

MLSTONES – an Organic Model for Detecting Cyber Events

What's in your haystack?

Elena Peterson

Elena@pnl.gov



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

The Challenge was Data ...

- ▶ The data tsunami is here – And getting bigger every day
- ▶ Approaches to dealing with massive data
 - Algorithms
 - Hardware
 - Programming models
- ▶ Despite ever growing data volumes, we need
 - Better answers
 - Faster
 - Cost-effective



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Problems to be solved

- ▶ Identify potential threats in a large body of data
 - Quickly (near-real time)
 - Not sure what you're looking for
- ▶ Characterize a set of unknown data
 - Simple characteristics aren't helpful (file extension, size, mime-type,...)
- ▶ Identify new threats that have not been identified before
 - Related digital artifacts, files, behavior



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

General Approach

- ▶ Apply a protein based model to the data
 - Structure infers function
- ▶ Use standard algorithms and tools from biology to process data
- ▶ Continue to apply biological concepts to cyber data
 - Family trees
 - Evolution
 - Motifs (signatures) – single representation of a family
 - Inference
 - Behavior
 - Lineage

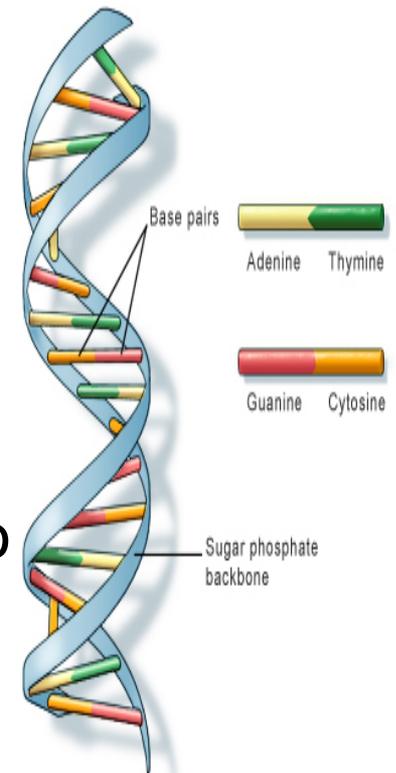


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Short Background on the Biology

- ▶ Information is encoded and passed down through genes to proteins
 - Proteins are the functional units performing most of the chemical and structural tasks needed for survival
 - The 20 amino acids can be mapped to 20 alpha characters and represented as text strings AGHTVFDS
- ▶ The sequence of a protein is highly related to it's function
 - So when sequences are similar the functions are very likely to be similar
 - Matches don't have to be "exact"



U.S. National Library of Medicine

Approach

Domain Specific

- ▶ Disassemble/Process
 - Domain dependent – need lowest level of representation
- ▶ Normalize and Mapping
 - Put “items” into 20 (or more) buckets

Repeatable Process

- ▶ Convert to protein format
- ▶ Find artifacts with conserved regions
 - Using BLAST – compare all things to all things
- ▶ Build families
 - Clustering based on various metrics
- ▶ Create motifs
 - Using multiple alignment concepts from biology



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Why does this apply to cyber events?

- ▶ We can do it at scale
 - Large data (biology is relatively small)
 - 7.5M binaries represented by 500 signatures
 - Fast (standard bio tools are slow)

Query size	MLSTONES	BIO
100 vs 100	.5 seconds	27.5 minutes
100 vs 1M	4 minutes	4 days
1M v 1M	4 hours	Never

- ▶ We can do it for various data types
 - Data streams, binaries, logs, network traffic, ...



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Benefits

- ▶ High applicability for various cyber event data –
 - Many cyber events can be represented as text strings
 - Order matters where frequency does not
- ▶ Provide a mechanism for quick/early decision making
- ▶ This process can ‘evolve’ and learn
- ▶ This method is *proactive* rather than reactive
 - Not rule-based
- ▶ Leverages well established algorithms from biology
- ▶ Most obfuscation techniques are mitigated naturally



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

More justification

- ▶ Pattern detection in biological systems provides
 - A rich base of knowledge for detecting correlations
 - A useful vocabulary for describing relationship and inheritance in those correlations
- ▶ Applying biological principles of inheritance makes it possible to
 - Annotate or infer function
 - Provide a framework for forensic analysis inference
 - Quantifying how network session profiles are related based on highly conserved traits



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

How do we Compare?

- ▶ Current methods for identifying signatures is reactive and requires human intervention to find meaningful pattern
 - MD5 hashes
 - Anti-virus
- ▶ General pattern matching in large data means we could find meaningless patterns everywhere
 - Haven't reduced the size of the problem
- ▶ Verifying that a pattern has *meaning* is difficult and time consuming
 - We automate that process
- ▶ MLSTONES is a fuzzy matching/lossy process that cuts through data quickly
 - Makes the playing field reasonable to play in



Risks

- ▶ Need Hardware to scale
 - Commodity hardware – no specialization required
- ▶ Some obfuscation techniques are not natively handled
 - Ex: Packers
 - Specific to the domain
- ▶ If you understand the algorithm well enough you can defeat it (in some cases)
 - It would take work and a true understanding
 - Ongoing process
- ▶ There will always be false negatives/false positives



Applications

- ▶ Malware classification & detection
- ▶ Netflow behavior classification & detection
- ▶ Document classification
- ▶ Server identification
- ▶ Fraud/Anomaly Detection
- ▶



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Plan to develop and deploy

- ▶ Further operationalize tools into a deployable framework
- ▶ Identify customers with needs and data suitable for the process
- ▶ Gather (or generate) representative data for testing and evaluation
 - Reduce false negative/false positive rates
- ▶ Modify current framework for client specifics
- ▶ Deploy to client for pilot testing and evaluation



Conclusion

- ▶ This methodology will be able to solve problems that can't currently be solved at least in a reasonable timeframe
- ▶ This can be added to an enterprise without replacing or surpassing other systems in place
 - Another tool in the arsenal of cyber defense



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965