

Genome-wide Analysis of Conserved DNA Signatures in the Anthrax Pathogen Identified by Pattern-based Comparative Genomics

Zayed I. Albertyn, Tan Ka Ju, Wong Chee San, M. Ramachandran and Robert G. Hercus

To counter the possible threat of bioterrorism and microbial pandemics the mechanisms pathogens use in infection and proliferation need to be clearly understood. Detection of signature patterns from genomic data provides the most comprehensive path to accurate and sensitive diagnosis. Microbial genome databases built using SynaBASE™ were used to analyse Bacillus anthracis Ames pathogen strain by comparison to ALL publicly available bacterial genomes. Results indicate that it is possible to identify the shared patterns between anthrax and all sequenced eubacteria in less than ten minutes per query. To our knowledge, referencing database pattern information in terms of this degree of scalability has not been attempted before. An analysis of three separate comparisons allowed identification of regions in the anthrax genome that are not unique to anthrax and are shared with other bacterial species. These regions of conservation in the anthrax pathogen showed significant matches with genes implicated in anti-microbial drug resistance. This approach provides a starting point for identifying unique DNA signatures specific to anthrax. Technologies based on SynaBASE may enable faster and more accurate analysis for high-throughput applications geared towards disease diagnostics and vaccine design.

Introduction

Microorganisms play an important role in our everyday lives; they are essential to our physiology, are exploited for their biotechnological value in multiple economically important industries, and are critical to the sustainability of the earth's ecology. However, a great variety of harmful strains or pathogens occur naturally and these are under intense scrutiny by the medical research community for disease eradication. Bacterial and viral pathogens have also been developed into biological weapons as they are relatively easy to manufacture and deliver, as well as being highly effective.

Accurate and rapid diagnosis of a pathogen is most important when faced with a critical situation that could result from a major food poisoning incident, a pandemic or a scenario suggesting a terrorist attack. The two most commonly used approaches to pathogen detection focus on identifying pathogen-specific DNA or protein regions; via DNA amplification and protein assays, respectively [1].

The optimal method for identification of unique, strain-specific and robust pathogen DNA signatures

is to screen every available microbial genome *in silico* to determine which sequence patterns need not be considered for PCR primer design. A proprietary structured network database technology-SynaBASE™, was used to identify and structure patterns from all publicly available microbial genomes.

The objective of this study was to highlight information content within SynaBASE applied to comparative microbial genomics, using the anthrax pathogen as a reference. This principle could be extended across the spectrum to viral and eukaryote parasite genomes.

Methods

The genome of *Bacillus anthracis* (the anthrax pathogen) was compared to those of all publicly available bacterial genomes to discover instances of intra- and inter- species sequence conservation.

Three SynaBASEs were built from sequence data

comprising of (i) all *Bacillus* species, (ii) *Bacillus anthracis* subspecies and (iii) the entire bacterial genome set downloaded from Genbank [4]. The genome of *Bacillus anthracis* was compared to all the genomes in the three databases stated above, using SynaCompare™ (SynaBASE's pairwise alignment tool) in a one versus all mode. A minimum sequence match length of 16bp was used to seed alignments. All comparisons were run on a single Linux HP® Intel Itanium™ architecture using a single 1.3GHz CPU accessing 64GB RAM.

The 5.2 Mbp *Bacillus anthracis* Ames strain's genome (Refseq ID NC_003366) was used as the query sequence for SynaCompare [2-3]. SynaCompare computes matches between a query and target sequence using SynaBASE patterns to seed and extend alignments. The final result is a pairwise alignment with the matching blocks coloured according to frequencies of patterns within a SynaBASE [3].

Selected regions of the *Bacillus anthracis* genome that exhibited significantly high pattern frequencies in a SynaCompare analysis against all bacterial genomes were extracted and searched against a SWISSPROT SynaBASE for gene annotations [5]. SynaSearch™, a database searching tool developed for SynaBASE, was used to find all significant alignments [6].

anthracis are shared with other bacterial species (see Figure 2 on the next page). These results may be further analysed to identify whether they represent annotated genes or non-coding regions.

Five regions of the anthrax genome that were highly frequent in all sequenced bacterial genomes were searched against a SWISSPROT SynaBASE to determine whether these contained any genes of biological interest. Results indicate that many important genes such as DNA gyrase, Inosine-5'-monophosphate dehydrogenase (IMPDH), Lysyl-tRNA synthetase and UDP-N-acetylmuramoyl ligase (see Table 1 on the next page) show significant matches to the five regions outlined above.

DNA gyrase is an enzyme specific to bacteria and introduces supercoils into genomic DNA. Gyrase enzymes are important drug targets and play a pivotal role in antimicrobial drug resistance to fluoroquinolones [7-8]. These enzymes are under investigation as targets for a broad range of bacterial pathogens including *E. coli* subspecies, *Neisseria gonorrhoeae* and *Mycobacterium tuberculosis* [9-10]. IMPDH is necessary for GTP energy synthesis and significant differences are observed between microbial and human orthologues at the enzyme's active site [11]. IMPDH genes are also implicated in mycophenol drug resistance exhibited by the

Results & Discussion

Genome alignments were rapidly computed from microbial sequence patterns stored in SynaBASE using SynaCompare. The largest comparison; *B. anthracis* Ames genome against all known bacterial genomes was processed in only 7 min 42s (see Figure 1). Relative pattern frequencies are shown colour coded for all pattern-derived alignments. Therefore a full alignment view in a one versus all comparison may be used to determine the relative level of sequence conservation within a database. From the pattern frequencies and horizontal alignment blocks it can be inferred that specific genomic DNA signatures in *Bacillus*

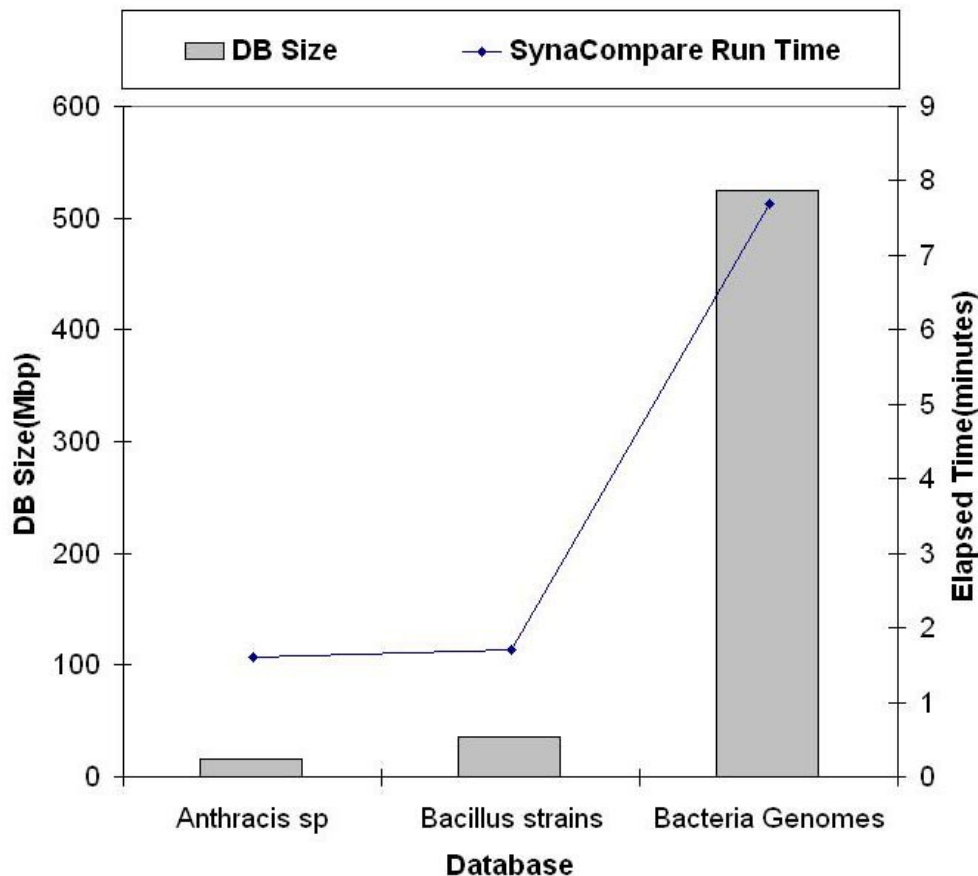


Figure 1. SynaCompare statistics for the genome comparison of the *Bacillus anthracis* Ames genome to all *Anthracis* strains, *Bacillus* species and all bacterial genomes. Runtime and database sizes are shown for each comparison.

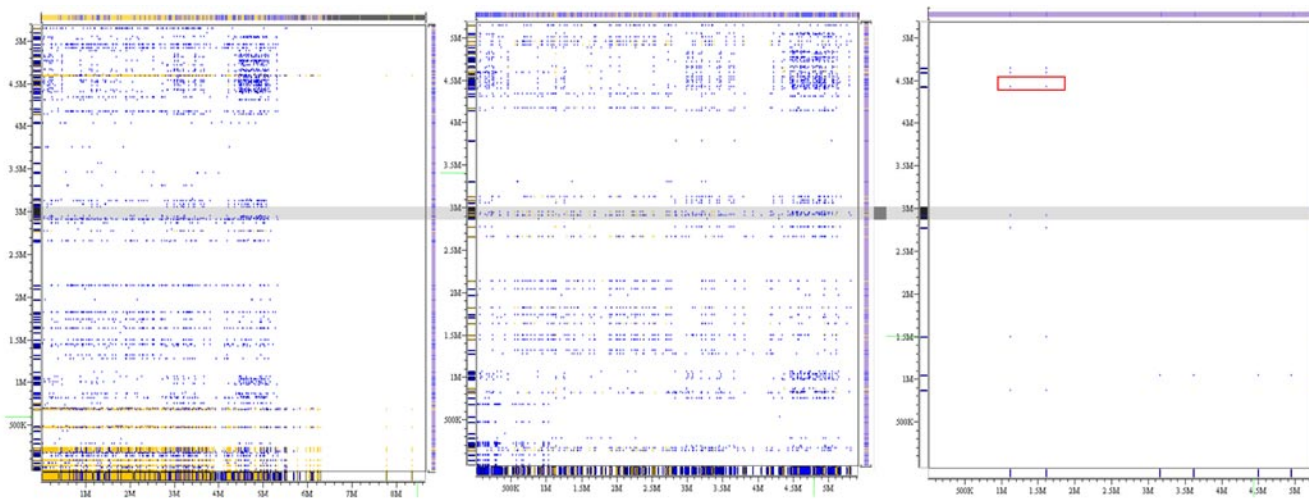


Figure 2. Multiple SynaCompare , one versus all alignment views for the *Bacillus Anthracis* Ames genome comparisons. Database pattern frequencies for each database are colour codes as high (yellow), medium (blue) and low (purple).

Genome Sequence Position (bp)	Length (bp)	Number of SynaSearch Matches	Score ¹	Protein Description	Search Speed (seconds)
1-10,282	10,282	50	405	a) Chromosomal replication initiator protein dnaA b) DNA gyrase subunit A	0.949
14,521-35,712	21,192	50	267	a) Inosine-5'-monophosphate dehydrogenase b) Seryl-tRNA synthetase	2.783
69,619-88,692	19,074	50	444	a) Lysyl-tRNA synthetase b) Cell division protein ftsH homolog	2.146
133,195-152,268	19,074	50	296	a) DNA-directed RNA polymerase alpha chain b) 30S ribosomal protein S9	2.108
232,797-247,631	14,835	38	116	a) UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase b) Alanine racemase	1.499

¹ Calculated as the number of matching bases in the highest scoring local alignment between a query and database target.

Table 1. Annotations for regions in the anthrax genome that share significant sequence conservation with all sequenced Eubacteria genomes in Genbank

pathogenic bacterium *Streptococcus pyogenes* [11]. Therefore a comparison of *Bacillus anthracis* to all bacterial genomes reveals conservation of known drug targets and resistance factors essential for pathogen survival.

The discovery that Lysyl-tRNA synthetase proteins match the anthrax pathogen's genome confirm previously published evidence of horizontal gene transfer of antimicrobial resistance factors [12]. UDP-N-acetylmuramoyl ligase facilitates peptidoglycan biosynthesis in *E. coli* and *Streptococcus pneumoniae* [13]. The conservation of these genes in the anthrax pathogen may imply shared biochemical pathways that lead to pathogenicity.

Conclusion

A comparison of three alignment views indicates sequence conservation across three databases of

microbial genomes when compared to anthrax. Recent work on microbial genomes has shown that common regions shared between and within bacterial strains can be used to identify possible virulence factors [14]. The SynaCompare results from all three comparisons confirmed the existence of these shared sites that may harbour genetic factors essential to survival of pathogen species.

A previous comparison between the *Bacillus anthracis* and *Bacillus cereus* genomes revealed common genes responsible for inhalation anthrax type symptoms caused by the former organism [14]. Therefore an exhaustive comparison using this method could be applied in a similar fashion to identify genomic patterns related to pathogen specific diseases. The discovery of genomic regions that harbour genes important to antimicrobial drug resistance in other bacteria could aid in identifying shared mechanisms of pathogenicity in anthrax.

Furthermore, the added advantage of significant speed improvements allows multiple iterative analyses to be conducted as well as enabling integration into more complex pipelines for the rapid identification and biological characterization of harmful microbes. Such ultra-high-throughput technology may prove essential for the present day fight against bioterrorism and new variants of diseases such as SARS and avian influenza.

REFERENCES

- Slezak T, Kuczumarski T, Ott L, Torres C, Medeiros D, et al (2003). Comparative genomics tools applied to bioterrorism defence. *Brief Bioinform* 4,133-49.
- Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT et al (2003). The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* 423, 23-5.
- Albertyn ZI, Anwar A., Ching S and Hercus, R (2004). Applications of a novel structured network database to genome-genome comparisons. Available online from www.synamatix.com, Research Newsletter April 2004.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (1999). GenBank. *Nucleic Acids Res.* 27,12-7
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC., Estreicher A., Gasteiger E., Martin MJ., Michoud K., O'Donovan C, Phan I, Pilbout S and Schneider M(2003).The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31,365-370.
- Albertyn ZI, Wong C, Liang Chung T, Teong Chuan N, Ramachandran M, Ching S,Anwar A and Hercus R (2004).SynaProbe™ - An Ultra High Speed Application for Whole Genome Microarray Probe Design
- Bellon S, Parsons JD, Wei Y, Hayakawa K, Swenson LL et al (2004)Crystal structures of *Escherichia coli* topoisomerase IV ParE subunit (24 and 43 kilodaltons): a single residue dictates differences in novobiocin potency against topoisomerase IV and DNA gyrase. *Antimicrob Agents Chemother.* 48 1856-64.
- Wall MK, Mitchenall LA, Maxwell A. (2004) *Arabidopsis thaliana* DNA gyrase is targeted to chloroplasts and mitochondria. *Proc Natl Acad Sc.*101,7821-6.
- Aubry A, Pan XS, Fisher LM, Jarlier V and Cambau E (2004) *Mycobacterium tuberculosis* DNA gyrase: interaction with quinolones and correlation with antimycobacterial drug activity. *Antimicrob Agents Chemother.* 48.1281-8.
- Dan M (2004).The use of fluoroquinolones in gonorrhoea: the increasing problem of resistance. *Expert Opin Pharmacother.*5,829-54.
- Zhang R, Evans G, Rotella FJ, Westbrook EM, Beno D, Huberman E, Joachimiak A and Collart FR (1999). Characteristics and crystal structure of bacterial inosine-5'-monophosphate dehydrogenase. *Biochemistry* 38, 4691-700.
- Brown JR, Gentry D, Becker JA, Ingraham K, Holmes DJ and Stanhope MJ (2003). Horizontal transfer of drug-resistant aminoacyl-transfer-RNA synthetases of anthrax and Gram-positive pathogens. *EMBO* 4,692-8.
- Ehmann DE, Demeritt JE, Hull KG and Fisher SL (2004). Biochemical characterization of an inhibitor of *Escherichia coli* UDP-N-acetylmuramyl-l-alanine ligase. *Biochim Biophys Acta* 1698,167-74.
- Hoffmaster AR, Ravel J, Rasko DA, Chapman GD, Chute MD, et al (2004). Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc Natl Acad Sci* 101,8449-54.

To request detailed technical information or feedback please send an email to tech@synamatix.com

This Research Newsletter is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes only and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaMine, and SynaSuite and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.