THE NATIONAL ACADEMIES

DIVISION ON EARTH AND LIFE STUDIES

Sequence-Based Classification for Select Agents: A Brighter Line

Committee on Scientific Milestones for the Development of a Gene-Sequence-Based Classification System for Oversight of Select Agents

Board on Life Sciences • National Research Council

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine
National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

Study background

- Regulation of dangerous pathogens and biological toxins is currently based on the Select Agent list.
- The Select Agent Regulations require users to monitor and track the possession, use, and transfer of listed agents.
- Synthetic biology now allows Select Agents to be synthesized based on knowledge of their gene sequences, and may enable production of novel microbes.
- Other dangerous pathogens, such as those causing emerging infectious diseases, may not be on the list and therefore work with them is not regulated.
- Because such novel pathogens are not listed as Select Agents, they are not subject to control under the Select Agent Regulations.

Study background

In its report, "Addressing Biosecurity Concerns Related to the synthesis of Select Agents", the National Science Advisory Board for Biosecurity (NSABB), considered the impact of synthetic biology and DNA synthesis technology on biosecurity and the current Select Agent regulations (SAR). The principal concerns it addressed were that:

- DNA synthesis technology is rapidly diminishing barriers to acquisition of pathogens, because an increasing variety of organisms may be instantiated by whole genome synthesis, rather than by acquisition of samples of existing organism stocks or cultures;
- Natural variation and intentional genetic modification blur the boundaries around any discrete list based on taxonomic names; and
- Synthetic biology may enable the accidental or deliberate construction of chimeric or entirely novel pathogens unrelated to current ones.

Study background

- Based on NSABB recommendations, our committee was asked to:

"identify the scientific advances that would be necessary to permit serious consideration of developing and implementing an oversight system for Select Agents that is based on predicted features and properties encoded by nucleic acids rather than a relatively static list of specific agents and taxonomic definitions".

- Study requested through a contract with the National Institutes of Health - Office of Biotechnology Activities .

NATIONAL ACADEMIES

Study committee

- **JAMES W. LEDUC *(Chair),*** *The University of Texas Medical Branch at Galveston*
- **RALPH BARIC,** *University of North Carolina at Chapel Hill School of Public Health*
- **ROGER G. BREEZE,** *Centaur Science Group*
- **R. MARK BULLER,** *Saint Louis University School of Medicine*
- **SEAN R. EDDY,** *Janelia Farm, Howard Hughes Medical Institute*
- **STANLEY FALKOW,** *Stanford University School of Medicine*
- **RACHEL LEVINSON,** *Arizona State University*
- **JOHN MULLIGAN,** *Blue Heron Biotechnology*
- **ALISON O'BRIEN,** *Uniformed Services University of the Health Sciences*
- **FRANCISCO OCHOA-CORONA,** *Oklahoma State University*
- **JANE S. RICHARDSON,** *Duke University Medical Center*
- **MARGARET RILEY,** *University of Massachusetts*
- **TOM SLEZAK,** *Lawrence Livermore National Laboratory*

STAFF

- **INDIA HOOK-BARNARD *(Study Director)***
- **CARL-GUSTAV ANDERSON**

Study process

- **Committee held three meetings - May, Sept and Oct of 2009**
- **Sept meeting was held in conjunction with a workshop:**
  - Julia Kiehlbauch, Robbin Weyant, Claudia Mickelson, Edward You and Amy Patterson helped us **understand the current structure for oversight of Select Agents and pathogens.**
  - Peter Pesenti, John Mulligan, Marcus Graf, Claes Gustafsson and Stephen Maurer discussed the **current mechanisms and criteria for screening and surveillance at the sequence level**.
  - Stanley Falkow, Jeffrey Taubenberger, Michael Katze, Ralph Baric and Ramon Felciano discussed **virulence**.
  - Sean Eddy, Jonathan Eisen, Elliot Lefkowitz, John Moult and Ian Lipkin addressed **gaps, challenges and potential milestones for predicting pathogenicity from sequence information**.
- **Meetings also included discussions with Mary Groesch (NIH, OBA); David Relman (NSABB), Jacqueline Corrigan-Curay (RAC); James Blaine (CDC) Carol Linden (HHS), and Arturo Casadevall (NSABB)**

Study findings

What is the aim of the Select Agent Regulations?

Finding:

The Select Agent Regulations are intended to restrict access of *known* agents that pose a threat to *biosecurity.*

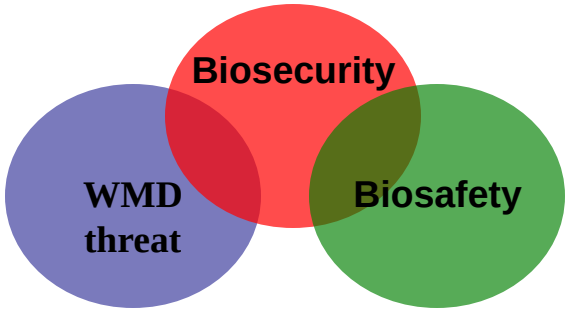
Study findings

What is the aim of the Select Agent Regulations?

Biosecurity: Select Agent Regulations aim to identify agents with a potential to be used as weapons; restrict access of Biological Select Agents and Toxins to certain individuals and facilities; and specify conditions under which research may occur.

SAR are NOT primarily concerned with:

- **Biosafety** – there is a well established framework of oversight for legitimate use of microorganisms - BMBL, RAC, BSL1-4
- **Biological Weapons threat** - BWC, and implementing legislation, prohibits development of biological agents as weapons.



What is the aim of the Select Agent Regulations?

Biosecurity: Select Agent Regulations aim to identify agents with a potential to be used as weapons; restrict access of Biological Select Agents and Toxins to certain individuals and facilities; and specify conditions under which research may occur.

- **SAR Requirements --** Handling of Select Agents requires controlled access, physical security, inventory control and site-specific risk assessments. Individuals with access to Select Agents must be cleared through FBI Criminal Justice Information Services Division background check.
- **SAR have serious consequences --** Select Agent Regulations are an instrument of law enforcement to facilitate attribution and prosecution. Failure to meet the requirements may result in criminal penalties of fines and up to ten years imprisonment.

What is the aim of the Select Agent Regulations?

focus on the *known*:

The Select Agent Regulations work primarily and most effectively in the context of possession and transfer of known stocks –providing a "chain of custody" – where names and Select Agent status are propagated in well-defined manner from registered sender to registered recipient of Select Agent cultures.

What is a Select Agent?

- A Select Agent is any biological agent or toxin that is named on the Select Agent list.
- Designation as a Select Agent is a careful and deliberate process
- Designating an unknown , novel agent as a Select Agent would require prediction of Select Agent status from sequence

Study findings

Can we predict Select Agents from Sequence? No.

Finding:

A sequence-based *prediction* system for oversight of Select Agents is not possible now or in the near future

Study findings

Can we predict Select Agents from Sequence? No.

There are two reasons for this:

1.) "Select Agent-ness" has both biological and non-biological components. Because the security threat posed by an agent is not determined by biological criteria alone, Select Agent status can never be predicted from sequence alone. "Select Agent" is not a scientific description, it is a regulatory designation.

Study findings

Can we predict Select Agents from Sequence? No.

1.) "Select Agent-ness" has biological and non-biological components.

Table 1.1: Prospects for *de novo* prediction of "Select Agent-ness" from sequence				
Property	**Predictable now?**	**Foreseeable future?**	**Maybe someday?**	**Never**
Pathogenicity			X	
Transmissibility			X	
Available treatments			X	
Ease of preparation			X	
Ease of dissemination			X	
Public perception				X
Historical bioweapon				X
Economic impact				X
Natural prevalence				X

Study findings

Can we predict Select Agents from Sequence? No.

2.) It is not feasible to predict the biological characteristics from sequence

- This is a prediction problem of the greatest complexity.
- There is no single characteristic that makes a microorganism a pathogen and no clear cut boundaries that separate a pathogen from a non-pathogen.
- Will require an extraordinarily detailed understanding of host, pathogen, and environment interactions integrated at the systems, organism, population, and ecosystem levels.
- High-level biological phenotypes like pathogenicity, transmissibility, and environmental stability, cannot plausibly be predicted with the degree of certainty required for legal purposes, either now or in the foreseeable future.

Study findings

Can we predict Select Agents from Sequence? No.

The goal of a predictive oversight system is so far out in front of current biological understanding that it would be unwise to attempt to address it in detail. Thus, we communicate only the following general 'milestones':

- Develop an accurate ability to predict the function of individual proteins from genome sequence.
- Develop an accurate ability to predict the output of biochemical, regulatory, and genetic pathways (modules) of several proteins acting together, from genome sequence.
- Develop an accurate ability to predict the behavior of a whole organism from its genome sequence.
- Develop an accurate ability to predict the interactions of organisms in their natural environment from their genome sequences, such as microbe/host symbioses or host/pathogen interactions.

Study findings

Good News and Bad News

Finding:

Prediction and design are linked

Good News and Bad News

The bad news:

Prediction and design are linked. Developing the ability to predict Select Agent pathogenicity from genome sequence raises serious dual-use concerns, because prediction and design go hand in hand. Accurate computational prediction of Select Agent characteristics from genome sequences enables computational design and optimization of bioweapon genome sequences.

Good News and Bad News

The good news:

Prediction and design are linked. Design and prediction go hand in hand; our lack of predictive ability in biology also means we cannot design genomes *de novo* at this time.

The feasibility of the problem and a solution are linked. Synthetic genomics poses three threat scenarios that would allow a "bad actor" to obtain a pathogenic organism with Select Agent properties; one of them is of most immediate concern and most readily addressed.

- "modification" of an existing Select Agent ---*feasible*
- "assembly" of a synthetic pathogen ---*possible*
- "de novo design" of a novel agent ---*improbable*

Study findings

Challenges to biosecurity and biosafety

Finding:

Synthetic genomics and the natural complexity of biology present challenges to biosecurity and biosafety that must be addressed well before prediction of biological function will be feasible.

Challenges to biosecurity and biosafety

Known, designated Select Agents: Need to provide increased clarity about which DNA sequences are subject to the Select Agent Regulations.

- boundaries around the taxonomic names of Select Agents on the list are unclear; i.e., how similar should two sequences be before labeling them with the same name.
- It is also unclear how much (which parts) of an agent must be present to be considered a Select Agent

Unknown, novel agents: Need to provide information and oversight for "sequences of concern" that are not themselves Select Agents, but that potentially could be used to produce a threat.

- to make it harder for individuals with nefarious intent to develop pathogens or toxins as weapons or as tools for bioterror without detection, and
- to avoid the accidental, inadvertent, or ill-advised production of hazardous constructs by well-meaning investigators.

Study findings

A Brighter Line

Finding:

A Sequence-Based Classification system for Select Agents and a "Yellow Flag" biosafety system for "sequences of concern" could be developed using current technologies.

A Brighter Line

Classification system for known Select Agents:

- **Sequence-based classification is strictly operational**, a set of tools for drawing decision boundaries around known sequences that do or do not belong to a desired classification. These tools are used now for robust and automatic classification of gene sequences into usefully annotated sequence families.
- **An operational definition of a complete Select Agent would not predict whether a sequence encodes a functional pathogen or not.** Sequence-based classification strategies would more sharply define the Select Agent Regulations to deal with issues raised by DNA synthesis, and natural variation.
- **Would establish a "brighter line":** an unambiguous procedure for deciding if a genome sequence is assigned one of the taxonomic names already on the Select Agent list.
- **Uses DNA Sequence information to better define Select Agents for regulatory purposes**

Yellow Flag

Yellow flag biosafety system for "sequences of concern":

- **Would function as an extension of biosafety.** A sequence-based classification system would help organize and condense knowledge about the genomic composition of dangerous pathogens and could be used to identify partial genomes and suspicious parts in the grey zone, triggering common-sense follow up.
- **Not regulatory in nature,** the yellow flag system could provide information relevant for biosecurity in a dynamic and timely fashion.

Study recommendations

Sequence-based classification is technologically feasible and may improve the current system; however, such a system does have limitations and potential negative consequences. Therefore, we do not specifically recommend that it be implemented. Rather, we make two recommendations:

- **The sequence space around each discrete taxonomic name on the Select Agent list should be clearly defined, so that Select Agent status can be unambiguously determined from a genome sequence (for example, by a DNA synthesis company). The sequence space should be broad enough to include the plausible modifications and chimeras that experts reasonably believe probably also act as Select Agents, without encompassing existing non-Select Agents.**
- **A sequence-based classification system could address this problem, and should be considered and weighed against the cost and complexity of implementing this technological augmentation to the current Select Agent Regulations.**

Study findings

NATIONAL ACADEMIES

Near-term Milestones for Sequence-Based Classification:

- **A sequence database with a Select Agent focus.**
- **An expanding sequence database of all biology.**
- **Define the criteria for Select Agent designation.**
- **Stratify the Select Agent list based on risk.**

Study findings

Long-term areas of research include:

- Protein structure and function
- Gene expression and regulation
- Pathogenic mechanisms
- Animal models of disease
- Data and information management for systems biology
- Synthetic biology
- Metagenomics and phylogenomics, including the human microbiome

Study conclusion

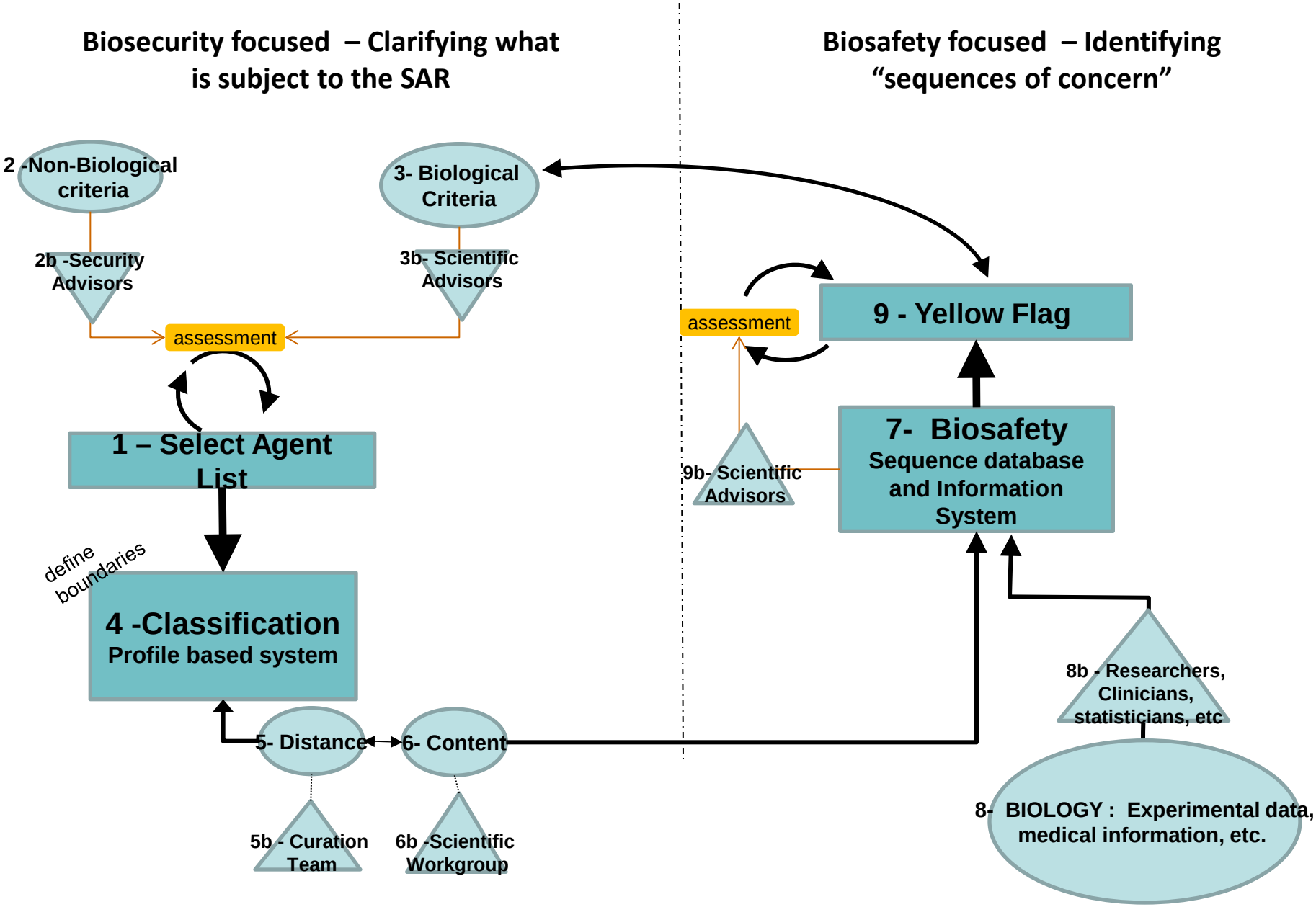
Our principal finding is that sequence-based prediction of Select Agent properties is not feasible, either now or in the foreseeable future; any dedicated research effort solely for this purpose is likely to have only negative consequences.

A sequence-based classification system and a yellow flag system are technologically feasible, but we have not carefully examined their cost nor their impact on basic research or national security .

Sequence-Based Classification for Select Agents: A Brighter Line

Thank you.

Sequence based framework with two complementary systems

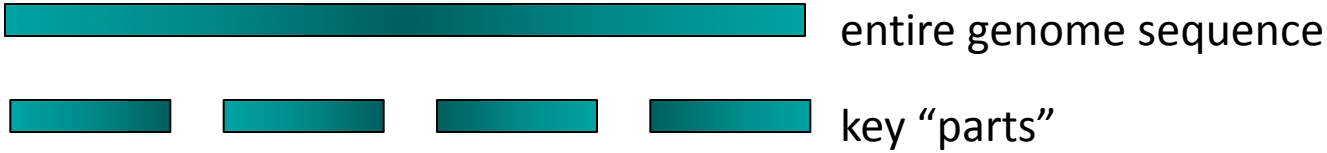


Study findings

Classification system – content and distance

Content -- what sequences must be present to qualify as a "complete" Select Agent?

SA1 – for example -*Bacillus anthracis*



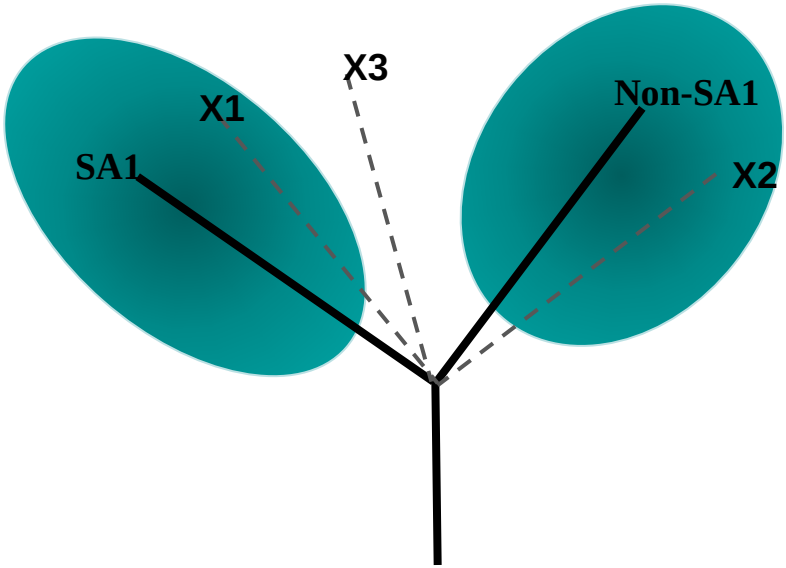
Content could be an entire genome sequence or might be considered "complete" content if key "parts" are present

Study findings

Classification system – content and distance

Distance -- how similar must the content be for a sequences to be classified as a known Select Agent?

SA1 – for example -*Bacillus anthracis*

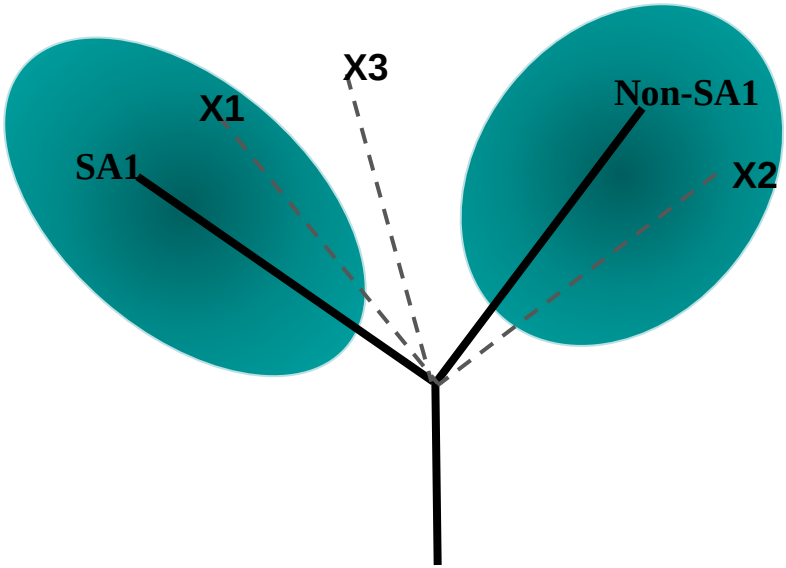


Study findings

Classification system – content and distance

Distance -- must be defined for the content of each Select Agent

SA1 – for example -*Bacillus anthracis*



SA2 – for example, *variola virus*