

Mathematics for Pathogenomics

PI: Andrzej K. Brodzik

Marc Colosimo (co-PI, CIIS)

Joe Francoeur (coding, CIIS)

Brian Sroka (consultant, CIIS)

Telephone: 781.271.6992

Email address: abrodzik@mitre.org

Sponsor/Funding: MSR

Problem

- ❑ **Bacillus anthracis is one of the best known and most lethal pathogens.**
- ❑ **Toxic and non-toxic anthrax strains are highly homologous.**
- ❑ **Similarity of strains significantly obstructs anthrax ID, which, in effect, has to rely on subtle, difficult to detect, differences among genomes.**
- ❑ **Current approaches are expensive, heuristic, and only partly effective in identifying anthracis strains.**
- ❑ **The problem is compounded as new strains are discovered and sequenced and as design of synthetic pathogenic sequences becomes possible.**

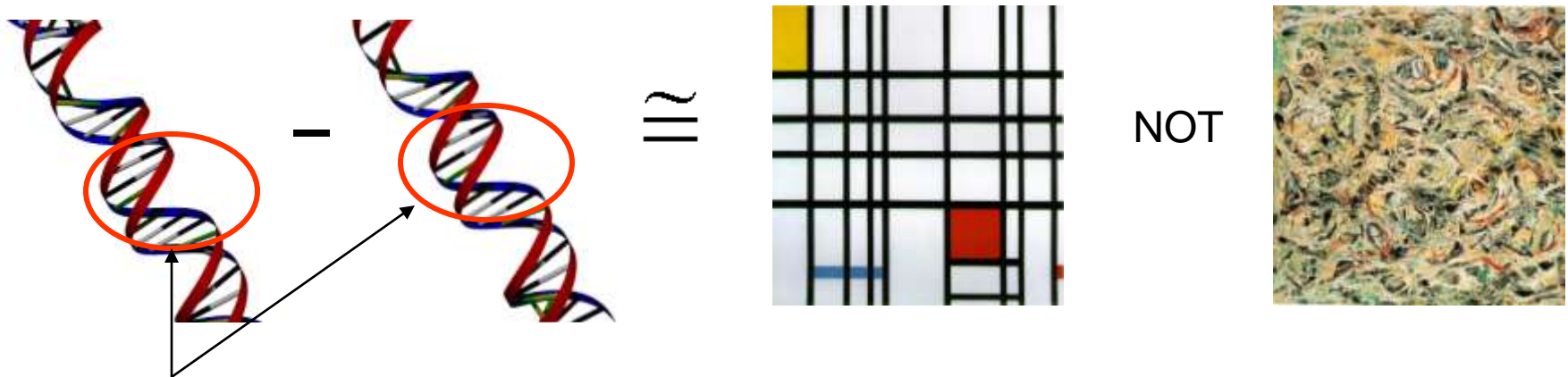
With *b. anthracis*, a potential biological weapon, the ability to identify, differentiate, and forensically track strains is crucial, and new research tools are critical.

NIAID Biodefense Agenda for CDC Category A Agents, 2006

Background

Tandem Repeats as Strain Markers

- ❑ The advantage of tandem repeats over other markers is in their high allelic diversity, resolving power, and mutation rates.
- ❑ These properties make tandem repeats well suitable for the analysis of highly homologous and evolutionary young genomes such as anthrax.

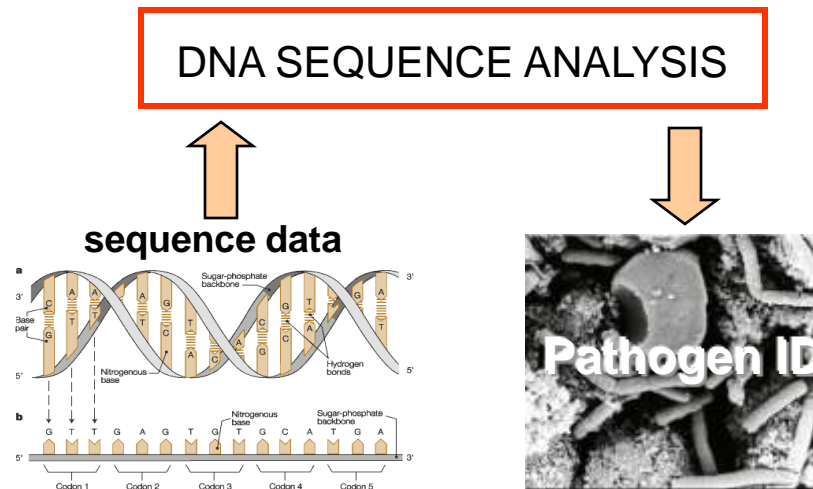


4 good reasons for relying on repeats:

- Prevalence
- Ease of modeling
- Biological significance
- Evolutionary variability

Objective

We explore novel mathematical approaches to DNA sequence alignment and pattern exploration to identify all tandem repeats in the anthrax genome and to design more effective strategies for anthrax strain ID.

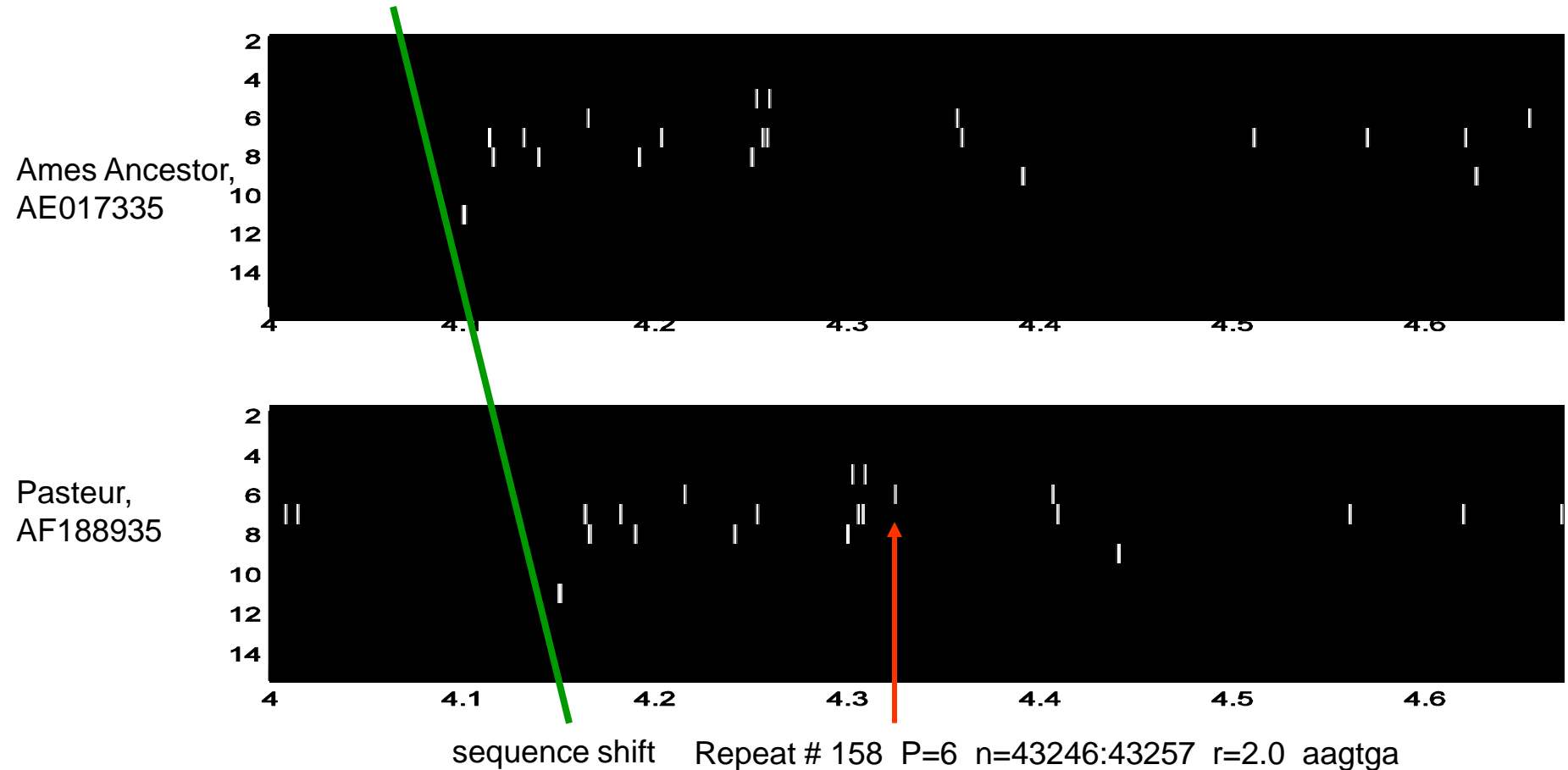


Activities

- Refinement of previously developed sequence analysis algorithms (Quaternionic Periodicity Transform).
- Generation of MATLAB and C codes.
- Characterization of a selected b. anthracis strain in terms of exact and approximate tandem repeats.
- Generation of a list of strain markers.
- Evaluation of effectiveness of the proposed method.

Highlight

Comparison of pXO2 Plasmids



QPT detected **339** VNTR:

Pasteur has **5** unique VNTRs: 2(AAGTGA)6, 2(TTTTCTT)6, 2(TGCTTC)7, 2(GTGACGTT)8, 2(GTATCACC)8

Ames has **2** unique VNTRs: 4(T)1, 2(GATTTTTTTT)9

Highlight (cont.)

Comparison of pXO1 Plasmids (cont.)

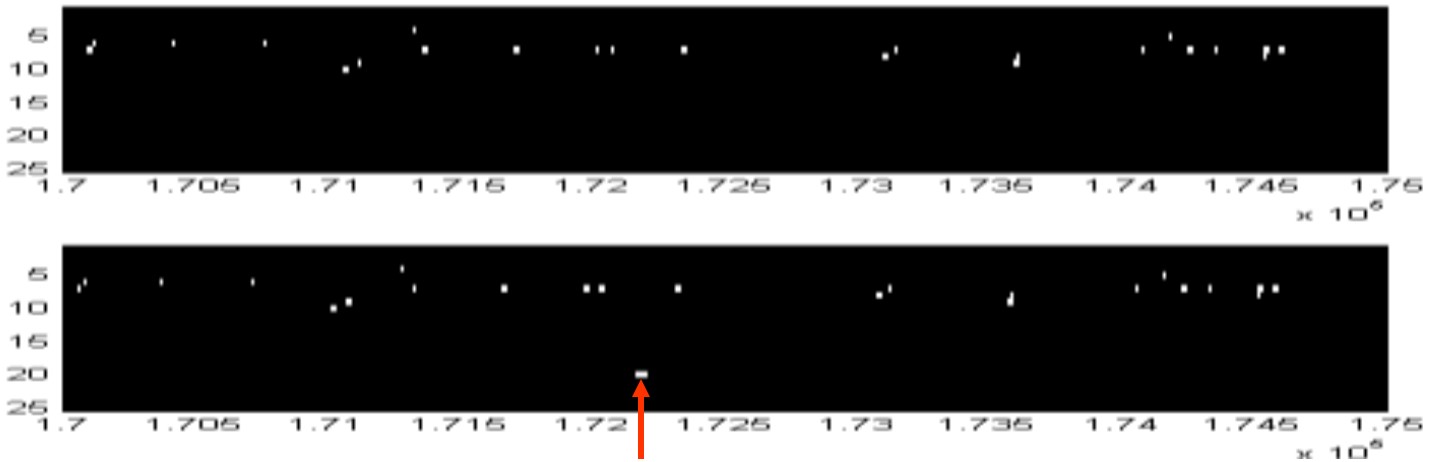
Repeat summary (578 repeats):

- #24, 7mer, n=5737, Ames
- #108, 7mer, n=36544, Ames
- #542**, 20mer, n=172169, Sterne
- #124-142, n=43328 (5.4K), Ames/Sterne
- #368-512, n=117285 (44K), Ames/Sterne

Ames Ancestor,
AE017336

Magnification:
170-175K

Sterne,
NC_001496



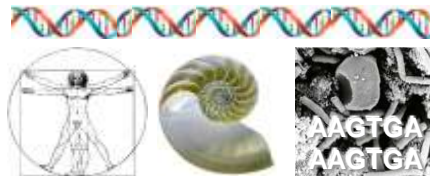
Single repeat # 542 P=20 n=172169:172215 r=2.4 GAAAAATAAATATATAATAG...

Impacts

- ❑ **A more efficacious forensic approach towards anthrax strain ID will be developed.**
- ❑ **The approach can be used to distinguish between naturally occurring and manufactured pathogens.**
- ❑ **Focus of the MSR is on pathogenic sequences. However, DNA sequence analysis tools are not anthrax-specific and can be used to investigate any genomic sequence. Application examples include:**
 - ❖ Personalized medicine (predisposition to genetic diseases)
 - ❖ *Ab novo* gene finding
 - ❖ Phylogenetic studies of genomes (“tree of life”)
 - ❖ Synthetic biology applications
- ❑ **If successful, this effort might lead to the formation of a center of expertise in CMB at MITRE (MB + MATH + HPC).**

Future Plans

- ❑ **Characterization of several anthracis strains and development of anthrax strain ID.**
- ❑ **A possibility to broaden the MSR goal, a multi-purpose sequence analysis workbench.**
 - ❖ **Pattern detection**
 - ❖ **Sequence alignment**
 - ❖ **Spectral analysis**
 - ❖ **Data compression**
- ❑ **Metagenomic and human DNA investigations.**



Mathematics for Metagenomics