interfarma; http://www.interfarma.org.br). This weakens the Brazilian industry by preventing both collaboration and pooling of complementary scientific and financial resources that might otherwise bankroll innovative drug development. Most local companies are insufficiently capitalized to carry out innovative R&D activity in the area of biopharmaceuticals, let alone invest over a billion dollars to fund the core process from target discovery to a regulatory approval or registration. As a result of the weakness of the pharmaceutical sector, not one blockbuster drug has been developed in Brazil throughout its history. Moreover, many ALANAC member companies are opting to produce less R&D-intensive products, such as generics, instead of innovative drugs.

Against this background, the Brazilian government has implemented several initiatives to create a local environment that is more conducive to innovative product development, thereby enriching the pool of partnering opportunities for pharmaceutical companies.

In 2004, the 'Innovation Law' (Law 10,973)[1] was introduced to encourage the sharing of intellectual property and other resources between public and private entities and allow direct support of R&D activities in private enterprises. Although the number of Brazilian biomedical inventions licensed at the US Patent & Trademark Office (Washington, DC) has doubled over the past two decades, it is still only a small number (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/cst_utl.pdf). The situation in Brazil is complicated further by the country's cumbersome patenting process. Under Patent Law 9,279, the National Institute of Industrial Property can grant a pharmaceutical patent related to a product only after agreement has been obtained from Brazil's National Health Surveillance Agency. This rule makes the Brazilian process longer and more unwieldy than that in any other territory in the world.

Even so, progress in fostering an innovation- and enterprise-friendly environment is being made. Two laws for creating favorable fiscal incentives for R&D investment (the 'Asset Law'; Law 11,196) and income tax exemptions for enterprises involved in R&D (Law 11,487) were introduced in 2005 and 2007, respectively. Although these laws had only a minor impact initially, in 2008 the income tax deduction derived from Law 11,196 amounted to ~0.05% of Brazilian gross domestic product (http://www.mct.gov.br). Even greater benefits could potentially be accrued if Law 11,487 could be extended to private enterprises, rather than applied solely to public research institutions, as it does at present.

More recently, the launch of the Brazilian Technology System (SIBRATEC[3]) has facilitated the identification and development of promising compounds