Whole Genome, Physics-based Sequence Alignment for Pathogen Signature Design

Jason Gans\(^1\), Wu-Chun Feng\(^2\), Murray Wolinsky\(^1\)

\(^1\)Los Alamos National Laboratory
\(^2\)Virginia Tech

The need for pathogen detection
Specific vs broad spectrum signatures

Specific Signature

GACTATA...  

Broad Spectrum Signature

ATGCCTAAT...

Detecting multiple targets with a single experiment reduces cost

DNA-based detection assays

DNA signature development can exploit the growing number of sequenced genomes
- ~ 400 bacterial genomes available
- ~ 8000 viral genomes available

Mature technologies for high throughput detection
- Gene chip, PCR
Parallel Computation (Moore’s Law is not enough)

Signature Detection Pipeline

1. Partition DNA sequence data
2. Identify sequence fragments unique to target species (i.e. not found in background sequences)
3. Identify unique fragments that are found in multiple target species
4. Compute minimal set of unique fragments required to span target species set
5. Convert unique fragments into biological assays
6. Computationally and experimentally validate assays
How to define sequence similarity/uniqueness?

The answer depends on how we ask the question

The assay format, e.g.
- PCR amplification
- DNA chip probe hybridization
- Single base extension

defines similarity.

A sequence similarity metric that supports common assay formats is DNA melting temperature, $T_m$

\[
T^B_{m,A} > T^H_{m} \implies \text{Sequences A and B are equivalent} \\
T^A_{m,B} < T^L_{m} \implies \text{Sequences A and B are dissimilar}
\]
How to define sequence similarity/uniqueness?

\[ T_M = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln \left( \frac{C_T}{N} \right) } \]

- \( R \) = Gas constant
- \( C_T \) = Strand concentration
- \( N \) = sequence dependant constant

**Nearest Neighbor Model**

\[ \Delta H^\circ = \Delta H_{\text{init}} + \Delta H_{\text{sym}} + \sum_{i=A,T,G,C}^{j=A,T,G,C} n_{ij} \Delta H_{ij} \]

\[ \Delta S^\circ = \Delta S_{\text{init}} + \Delta S_{\text{sym}} + \sum_{i=A,T,G,C}^{j=A,T,G,C} n_{ij} \Delta S_{ij} \]

---

**Partition Target Sequences**

1. **Background** (related non-pathogens)
2. **All genome sequences**
3. **Target** (pathogens of interest)
4. **Ignore** (distantly related)
**Isolate sequences unique to targets**

**Why search for unique sequence?**

Non-unique signatures susceptible to false positives

\[ S = \sigma_1 \text{ AND } \sigma_2 \]

- \( \sigma_1 = \text{ATGACCATGA} \ldots \)
- \( \sigma_2 = \text{GCATTCAGTA} \ldots \)

**Isolate sequences unique to targets**

Detection Signature

\[ S = \sigma_1 \text{ OR } \sigma_2 \]

- \( \sigma_1 = \text{ATGACCATGA} \ldots \)
- \( \sigma_2 = \text{GCATTCAGTA} \ldots \)
Isolate sequence fragments unique to target species

Target sequence

Background sequence

Sequence fragments unique to target species

Sequence fragments common to target and background

1. Use mpiBLAST to identify regions of similarity between \( U_i \) and \( B \).
2. If a sequence fragment has a \( T_m > 55 \text{ °C} \) (over a \( \leq 21 \) base window) then it is considered a match.

Using mpiBLAST to compute the set difference.

\[
U = T - (T \cap B) \\
U \approx U_f \\
U_{i+1} = U_i - (U_i \cap B) \\
U_0 = T \\
U_i \approx U_3
\]
Isolate sequences unique to targets: 1st generation

mpiBLAST
- Pro: Allowed signature design to use all available bacterial genomes
- Con: Required iteration and separate computation of set difference

2nd generation implementation
- Use high level structure of mpiBLAST (i.e. master/worker, database segmentation, query segmentation, MPI/C++)
- Directly compute set difference: *in silico* subtractive hybridization

In *silico* subtractive hybridization

*Diagram showing the process of subtractive hybridization with roles and sequence interactions.*
**Insilico subtractive hybridization**

- portion of background sequence in ram
- background sequence database
- query sequence
- worker
- master

**In silico subtractive hybridization**

- portion of background sequence in ram
- background sequence database
- query sequence
- worker
- master
Target fragment alignment

*In silico* subtractive hybridization produces fragments unique to every target species

- \( \text{species A} \rightarrow \{\text{ACTGGATCGATC, GGCTGGATTCTAGG, TAGGCTTAGGCTTA, ATTCGGCCCAGATAG, ...}\} \)
- \( \text{species B} \rightarrow \{\text{GCTTCTAGACAAAC, ATGGCGATTAGCCA, GCTAAGCCTAGCTA, TTTGACTAGATCAC, ...}\} \)
- \( \text{species C} \rightarrow \{\text{GGCTAGGCCCA, GGCTCTAGGATATAC, CGTCTAGGCTTATAT, CGGATTCGGCTTAG, ...}\} \)

Need to associate every unique target fragment with one or more target genomes

- \( \text{CGTAGGCTTAGGATAT} \downarrow \) \( \text{(species A, species B, species C)} \)
- \( \text{AGGCTATAGCGGA} \downarrow \) \( \text{(species B, species D, species E, species F)} \)
- \( \text{CAATAGGCTTATAAGGATTA} \downarrow \) \( \text{(species A, species F)} \)

Target fragment alignment

- portion of unique target sequences
- In ram
- unique target fragments
- worker
- master
Target fragment alignment

portion of unique target sequences in ram
unique target fragments

worker
master

Find the minimal set of signatures that will detect all target species

\( s_i \) is a set of species that contain a common sequence fragment
Find the minimal set of sequence fragments that will detect all target species

Given:

\[ T = \{A, B, C, D, E, F, G, \ldots, \} \]
\[ s_0 = \{E\} \]
\[ s_1 = \{D, E, G\} \]
\[ s_2 = \{B, C, F, H\} \]

Our task is to find \( G \equiv \{s_i, s_j, s_k, \ldots\} \)

Where \( T = U G_i \)
and
\( |G| \) is a global minimum

The set coverage problem is NP hard!

---

Find the minimal set of sequence fragments that will detect all target taxa

Approximate solution

Solve for \( G \) using Metropolis Monte-Carlo driven simulated annealing

Trial move generation: \( G_i \rightarrow G_{i+1} \)
Add random \((0,n)\) number of sets
and
Delete random \((0,n)\) number of sets

Acceptance criteria: \( G_{i+1} = \)
\[ G'_{i+1} \]
\[ |G'_{i+1}| \leq |G_i| \]
or
\[ \pi \leq \exp(-|G'_{i+1}| - |G_i|)/T) \]
\[ \pi \in [0,1] \]
Filter signatures with experimental constraints: PCR Primers

DNA Amplification Using PCR

1. Reaction mixture contains target DNA sequence to be amplified.
2. Reaction mixture is heated to 95°C to denature target DNA.
3. Reaction mixture is cooled to 50°C to allow primers to hybridize to complementary sequences in target DNA.
4. Reaction mixture is heated to 72°C, Two primers anneal, resulting in two complementary strands from primers.
5. First synthesis cycle results in formation of new copies of target DNA sequence.
6. Reaction mixture is heated to 95°C, DNA strands are denatured.
7. Reaction mixture is cooled to 50°C, primers anneal.
8. Second synthesis cycle results in four copies of target DNA sequence.

Forward primer: 5' - [Sequence] - 3'
Reverse primer: 5' - [Sequence] - 3'

Amplicon: 5' - [Sequence] - 3'

Filter signatures with experimental constraints: PCR Primers

Unique signature = Forward primer
“Consensus” region = Reverse primer

Maximum amplicon length
61 taxa detected using 15 PCR primer pairs
Family

Biothreat Viral Signatures
(1565 targets, 97 PCR primer pairs)
Biothreat Viral Signatures
(1565 targets, 97 PCR primer pairs)

PCR Primers and TaqMan® Probes: HIV-1

1005 strains detected with 22 PCR primer/probe sets
PCR Primers and TaqMan® Probes: HIV-1

Subtype M

Subtype N

Subtype O

Recombinant M/O strain

Acknowledgements

Informatics Team
Byron Delaney
Norman Dogget
Rob Leach
Jian Song
Chris Stubben
Murray Wolinsky

Experimental Validation
John Dunbar
Lance Green
Scott White

mpiBLAST
Lucas Carey
Aaron Darling
Wu-Chun Feng

Funding
DHS
TSWG