Probabilistic Latent Semantic Analysis Applied to Whole Bacterial Genomes Identifies Common Genomic Features

J. Rusakovica, J. Hallinan, A. Wipat* and P. Zuliani

School of Computing Science, and Centre for Synthetic Biology and Bioexploitation, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Summary

The spread of drug resistance amongst clinically-important bacteria is a serious, and growing, problem [1]. However, the analysis of entire genomes requires considerable computational effort, usually including the assembly of the genome and subsequent identification of genes known to be important in pathology. An alternative approach is to use computational algorithms to identify genomic differences between pathogenic and non-pathogenic bacteria, even without knowing the biological meaning of those differences. To overcome this problem, a range of techniques for dimensionality reduction have been developed. One such approach is known as latent-variable models [2]. In latent-variable models dimensionality reduction is achieved by representing a high-dimensional data by a few hidden or latent variables, which are not directly observed but inferred from the observed variables present in the model. Probabilistic Latent Semantic Indexing (PLSA) is an extention of LSA [3]. PLSA is based on a mixture decomposition derived from a latent class model. The main objective of the algorithm, as in LSA, is to represent high-dimensional co-occurrence information in a lower-dimensional way in order to discover the hidden semantic structure of the data using a probabilistic framework.

In this work we applied the PLSA approach to analyse the common genomic features in methicillin resistant *Staphylococcus aureus*, using tokens derived from amino acid sequences rather than DNA. We characterised genome-scale amino acid sequences in terms of their components, and then investigated the relationships between genomes and tokens and the phenotypes they generated. As a control we used the non-pathogenic model Gram-positive bacterium *Bacillus subtilis*.

1 Introduction

The spread of drug resistance amongst clinically-important bacteria is a serious, and growing, problem [1]. Bacteria can acquire resistance to a range of drugs via the transmission of cassettes of drug resistance genes carried on plasmids. In order to treat patients effectively, and reduce the unnecessary use of antibiotics, rapid identification of the resistance status of the bacteria involved in an infection is essential. Currently, this process involves taking samples from the patient and growing up the bacteria involved in the presence of a range of different antibiotics. This process is slow and labour intensive, and provides information only about the specific antibiotics tested. The advent of relatively cheap, rapid DNA sequencing has led to the possibility of using this information for analysing the antibiotic resistance status of bacteria directly from their genomes [4, 5]. However, the analysis of entire genomes requires considerable computational effort, usually including the assembly of the genome and subsequent identification of genes known to be important in pathology. An alternative approach is to use computational algorithms to identify genomic differences between

* To whom correspondence should be addressed. Email: anil.wipat@ncl.ac.uk

pathogenic and non-pathogenic bacteria, even without knowing the biological meaning of those differences.

The concept of a pan-genome was first introduced in 2005 by Medini et al. [6], who proposed that a microbial assemblage can be considered to comprise a single common genome, rather than being a collection of individual genomes. Medini et al. divided the pan-genome into two groups: genes common to all species, the core genome; and genes present in only some of the species, the peripheral genome. The peripheral genome is further divided into genes unique to a particular species or strain, and those shared between a fraction of the population. Genes in the core genome are those responsible for basic maintenance common to all bacteria, while those in the peripheral genome adapt individual species to the particular environmental niche in which they exist [7]. As environments change, peripheral genes may be gained or lost, while core genes tend to be more resistant to evolutionary change.

Applying the pan-genome concept, it should be possible, in principle, to determine the degree of similarity between different bacterial species or strains by identifying genome-level patterns to find conserved genome fragments, the proportion of individual genomes which they occupy, and their correlation with environmental factors. Pattern recognition of this type has long been a subject of research [8], and is increasingly being applied to biological problems [9]. However, many pattern recognition techniques are computationally expensive, and the computational effort increases exponentially with the dimensionality of the datasets.

1.1 Sliding Window Approaches to Sequence Analysis

Biological sequences, whether DNA, RNA or amino acid are frequently analysed one letter at a time. However, more genetic context may be gained by analysis at the level of tokens: *n*-mers of bases or amino acids. For a token of length five there are 1024 possible tokens from a DNA sequence, but 3,200,000 in an amino acid sequence. The analytical space of an amino acid sequence is therefore far richer and more complex than that of a DNA sequence.

A genome can be converted into a set of tokens using a sliding window of length *n* over the sequence. The window is positioned at the start of the sequence, and the first *n* letters (positions 0 to *n*-1) recorded as a token. The window is then moved forward by one letter, and the next token (positions 1 to *n*) recorded. In this way the entire genome sequence can be converted into a set of overlapping tokens. Ignoring position information, the entire set of tokens may be treated as a bag of words (BOW), following approaches previously used in text mining [10]. Each genome can be represented using a vector space model, in which each vector is a token, and each token has a weight, representing the number of times it occurred in that genome.

Vector spaces may be very high dimensional. For example, 2,000 tokens may form 2000^{2000} different vectors. Bacterial genomes may be dissected into over 2,000,000 tokens, making the associated vector space extremely large. Such spaces are subject to the "curse of dimensionality". This evocative term was coined by Bellman [11], who obviously felt deeply about the topic, characterizing it as "a malediction that has plagued the scientist from earliest days". Bellman was referring to the exponential growth in computational complexity that accompanies a linear growth in the number of dimensions to a problem. The meaning of the term has changed slightly over time; in current common statistical usage the term reflects the problems associated with the sparsity of data in multiple dimensions [12]. Sparse data spaces are difficult, and computationally expensive, to search.

To overcome this problem, a range of techniques for dimensionality reduction have been developed. One such approach is known as latent-variable models [2]. In latent-variable models dimensionality reduction is achieved by representing a high-dimensional data by a

few hidden or latent variables, which are not directly observed but inferred from the observed variables present in the model. In the latent-variable models we assume that a small number of hidden causes combine to give rise to the complexity of the data. Both the original data and the latent variables may be either continuous or discrete.

When both types of variable—original and latent—are discrete, an appropriate method to apply is latent class analysis (LCA) [13]. LCA is a statistical method, which aims to find groups of related items (latent classes or components) from observed multivariate data. Each latent variable represents a class that is characterised by a conditional probability distribution describing the chance that each observed variable is in that class. LCA can be regarded as a form of cluster analysis, as it finds a small number of latent classes that explain the observed data. However, cluster analysis assigns each item to a single cluster, whereas, in LCA one item can be allocated to multiple groups, and is associated with corresponding conditional probabilities [3].

An example of a latent-class model is the latent topic model, widely used in information retrieval [14]. This algorithm discovers the abstract topics present in a collection of documents. The underlying assumption is that each document can be represented at a topic level instead of a word level. Each document is assigned to each topic with a different weight, which reflects the extent of its membership to the topic, including the coordinates of that document in a reduced latent semantic space. Topic modelling in essence integrates soft clustering with dimension reduction [10].

One version of latent-class models is latent semantic analysis (LSA). The LSA algorithm was designed for the fast retrieval of information in documents. The goal of the LSA is to represent the semantic relationships between words and documents by projecting documents into a semantic space [3]. This projection permits the analysis of documents at a conceptual level. This transformation also helps in handling occurrences of polysemy—the existence of many possible meanings for a word or phrase—and synonyms, which are mapped to similar locations in the semantic space.

The algorithm is based on the construction of a co-occurrence matrix for the vectors. A singular value decomposition (SVD) of the co-occurrence data is then performed, factorising the co-occurrence matrix into a product of matrices (Eqn 1).

$$N = U\Sigma V^t \tag{1}$$

where U and V are matrices with orthonormal columns and Σ is a diagonal matrix.

Data dimensionality is then reduced by mapping BOW vectors, which are restricted to be linear in the lower dimensional representation, onto a latent semantic space. A latent sematic space is a mathematical representation of a large body of text, in which every term, every text, and every novel combination of terms has a high dimensional vector representation.

LSA has been used to analyse genomic patterns in *E. coli* and *P. marinus*. Each token was a short nucleotide fragment. It was found that *E. coli* genomes rarely share identical components. In contrast, in *P. marinus* most of the tokens are shared between strains. Based on the functional analysis of each gene region present in different components, it was concluded that the LSA analysis is capable of characterizing a set of core and dispensable genes in these two bacteria [15].

Probabilistic Latent Semantic Indexing (PLSA) is an extention of LSA [3]. PLSA is based on a mixture decomposition derived from a latent class model. The main objective of the algorithm, as in LSA, is to represent high-dimensional co-occurrence information in a lower-dimensional way in order to discover the hidden semantic structure of the data using a probabilistic framework. The algorithm relates a set of observed multivariate data (words and

documents) to a set of latent variables (represented by topics or components). The introduction of latent variables acts as a bottleneck and results in dimensionality reduction (Figure 1).

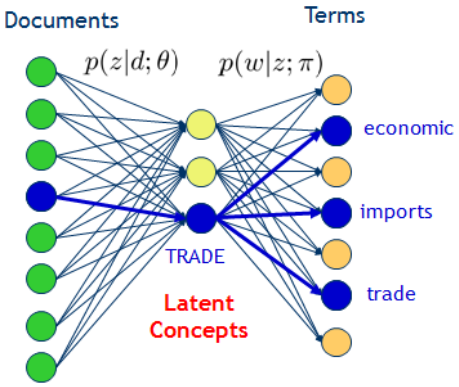

Fig. 1: Schematic representation of PLSA general structure. It shows the latent topics that links documents and words into common topics with the associated probabilities. Source: [3].

The second representation of the PLSA algorithm is related to a matrix decomposition. So the algorithm can be viewed as a factorisation of the sparse co-occurrence matrix with the aim to reduce its dimensionality. This can be achieved by approximation of a co-occurrence matrix into a product of low-rank matrices:

$$A = U\Sigma V^t \quad (2)$$

Where U is the conditional probability of a document given a component $P(d/z)$. Σ is the diagonal matrix of the prior probabilities of the topics, $P(z)$, and V is $P(w/z)$. PLSA is sometimes compared to a Non-negative matrix factorization (NMF). It has been shown that PLSA is a NMF with Kullback-Leibler (KL) divergence [16]. These authors state that any local maximum likelihood solution of PLSA is a solution of a NMF with KL divergence.

PLSA has been applied to a range of problems, including text indexing, retrieval and clustering [17-19]; image analysis [20-22]; and the analysis of mass spectroscopy data [23]. In the field of bioinformatics, PLSA has been applied to problems such as the prediction of protein subcellular localization [24] and the joint analysis of multiple metagenomic samples [25]. The latter study aimed to find hidden components shared across metagenomic datasets, in order to infer associations between microbial sequences and phenotypes. The tokens used in this study were of length 4. Some interesting, significant correlations were found between phenotypes and the distribution of phenotypes.

1.2 *Staphylococcus aureus* as a Pathogen

Staphylococcus aureus is a common human commensal, with up to 30% of people carrying the bacterium in their nasal passages [26]. It is usually harmless, but when it moves into sites such as the lungs or bloodstream it can cause a range of pathologies. *S. aureus* is the most frequently occurring bacterium in clinical isolates from hospital inpatients [27]. Many strains are resistant to widely-used drugs, and the steady rise in the incidence of *S. aureus* bacteraemia due to the increased frequency of invasive procedures, increased numbers of immunocompromised patients, and increased resistance of *S. aureus* strains to available antibiotics, including methicillin, [27] has caused widespread concern. *S. aureus* infection can lead to a wide range of pathologies, including: skin infections (e.g., folliculitis, furuncles, impetigo, wound infections, scalded skin syndrome); soft-tissue infections (e.g., pyomyositis, septic bursitis, septic arthritis); toxic shock syndrome; purpura fulminans; endocarditis; osteomyelitis; pneumonia; food poisoning; infections related to prosthetic devices (e.g.,

prosthetic joints and heart valves; vascular shunts, grafts, catheters); and urinary tract infection. The development of algorithms to identify the potential clinical phenotype of the bacteria from their genome sequence would facilitate rapid diagnosis and guide treatment decisions.

We applied the PLSA approach to *S. aureus*, using tokens derived from amino acid sequences rather than DNA. We characterised genome-scale amino acid sequences in terms of their components, and then investigated the relationships between genomes and tokens and the phenotypes they generated. As a control we used the non-pathogenic model Gram-positive bacterium *Bacillus subtilis*.

2 Methods

2.1 Dataset

Perhaps the most clinically important phenotypic trait in *S. aureus* is resistance to the antibiotic methycillin, which is often used when other antibiotics fail. Methycillin resistant *S. aureus* (MRSA) carry a gene called *mecA*, which also confers resistance to penicillin and related antibiotics [28]. Ten *S. aureus* strains were downloaded from GenBank[†]. Five of the genomes are known to contain *mecA*, and five do not (Table 1).

Tab. 1: Bacterial strains used in the analysis.

Strain	***mecA* status**
S. aureus N315	+
S. aureus Mu50	+
S. aureus MW2	+
S. aureus COL	+
S. aureus TW20	+
S. aureus Newman	-
S. aureus VC40	-
S. aureus JH1	-
S. aureus RF122	-
S. aureus NCTC 8325	-

2.2 Model Construction

Each genome was read and translated into all six reading frames. Each genome was then converted into a set of tokens of length five, using the sliding window approach described above. This window length was selected because the size of the possible search space increases exponentially with window length, making computational analysis prohibitively lengthy. A token length of five is the smallest at which "forbidden penta-peptides", which are known to be absent in the proteome, can appear in the analysis [29].

[†] http://www.ncbi.nlm.nih.gov/genbank/

In the latent class model, genomes can be represented as a collection of samples $D = d_1, d_2, \ldots, d_n$ with tokens from a vocabulary $W = w_1, w_2 \ldots, w_n$. Ignoring the sequence order in which tokens occur in samples, we can represent the data as a $N*M$ sample token co-occurrence matrix M, where the element M_{ij} indicates the number of occurrences of a particular token w_j in a sample d_i . This representation of samples is called a vector-space model, where each column of a matrix M represents a vector of a particular sample (Figure 2).

$$M = \begin{array}{c} \\ w_1 \\ w_2 \\ \ldots \\ w_n \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \ldots & d_n \end{array} \\ \begin{pmatrix} M_{11} & M_{12} & \ldots & M_{1n} \\ M_{21} & M_{22} & \ldots & M_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ M_{n1} & M_{n2} & \ldots & M_{nn} \end{pmatrix} \end{array}$$

Fig. 2: Representation of a co-occurrence matrix M, each column represents a vector of a particular sample and each row represents a vector of a particular token.

Formally, for a given value of n, the number of occurrences of all tokens is normalised across samples in order to derive sample-specific relative occurrences. We denote by x_{in} the relative abundance of token n in sample i, and by y_i the phenotype of sample i.

We used a nonparametric Mann-Whitney test to compare the relative abundance of each token in the different phenotypic groups.

2.2.1 Probabilistic Latent Semantic Analysis

The PLSA algorithm is based on a statistical model, called the aspect model [3]. The aspect model is a latent variable model for co-occurrence data, such as documents and terms, that associates an unobserved class variable or component z_k, where k is the number of components in each observation, where an observation is occurrence of a particular token. The aspect model has two forms, namely symmetric and asymmetric forms. We use the symmetric form, as it models samples and tokens in a symmetric manner (Figure 3).

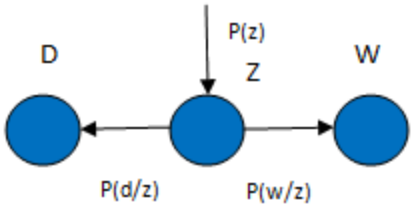

Fig. 3: Graphical representation of the aspect model in a symmetric form. D represents samples, W tokens and Z components.

The algorithm uses the following probabilities: $P(z_k)$ is the probability that a token will be observed in a particular component; $P(w_j \mid z_k)$ is the class-conditional probability of a certain token conditioned on the unobserved component z; and $P(d_i \mid z_k)$ is the conditional probability of a particular sample given the distribution of components, giving a component-specific probability distribution over the samples.

PLSA is therefore a generative model with the following steps:

- Select a component z with probability $P(z)$
- Choose a sample d with probability $P(d \mid z)$
- Generate a token w with probability $P(w \mid z)$

The resulting outcome is an observation pair (*d*,*w*), which can be considered to be a joint probability *P(d | w)* between a token and a sample:

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k) P(d_i \mid z_k) P(w_j \mid z_k) \quad (3)$$

The algorithm was applied to the ten *S. aureus* genomes listed in Table 1.

3 Results

The ten *S. aureus* genomes were used to generate tokens, as described above. A total of 2,035,253 million unique tokens were identified (Table 2), of which 12,319 tokens are different between methicillin resistant and methicillin susceptible phenotypic groups.

Tab. 2: List of mecA gene tokens present and absent in *S. aureus* strains. A total number of tokens of length 5 is specified for each strain.

Strain	Reference	*mecA* status	Number of tokens
N315	[30]	+	1,730,352
Mu50	[31]	+	1,752,841
MW2	[32]	+	1,737,534
COL	[33]	+	1,732,918
TW20		+	1,788,387
Newman	[34]	-	1,739,876
VC40	[35]	-	1,704,891
JH1	[36]	-	1,556,066
RF122		-	1,718,817
NCTC 8325	[37]	-	1,740,199

The distribution of tokens for strain N315 is shown in Figure 4.

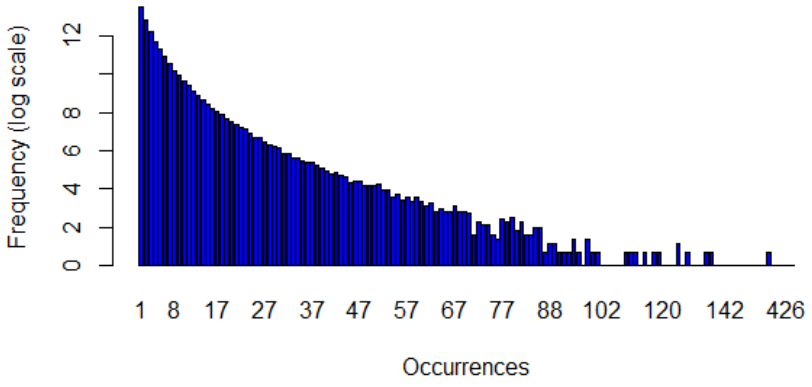

Fig. 4: Distribution of token occurrences in *S. aureus* strain N315. All strains had a similar token distribution.

Of the 2,035,253 unique tokens identified, 12,319 were found to be unique to one of the two phenotypic groups: methycillin resistant or non-resistant (Mann-Whitney U test, $p < 0.05$). Although some of these tokens will undoubtedly be derived from the *mecA* gene, which we know to be present in the resistant group and absent in the sensitive group, it is highly likely that there are other systematic differences between the two groups, of which we know little or nothing. From a practical point of view, we are interested in tokens which are highly correlated with the phenotype. We therefore assume that there exists a group of common bacterial components and that each bacterium is a mixture of these components. In order to identify the components a PLSA model was derived, using an Expectation-Maximisation (EM) algorithm. The EM algorithm converges to a local maximum of its fitness value, and must therefore be run multiple times from random starting points in order to more fully explore the solution space. Three components were identified (Figure 5).

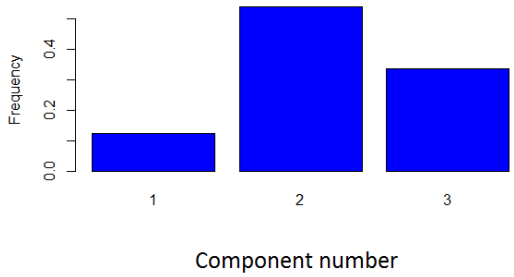

Fig. 5: *P(z)* matrix generated by the PLSA algorithm for the ten *S. aureus* strains

The distribution of tokens was not statistically significantly different across the three components, indicating that, as expected, presence or absence of *mecA* is not the only difference between the resistant and sensitive strains (Table 3).

Tab. 3: The five most highly-ranked tokens in each component

Component 1		Component 2		Component 3	
Token	$P(w/z_1)$	Token	$P(w/z_1)$	Token	$P(w/z_1)$
SDSDS	6.14E-5	SDSDS	8.23E-5	SDSDS	8.08E-5
DSDSD	5.63E-5	DSDSD	8.06E-5	DSDSD	7.47E-5
LLLLL	5.47E-5	LLLLL	6.21E-5	LLLLL	3.47E-5
LILLL	2.93E-5	SESLS	3.19E-5	LFLLL	3.15E-5

However, differences between strains were apparent: it appears that strains N315, Mu50 and JH1 are more closely related to each other than to the rest of the strains. In addition, strain RF122 has a very different token distribution from the other strains (Figure 6).

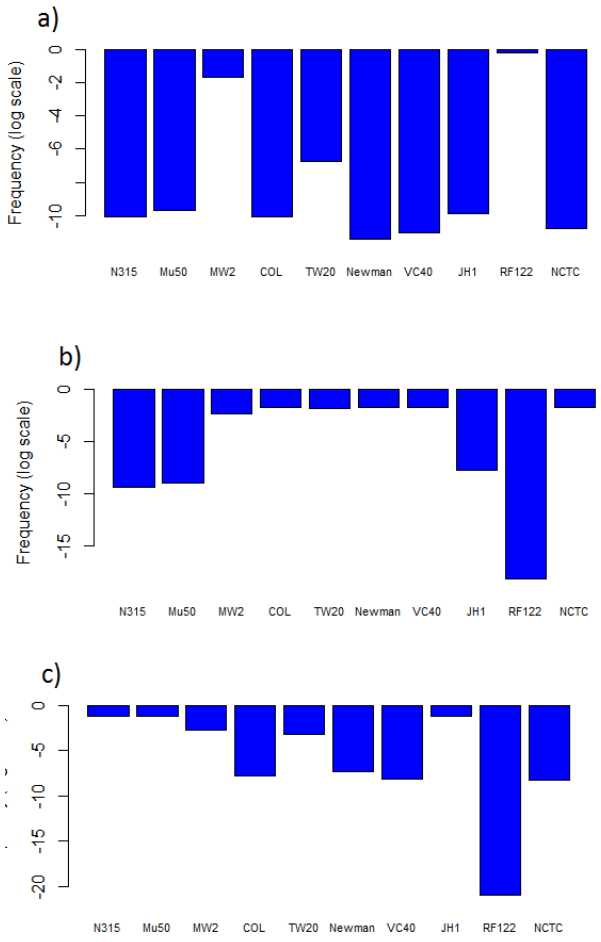

Fig. 6: Distribution of tokens in a) Component 1, b) Component 2, and c) Component 3. The first five strains are methycillin resistant, and the remainder are sensitive

Of the 2,035,253 obtained, only 9,613 are significant at the 5% level, compared with 12,319 identified as significant using the co-occurrence approach. It appears, then, that 2,706 of the tokens are false positives, and not related to the methycillin resistance phenotype.

The algorithm was then applied to the comparison of tokens derived from the methycillin-resistant *S. aureus* strains with those from five strains of *Bacillus subtilis*: 168; 6051 HGW; BSn5; QB928 and XF 1. Regression analysis indicates that the distributions of tokens in components 1 and 3 are significantly different at the 5% level (Figure 7), so that our approach can effectively distinguish between MRSA and *B. subtilis*.

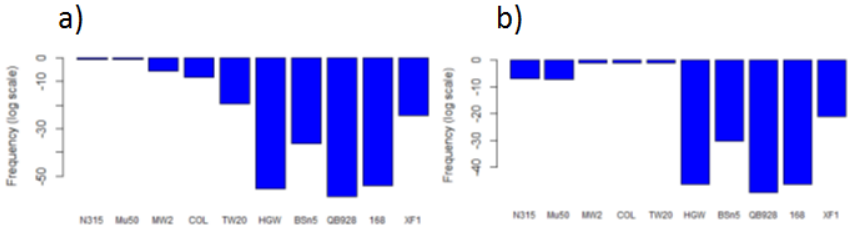

Fig. 7: Distribution of tokens from MRSA and *B. subtilis* strains in a) Component 1 and b) Component 3. Strains N315, Mu50, MW2, COL and TW20 are MRSA strains, while the remainder are *B. subtilis*

4 Discussion

The rapid identification of bacterial pathogens, and their antibiotic resistance status, is important in order to treat patents effectively, and to reduce the use of ineffective drugs. Although the mechanism of drug resistance in bacteria has been an active area of research for many decades, our understanding of the genes and proteins involved in the thousands of bacteria responding to hundreds of antibiotics is incomplete. The information that we do have is largely contained in the literature, couched in abstruse technical jargon, and not easily available for computational analysis.

An increasingly attractive alternative approach is to use the rapidly-growing body of genome data being generated by recently-developed high throughput Next Generation Sequencing technologies. Although much of this data lacks annotation, it is directly linked to clinical phenotype, and then can be used in a "black box" manner, with important differences learned by computational algorithms. Once trained, such algorithms can be used to rapidly identify organisms and their drug resistance status. The differences identified can also be used to direct further laboratory investigations, by generating testable hypotheses about the areas of the genome which differ between resistant and sensitive strains.

We applied an algorithm developed for the analysis of large text corpuses, latent sematic indexing, to the analysis of differences between the genomes of methycillin resistant and methycillin sensitive *Staphylococcus aureus*. Rather than apply the algorithm directly to the DNA sequence, with its limited four-letter alphabet, we computationally translated the genomes in all six reading frames, producing a much larger search space from the 20-letter amino acid alphabet. Using a sliding window approach we generated all possible tokens of length *n* from each genome.

The PLSA algorithm generated three components. Although there were visible differences between the distribution of tokens in the components according to methycillin resistance status, these differences are not statistically significant. This finding is not entirely surprising, since we based our phenotypic groupings on the presence or absence of a single gene, *mecA*, known to be important in methycillin resistance. Although *mecA* is a major player, there are undoubtedly many other factors affecting methycillin resistance, including other genes, environmental factors, and possibly issues such as post-translational modifications to amino acid strings. Application of PLSA to tokens derived from a different bacterial species, *Bacillus subtilis*, did identify significant differences, indicating that the approach is valid, given the appropriate identification of indicators for the clinical phenotype. The information already present in the literature could be used to further refine this identification. Existing information could also be integrated into the analysis by incorporating annotations from a database such as the Gene Ontology [38], based upon the chromosomal location of a token.

The amount of genome sequence data available in public databases is very large, and increasing exponentially. Although data is deposited faster than it can be annotated, it can still be used for applications including clinical assessment, using approaches such as latent semantic analysis.

References

[1] B. Spellberg, R. Guidos, D. Gilbert, J. Bradley, H. W. Boucher, W. M. Scheld, J. G. Bartlett, and J. Edwards. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. *Clinical Infectious Diseases*, 46:155-164, 2008.

[2] M. A. Carreira-Perpinan. A review of dimension reduction techniques. University of Sheffield, Department of Computer Science1997.

[3] T. Hofmann. Probabilistic latent semantic analysis. presented at the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[4] L. Poirel, R. A. Bonnin, and P. Nordmann. Analysis of the resistome of a multidrug-resistant NDM-1-producing *Escherichia coli* strain by high-throughput genome sequencing. *Antimicrobial Agents and Chemotherapy,* 55:4224-4229, 2011.

[5] M. D. Adams, E. R. Chan, N. D. Molyneaux, and R. A. Bonomo. Genomewide analysis of divergence of antibiotic resistance determinants in closely related isolates of *Acinetobacter baumannii*. *Antimicrobial Agents and Chemotherapy,* 54:3569-3577, 2010.

[6] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development,* 15:589 - 594, 2005.

[7] A. M. Lesk. *Introduction to Bioinformatics*. New York: Oxford University Press, 2008.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*: Springer, 2007.

[9] S. Sauer and M. Kliem. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* 8:74-82, 2010.

[10] Crain, S., K. Zhou, S.-H. Yang, and H. Zha. Dimensionality reduction and topic modelling: From latent semantic indexing to latent Dirichlet allocation and beyond. in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds., ed: Springer US, 2012, pp. 129 - 161.

[11] R. E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey: Princeton University Press, 1961.

[12] D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. New York.: John Wiley & Sons, Inc,, 1992.

[13] P. F. Lazarsfeld and N. W. Henry. *Latent structure analysis*. Boston: Houghton Mifflin,, 1968.

[14] M. Steyvers and T. Griffiths. Probabilistic topic models. in *Handbook of Latent Semantic Analysis*, ed: Psychology Press, 2007.

[15] T. Y. Lim, X. Hu, C. Xin, S. Xiajiong, E. K. Park, and G. L. Rosen. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 9:980 - 991, 2012.

[16] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. presented at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil. , 2005.

[17] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. in *17th International Conference on Machine Learning*, Stanford University, Stanford, CA, USA, 2000, pp. 167 - 174.

[18] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. in *NIPS '00: Advances in Neural Information Processing Systems*, 2000.

[19] T. Brants, F. Chen, and I. Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis. presented at the Eleventh International Conference on Information and Knowledge Management, CIKM '02,, New York, NY, 2002.

[20] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. 370 - 377, 2005.

[21] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. presented at the 12th Annual ACM International Conference on Multimedia, Multimedia '04, New York, NY, 2004.

[22] J. Li, Gong, S. and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. presented at the British Machine Vision Conference, 2008.

[23] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical Chemistry,* 80:9649 - 9658, 2008.

[24] J.-M. Chang, E. Chia-Yu Su, A. Lo, H.-S. Chiu, T.-Y. Sung, and W.-L. Hsu. Psldoc: Protein subcellular localization prediction based on gapped dipeptides and probabilistic latent semantic analysis. *Proteins: Structure, Function, and Bioinformatics,* 72:693710, 2008.

[25] Y. Baran and E. Halperin. Joint analysis of multiple metagenomic samples. *PLoS Computational Biology,* 8:2012.

[26] C. D. den Heijer, E. M. van Bejnen, W. J. Paget, M. Pringle, H. Goossens, C. A. Bruggeman, F. G. Schellevis, E. E. Stobberingh, and A. S. Team. Prevalence and resistance of commensal *Staphylococcus aureus*, including methicillin-resistant *S. aureus,* in nine European countries: A cross-sectional study. *Lancet Infectious Diseases,* 13:1011, 2013.

[27] C. K. Naber. *Staphylococcus aureus* bacteremia: Epidemiology, pathophysiology, and management strategies. *Clinical Infectious Diseases,* 48:S231-S237, 2009.

[28] K. Ubukata, R. Nonoguchi, M. Matsuhashi, and M. Konno. Expression and inducibility in *Staphylococcus aureus* of the meca gene, which encodes a methicillin resistant *S. aureus*-specific penicillin-binding protein. *Journal of Bacteriology,* 171:2882 - 2885, 1989.

[29] T. Tuller, B. Chor, and N. Nelson. Forbidden penta-peptides. *Protein Science,* 16:2251 - 2259, 2007.

[30] T. Ito, Y. Katayama, and K. Hiramatsu. Cloning and nucleotide sequence determination of the entire mec DNA of pre-methicillin-resistant *Staphylococcus aureus* N315. *Antimicrobial Agents and Chemotherapy,* 43:1449-1458, 1999.

[31] K. Hiramatsu, L. Cui, M. Kuroda, and T. Ito. The emergence and evolution of methicillin-resistant *Staphylococcus aureus. Trends in Microbiology,* 9:486-493, 2001.

[32] L.-c. C. WU. DNA sequences and primers for identifying methicillin-resistent *Staphylococcus aureus* MW2 and USA300 strains. 2010.

[33] S. Ravipaty and J. P. Reilly. Comprehensive characterization of methicillin-resistant *Staphylococcus aureus* subsp. *aureus* COL secretome by two-dimensional liquid chromatography and mass spectrometry. *Molecular & Cellular Proteomics,* 9:1898-1919, 2010.

[34] T. Baba, T. Bae, O. Schneewind, F. Takeuchi, and K. Hiramatsu. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *Journal of Bacteriology,* 190:300-310, 2008.

[35] R. Rosenstein and F. Götz. What distinguishes highly pathogenic staphylococci from medium-and non-pathogenic? in *Between Pathogenicity and Commensalism*, ed: Springer, 2013, pp. 33-89.

[36] S. Matyi, J. Dupre, W. Johnson, P. Hoyt, D. White, T. Brody, W. Odenwald, and J. Gustafson. Isolation and characterization of *Staphylococcus aureus* strains from a Paso del Norte dairy. *Journal of Dairy Science,* 96:3535-3542, 2013.

[37] J. Štěpán, R. Pantůček, V. Růžičková, S. Rosypal, V. Hájek, and J. Doškař. Identification of *Staphylococcus aureus* based on PCR amplification of species specific genomic 826 bp sequence derived from a common 44-kb *Sma I* restriction fragment. *Molecular and Cellular Probes,* 15:249-257, 2001.

[38] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.* Gene Ontology: Tool for the unification of biology. *Nature Genetics,* 25:25 - 29, 2000.