Chapter 23

Data Mining to Discover Emerging Patterns of Antimicrobic Resistance

J. A. Poupard, R. C. Gagnon, and M. J. Stanhope
GlaxoSmithKline, Collegeville, PA, USA

1. INTRODUCTION

Antimicrobial susceptibility testing results are typically presented as summary information in the form of percent susceptible-intermediate-resistant (SIR), as the minimum concentrations required to inhibit 50% or 90% of isolates ($MIC_{50/90}$s) or as MIC frequency distributions. However, extracting additional information from large databases, involving thousands of isolates tested against more than 20 antimicrobial agents containing 6–10 individual dilutions, involve data points numbering in the millions. In such cases, traditional methods of analysis are insufficient. One approach to dealing with these levels of complexity is by applying novel data mining procedures. Although it should be noted that there is no universal definition of the term "data mining," for the purposes of this chapter it is defined as:

A new discipline lying at the interface of statistics, database technology, pattern recognition and machine learning, and concerned with secondary analysis of large databases in order to find previously unsuspected relationships, which are of interest or value to their owners. (Hand, American Statistician, 1998 [Hand, 1998]).

Due to the complex nature of data mining in this context, a team effort is required involving experts from various fields, including specialists in computer/bioinformatics. Regardless of the number of disciplines involved, it is critical that someone specializing in microbiology or infectious diseases is

Antibiotic Policies: Theory and Practice. Edited by Gould and van der Meer
Kluwer Academic / Plenum Publishers, New York, 2005

included. This person should have full knowledge of the limitations, in design and execution, of the susceptibility testing procedures used to generate the data. In this capacity, the microbiologist evaluates the data generated and searches out inconsistencies in those data as well as in the quality control procedures associated with the primary database. Any inconsistencies that are revealed need to be pursued and resolved in order to assure validation of the primary database. Failure to do this thoroughly will undermine any further analysis conducted. As data mining of large multinational resistance databases is relatively novel, the complexity of these procedures is just now becoming apparent.

The goal of this chapter is to focus attention on methods for identifying new or unrecognized resistance patterns in large surveillance study databases. Application of these data for use in resistance modeling and infection control is addressed elsewhere in this book. Although methods applicable to these large databases certainly apply to information generated in individual hospital laboratories, the large surveillance databases pose a special problem because of the volume of data involved.

2. BRIEF LITERATURE REVIEW

Although it may not have been called data mining, the extraction of information from large databases has been a hallmark of epidemiology studies for a long period of time (Kaslow and Moser, 2000). In the 1980s investigators started to apply data mining techniques in attempts to combine antimicrobial susceptibility testing information obtained from the clinical microbiology laboratory with hospital/medical center information systems to help identify specific hospital infection control problems. These studies attempted to characterize and rapidly identify outbreaks of infection, particularly in locations such as intensive care units. A series of papers associated with investigators at the University of Alabama focused on many of these and related issues (Brossette *et al*., 1998, 2000; Moser *et al*., 1999). In a 1998 paper, Brossette *et al*. described a concept they called data mining surveillance system (DMSS) to encourage the application of rules of association in identifying new patterns within large infection control and public health surveillance data. The DMSS was further refined in subsequent papers through its application to intensive care units and for infection control surveillance. More recently, Peterson and Brossette introduced the concept of virtual surveillance to encourage the application of data mining techniques on an ongoing basis (Peterson and Brossette, 2002).

Two significant data mining studies on antibiotic use and drug resistance in a hospital setting have been conducted (Lopez-Lazano *et al*., 2000; Monnet *et al*., 2001). These studies combined antimicrobial susceptibility testing data

with information on antimicrobial use obtained from the hospital pharmacy to conduct time-series analysis employing data mining techniques based on the ARIMA (autoregressive integrated moving average) model proposed by Box and Jenkins (Box and Jenkins, 1976). They succeeded in demonstrating a temporal relationship between the use of two antibiotics in the hospital and the percentage of Gram-negative bacilli resistant or intermediate to those antibiotics. In somewhat similar studies, Brown *et al*. evaluated the use of binary cumulative sums (CUSUMs) and moving average (MA) control charts to identify clusters of nosocomial infections using changes in antimicrobial resistance of isolates (Brown *et al*., 2002). In 2002 Poupard *et al*. summarized three methods for data mining of large multinational surveillance databases (Poupard *et al*., 2002). It is becoming apparent that as hospital and third party payer databases expand, the use of novel applications of data mining from other fields will become increasingly applied to resistance surveillance databases.

3. PRIMARY RESISTANCE SURVEILLANCE DATA

When planning surveillance studies, the basic database for analysis will need to contain line listings of MICs for the individual isolates. This is important, not only because breakpoints (SIR) change over time, but because the investigator may want to apply unique breakpoints that differ from the standard breakpoints; for example, in order to collect information on possible first-step mutations prior to an organism becoming resistant.

It should also be noted that as much patient demographic information and specific isolate information as possible is always preferred, regardless of the original intent of the planners. These kinds of information are often invaluable when resolving issues of unusual or interesting results from data mining procedures.

4. THREE APPROACHES TO RESISTANCE INFORMATION DATA MINING

Three main approaches will be discussed for data mining of large databases containing drug susceptibility/resistance information to search for novel information or patterns of antimicrobial resistance: (1) the antibiotype method, (2) multivariate analysis, and (3) evolutionary genetics approaches. These methods were previously presented in a summary paper analyzing information

generated by the Alexander Project, a 10-year multinational surveillance study of upper respiratory isolates (Poupard *et al*., 2002).

4.1. The antibiotype approach

This is a method that first converts the long string of MICs to any number of drugs into a series of 0s and 1s with 0s representing a susceptible result and 1s representing non-susceptible or resistant results; the basic antibiotype. It should be noted that the 1 can be based on selected, published breakpoints, or an artificial breakpoint such as an MIC result that the investigator determines to be a first-step mutation. The string of S (susceptible) and R (non-susceptible) results of an isolate tested against individual drugs is converted into a string of 0s and 1s. The use of the binary code is adaptable to all computer programs and enables the investigator to search the database for novel patterns. A string of all 0s indicates an isolate is susceptible (based on the chosen breakpoints) to all drugs tested which is often lost when one prepares summary tables based on percent SIR or on $MIC_{50/90}$s.

In order to perform more sophisticated analysis, the long string of numbers can be converted into a two- or three-digit number. This is done by grouping the string of 0s and 1s into subsets of three numbers. Each consecutive number in the set of three is assigned a value of 1, 2, or 4, with 0s remaining 0s. For example, 100 = 100; 010 = 020; 001 = 004; 111 = 124, etc. Once converted in this way, each set of three numbers can then be transformed to a single number derived from the sum of the values for the non-susceptible results, so: 100 = 100 = 1; 010 = 020 = 2; 001 = 004 = 4; 111 = 124 = 7, etc. It is also possible to assign a two- or three-digit hyphenated code based on the number of non-susceptible results (1s) in the binary string. Each unique string with the same number of non-susceptible results would give a new second digit, and the process would be continued until each antibiotype has a unique number designation.

Use of this method permits the determination of the predominant antibiotype, as well as rare antibiotypes, and enables the evolution of these antibiotypes to be tracked over time or by designated locations. It also permits analysis of variability in a population of unique antibiotypes over time and can show the rise or decline in the all zero antibiotype within the population. Specific applications of this methodology to isolates of *Streptococcus pneumoniae* and *Haemophilus influenzae* can be found in the previously cited paper by Poupard *et al*. (2002).

4.2. Multivariate analysis methods

Multivariate projection methods are applicable for obtaining a broad overview of large, complex (multidimensional) data. Projection methods are

powerful tools for discovering patterns in such data and have been applied in many situations, for example, genetics (Nguyen and Rocke, 2002), cheminformatics (Janne *et al*., 2001), and statistical process control (MacGregor and Kourti, 1995), among others. In this chapter, we describe their application to multinational surveillance data. These methods are not meant to supersede the traditional univariate approach to analysis of surveillance data. Rather, they are meant to enhance the univariate analysis and to enable a greater understanding of the underlying patterns of variability among the antibiotics, countries, dates of collection, isolate sources, etc. In addition to this greater understanding, identification of interesting isolates, which may have escaped detection using the univariate approach, is likely. Thus, multivariate processes extend the level of understanding of the data beyond that obtained from the standard univariate approaches and provide a framework for additional analysis.

We applied multivariate analysis to the 1998 Alexander Project collection of 8,952 *S. pneumoniae* isolates from 24 countries. Isolates were tested against 20 antibiotics and data were available for age, gender, and isolate source (Table 1). While it is possible, and generally of interest, to apply multivariate techniques to the entire dataset, such an application may give rise to results which are artifacts of the sampling, rather than reflecting true patterns of resistance. For example, preliminary examination showed that data were collected from any of six sources (throat, ear, sputum, blood, nasopharynx, sinus) and from patients of any age. However, these data were not well distributed across the countries of origin. Some countries had no ear isolates (e.g., Austria, France, Germany) whereas others had many (e.g., United States). Similarly, some countries have very few isolates from patients of 5 years of age or less

Table 1. Antibiotics (abbreviation used) analyzed from the 1998 Alexander Project

β-Lactams	Macrolides	Quinolones	Others
Penicillins Penicillin (Pen) Amoxicillin (Amx) Amoxicillin/ clavulanic acid (Aug)	Erythromycin (Ery) Clarithromycin (Cla) Azithromycin (Azi)	Ciprofloxacin (Cip) Ofloxacin (Ofl) Gemifloxacin (Gem)	Clindamycin (Cli) Chloramphenicol (Chl) Doxycycline (Dox) Co-trimoxazole (Cot)
Cephalosporins and loracarbef Cefaclor (Fac) Loracarbef (Lor) Cefuroxime (Fur) Cefixime (Fix) Cefotaxime (Tax) Ceftriaxone (Axo) Cefprozil (Cpz)			

(e.g., Switzerland, Austria) and others had many (e.g., United States, Japan). As the prevalence of resistance has been shown to vary based on isolate source and age, particularly in patients of 5 years of age or younger (Sahm *et al*., 2000; Thornsberry *et al*., 1999), there was a risk that the uneven distribution of these data would produce misleading results. However, all the countries had large numbers of sputum isolates from patients older than 5 years of age and the analysis was therefore restricted to this subset, including 1,295 bacterial isolates.

For the 1998-subset data, 20 antibiotics were tested representing 20 dimensions (variables) against 1,295 bacterial isolates (observations). It is, of course, not possible to visually examine 20-dimensional data. The concept of multivariate projection is that high dimensional data are transformed into a lower dimensional space, allowing data to be examined visually while at the same time explaining the variation in the data. The percentage of the total information in the data that is represented by the lower dimensional space is determined by R^2, analogous to the familiar R^2 value from simple linear regression. Distributions of MIC data are generally not symmetric. For most modeling procedures, whether univariate or multivariate, transformation of the MIC data to achieve a close-to-symmetric distribution is common. For example, MIC distributions are sometimes summarized using the geometric mean, which is based on log MIC, a close-to-symmetric transformation of the MIC data. Such transformations are essential for data modeling, mainly to make the models more efficient (reliable) and to remove undue influence on the model from relatively few extreme values, in this case extreme MICs.

For our analysis, MIC data were transformed to achieve a close-to-symmetric distribution using a log transformation. Data were first modeled using principal components analysis (PCA). The mathematical complexities of PCA are discussed in many statistical texts; see, for example, Morrison (1990), or Eriksson *et al*. (2001). PCA was carried out on the log MICs using SIMCA (2000) software. The principal components were then summarized graphically using score and loading plots. In their lower dimensional space, score plots describe the coordinates of the observations (isolates) while loading plots describe the coordinates of the variables (antibiotics). In order to interpret the principal components, it is necessary to examine loading plots. The loading plot for each principal component describes the structure being revealed by the component. The largest component corresponds to the highest proportion of total R^2 explained and is called the principal component 1, or p[1]. In addition to plots of individual loadings, two-dimensional or three-dimensional plots of combinations of the loadings are also very informative. Variables contributing similar information, for example, those that are correlated, are grouped together in two-dimensional or three-dimensional loading plots. Variables that are negatively correlated are located on opposite sides of the plot. The further from the plot origin that a variable lies, the greater its impact on the PCA model.

For the 1998 MIC data, the 20 variables and 1,295 observations were represented by four principal components, accounting for 88% of the total information in the MICs. Thus, the 20-dimensional data are summarized and interpretable in only four dimensions. The loadings for the first four principal components for *S. pneumoniae* in 1998 are shown in Figure 1. As stated, the loadings reflect structure among the variables, in this case antibiotics.

- The first principal component, p[1], explained 61% of the information in the MIC data and described isolates with high MICs among all antibiotic classes except the quinolones (Figure 1a).
- The second component explained an additional 13% of the MIC information, separated macrolides, chloramphenicol, and doxycycline from the β-lactams and co-trimoxazole, but contained relatively little information about the quinolones (Figure 1b).

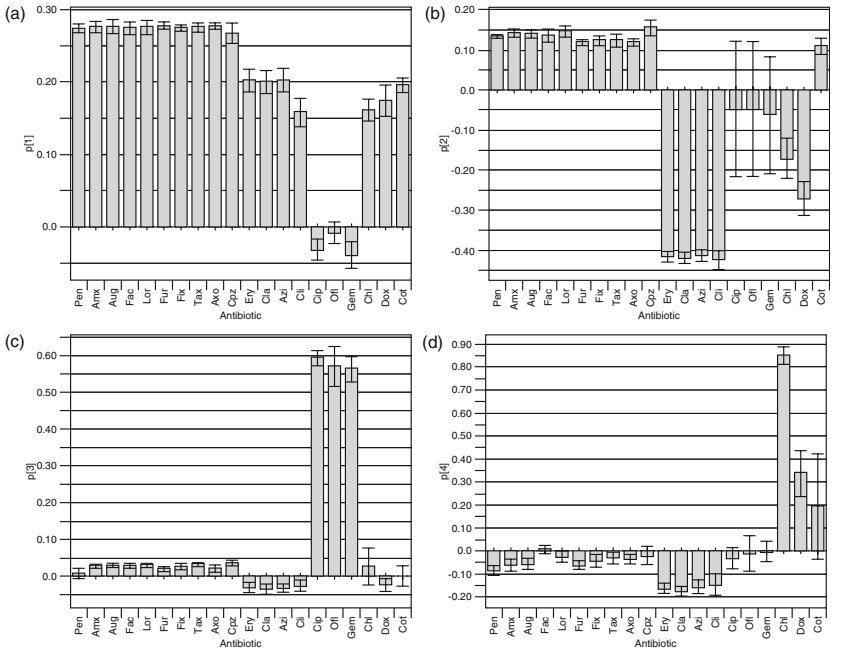

Figure 1. Loading plots for the first four principal components. (a) Antibiotic loadings for principal component 1 (p[1]); (b) antibiotic loadings for principal component 2 (p[2]); (c) antibiotic loadings for principal component 3 (p[3]); (d) antibiotic loadings for principal component 4 (p[4]).

- The third component (11% of information) uniquely described isolates with high quinolone MICs (Figure 1c).
- The fourth component (3% of information) picked up isolates with high chloramphenicol, doxycycline, and co-trimoxazole MICs, with relatively low macrolide MICs (Figure 1d).

The 1998 MIC data therefore can be described by four components, which, perhaps not surprisingly, closely follow the antibiotic classes.

Plots of two-dimensional loadings (antibiotics), combined with two-dimensional score plots (isolates), reveal unusual isolates, patterns among the isolates and relationships between the isolates and antibiotics (Figure 2). A two-dimensional plot of components p[1] vs p[2] revealed clustering by antibiotic class, with β-lactams clustered in the upper right-hand corner (Figure 2a). This cluster was due to the strong correlation among the β-lactam MICs (co-trimoxazole MICs were closely correlated with the β-lactams). The macrolides (plus clindamycin) were also clustered, but were separate from the β-lactams. Chloramphenicol and doxycycline were closely related to each other, but dissimilar from the other classes of antibiotics. The quinolones, being close to the origin, exerted no influence in the first two components. The two-dimensional score plot for the first two dimensions is shown in Figure 2b. The scores in dimension 1 are denoted t[1], and in dimension 2 are denoted t[2]. The score plot revealed distinct clusters among the isolates and was interpreted by relating the position of observations in the plot (which represent individual bacterial isolates) to the positions of variables in the p[1], p[2] loading plot. For example, the upper right quadrant of Figure 2b represents a cluster of isolates with high β-lactam and co-trimoxazole MICs (cluster 1). These isolates are associated with low macrolide MICs, and low to midrange doxycycline and chloramphenicol MICs, because these drugs are located in different regions of the loadings plot.

The median MICs from cluster 1 in Figure 2b were plotted in Figure 3a expressed as the number of dilutions from each antibiotic's respective MIC_{90}. As expected, the β-lactam median MICs were at their respective MIC_{90}s (i.e., 0 dilutions from the MIC_{90}) and macrolide, clindamycin MICs were between 9 and 11 dilutions below their MIC_{90}s. Cluster 2, at the bottom of the score plot Figure 2b, was associated with high macrolide and low β-lactam MICs; this was reflected by median MICs in that cluster, compared to the MIC_{90}s for each drug (Figure 3b). Clusters 3 and 4 were both high in the first component, the difference being the relative positions in the second component, and hence these clusters differed primarily by macrolide MICs (Figure 3c and 3d, respectively). Clearly the MICs were high for all drugs (except quinolones) in cluster 4, and in cluster 3 the increase in MICs among the macrolides was evident (compare with Figure 3a).

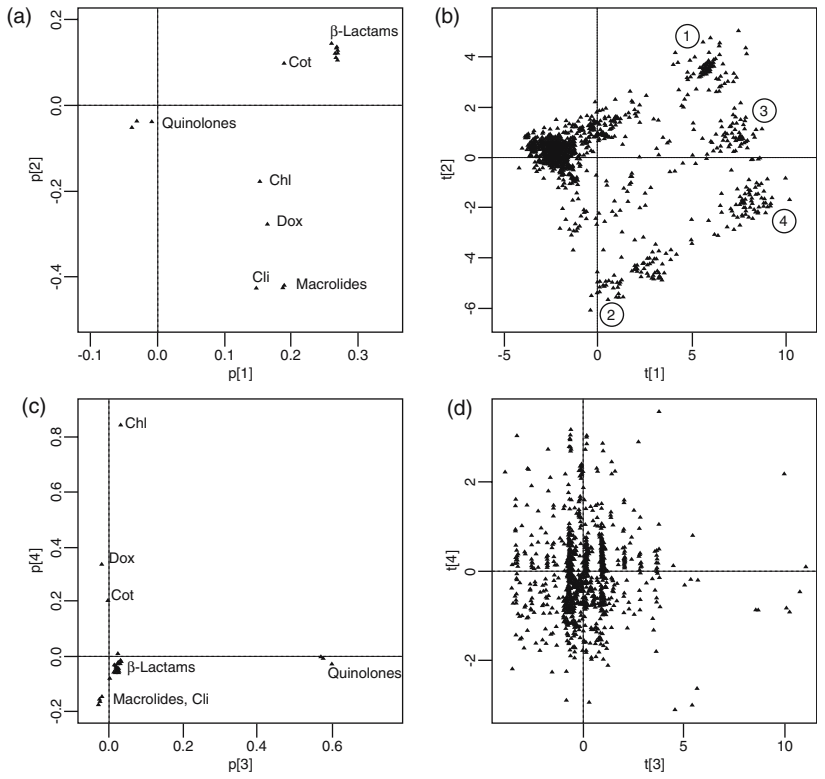

Figure 2. Two-dimensional loading and score plots for the first four principal components at the MIC level. (a) p[1] vs p[2] loading plots, showing clustering among antibiotics for the first two components; (b) t[1] vs t[2] score plots, showing major clusters among isolates for the first two components; (c) p[3] vs p[4] loading plots, showing clustering among antibiotics in the last two components; (d) t[3] vs t[4] score plots, showing major clusters among isolates in the last two components.

Figure 1c showed that the third principal component represented the quinolones and Figure 1d showed that the fourth component was dominated by chloramphenicol. Two-dimensional loading plots and corresponding two-dimensional score plots help to reveal how the isolates with high quinolone or chloramphenicol MICs cluster with the other isolates (Figure 2c and 2d). Isolates with high quinolone MICs are easily identified (on the far right of Figure 2d), and other outlying isolates, driven by high or low chloramphenicol MICs can be seen clearly. Note the vertical banding among the isolates of Figure 2d; these bands correspond to MICs of the quinolones, with the highest density bands being in the mid-MIC range.

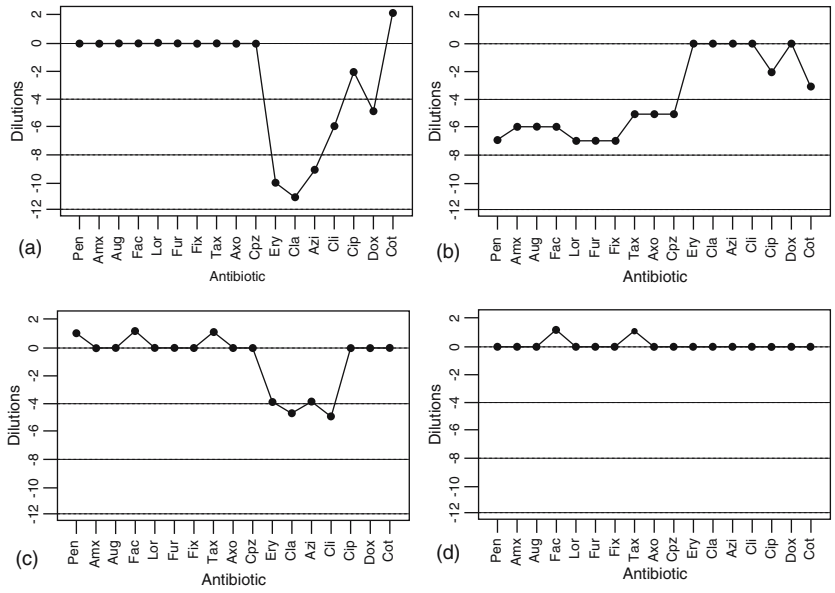

Figure 3. Median number of dilutions from each antibiotic's MIC_{90}, for major clusters in the first two principal components (p[1] and p[2]). (a) Cluster 1, median MICs for isolates in cluster 1 are at the MIC_{90} for β-lactams, but are between 9 and 11 dilutions below the MIC_{90} for the macrolides, between 3 and 5 dilutions below the MIC_{90} for clindamycin and doxycycline, and 2 dilutions below for chloramphenicol; (b) cluster 2, this is the inverse of cluster 1 with the β-lactams having low MICs relative to their MIC_{90}s, whereas macrolides and clindamycin are at their MIC_{90}s; (c) cluster 3, this shows MICs which are similar to cluster 1 except that macrolide, clindamycin MICs have increased compared with cluster 1; (d) cluster 4, this corresponds to high MICs for all non-quinolones, with MICs at the MIC_{90}s. Note that quinolones are not included in this figure as there was no information about the quinolones in the first two principal components.

The methodology described above can also be applied to the antibiotypes, and, as expected, the results were similar, though with some notable exceptions. First, the two-dimensional loadings plot for the first two components (Figure 4a) illustrates that antibiotic groupings were similar to the MIC groupings (Figure 2a), with the exception that amoxicillin and amoxicillin/clavulanic acid, at the antibiotype level, were distinct from the other β-lactams in p[1]. This distinction is due to the comparatively increased susceptibility of *S. pneumoniae* strains to amoxicillin vs other β-lactams when NCCLS breakpoints are used to determine the antibiotype (NCCLS, 2000). Figure 4b describes the distribution of isolates at the antibiotype level. The interpretation was similar to that for isolates at the MIC level (Figure 2b), although here there appeared to be a greater variety in the clustering pattern. The third and

fourth components (Figure 4c) described (1) isolates with resistance to the quinolones ciprofloxacin and ofloxacin (there was no resistance to gemifloxacin in the isolates analyzed), and (2) the isolates with resistance to amoxicillin and amoxicillin/clavulanic acid (Figure 4c). Figure 4d shows the pattern of isolates in the third and fourth components. Isolates resistant to both amoxicillin and amoxicillin/clavulanic acid but not the two quinolones were clustered at the very top of the plot (cluster 1). There were two isolates with resistance to both amoxicillin and amoxicillin/clavulanic acid and one of the two quinolones, ciprofloxacin, or ofloxacin (cluster 2). A small group of six isolates were resistant to both quinolones, as well as to amoxicillin and

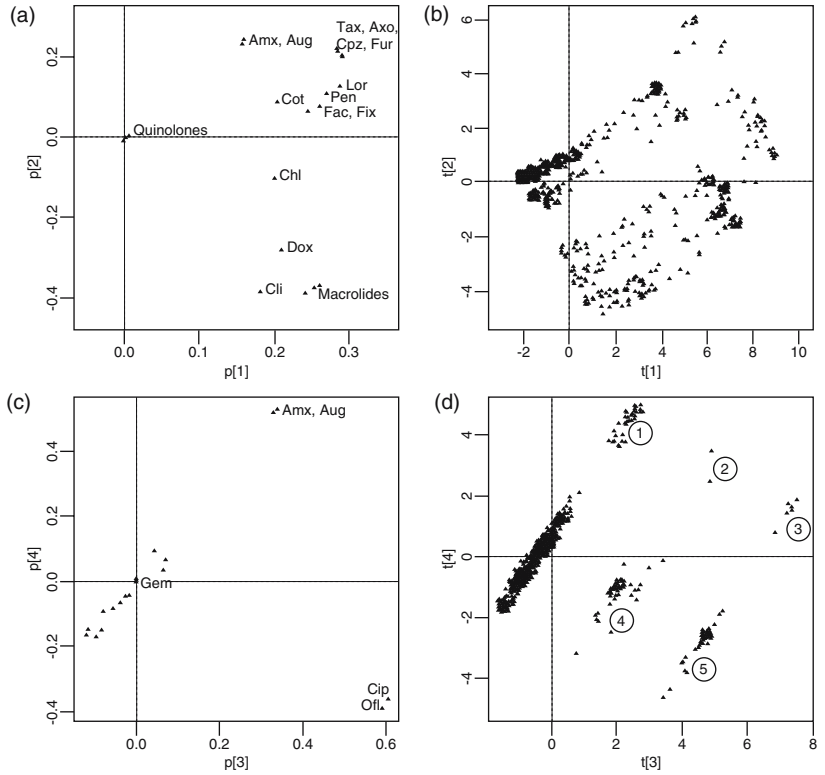

Figure 4. Two-dimensional loading and score plots for the first four principal components at the antibiotype level. (a) p[1] vs p[2] loading plots showing clustering among antibiotics in the first two components; (b) t[1] vs t[2] score plots, showing major clusters among isolates in the first two components; (c) p[3] vs p[4] loading plots, showing clustering among antibiotics in the last two components; (d) t[3] vs t[4] score plots, showing major clusters among isolates in the last two components.

amoxicillin/clavulanic acid (cluster 3). Cluster 4 included a group of isolates resistant to one of the two quinolones but not amoxicillin or amoxicillin/clavulanic acid. Finally, cluster 5 was a group of isolates resistant to both quinolones but not to amoxicillin and amoxicillin/clavulanic acid.

Other interactions among the antibiotics and isolates may be obtained by other two-dimensional and/or three-dimensional views of the components. The four components together explained over 80% of the information in the antibiotypes. In the above analysis, the country-to-country effect was not included. However, it is straightforward to include country as a variable. One way is to fit the same overall PCA model but use different colors or graph symbols in the score plots to depict the distribution of isolates for the different countries, and/or use graphical tools to display the countries individually. This works well and is a rather striking way to examine countries. Another approach is to use a related multivariate projection to model the relationship between countries and isolates; in this case the technique known as projection to latent structures (PLS) works well. The technical details of PLS have been well described (Eriksson *et al*., 2001). With PLS, countries are considered predictor variables (or X variables) and antibiotics are considered response variables (or Y variables), with the isolates as observations. Both the X and Y, which are matrices with rows as isolates and columns as variables (countries in the X matrix and antibiotics in the Y matrix), are projected into lower dimensional space similar to a PCA projection, with the projections modified slightly to maximize the correlations between the X and Y variables. PLS models are easily fitted with the SIMCA software package.

The results of our PLS modeling have shown which countries are most (or least) highly associated with the antibiotic MIC or antibiotype patterns. As an example, we consider the 1998 antibiotype data previously modeled. In PLS we examine loadings for both X and Y variables to learn about their associations. To distinguish the PLS loadings from the PCA loadings, we used w[*i*] and c[*i*] for X and Y loadings, respectively, for component *i*. The loading plots for the first two components of the PLS projections are shown in Figure 5a (countries) and 5b (antibiotics). Figure 5b shows that the first Y component had high loadings for all antibiotics except the quinolones, and to a lesser extent amoxicillin, amoxicillin/clavulanate, clindamycin, and co-trimoxazole. Thus, this component picked up countries with resistance patterns dominated by most of the non-quinolone agents. X variables with high loadings are highly correlated with these Y loadings—hence from Figure 5a we see that France, Hong Kong, and Japan were the countries most highly associated with resistance to most non-quinolone agents. The next PLS component, captured in Figure 5c and 5d, was primarily driven by the United States, Japan, and the Slovak Republic, and picked up isolates that were differentiated based upon macrolide/β-lactam/doxycycline/co-trimoxazole resistance. Note that Figure

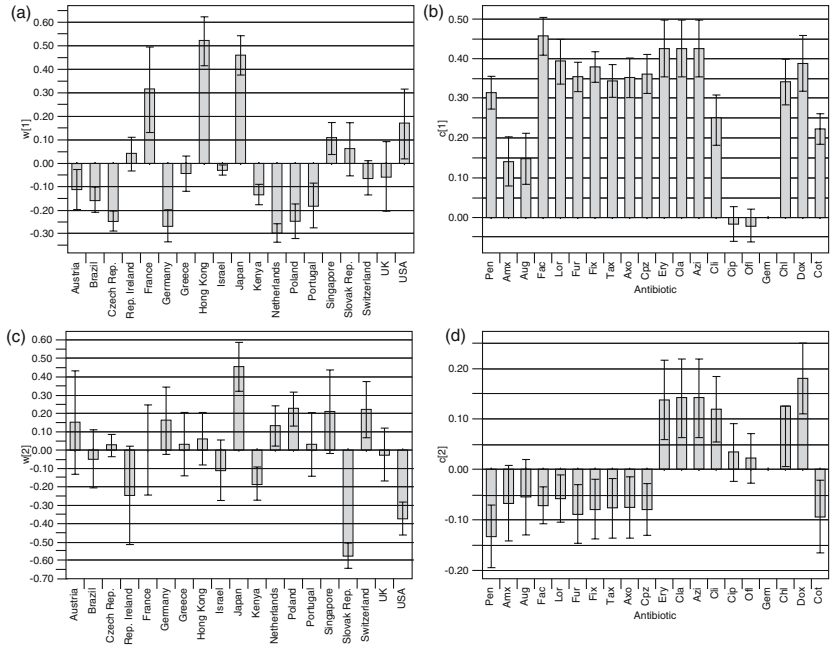

Figure 5. Loading plots for PLS models, relating country to antibiotype. (a) X matrix loading for component 1; (b) Y matrix loading for component 1; (c) X matrix loading for component 2; (d) Y matrix loading for component 2.

5b and 5d shows that there was no information about the quinolones in the first two PLS components.

As with PCA, it is of interest to plot two-dimensional plots. In Figure 6a we plotted w[1] and w[2] for the country loadings, and in Figure 6b c[1] and c[2] for the antibiotic loadings. Figure 6a shows which countries had the most extreme patterns of resistance (The Slovak Republic, United States, Japan, Hong Kong, and France) compared with the rest of the countries sampled. Figure 6b shows the projection into the Y space for the antibiotics, and spatially relating positions of antibiotics and countries in Figure 6a and 6b provides an understanding of the relationship between antibiotic resistance and country. Hong Kong, with high loadings in component 1 and loadings close to zero in component 2 had a relatively large number of isolates with resistance across all antibiotics. France was similar to Hong Kong, but a smaller component 1 loading suggests that this type of resistance was not as prevalent as in Hong Kong. Japan, high in component 1, but low in component 2 had isolates with resistance across all antibiotics, but also had a set of isolates with high

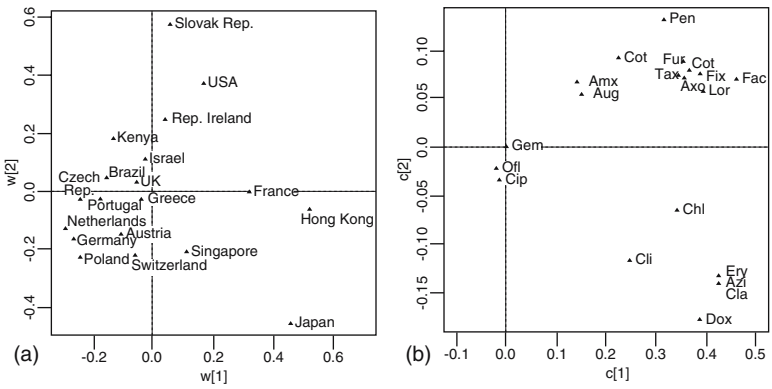

Figure 6. Two-dimensional loading plots for PLS models relating country to antibiotype. (a) X variable loadings for the first two components; (b) Y variable loadings for the first two components.

resistance among the macrolides, chloramphenicol, doxycycline, and clindamycin with low resistance among the β-lactams. Japan's second component of resistance was almost the opposite of the United States, which appeared to have isolates with resistance among the β-lactams but not among other antibiotics. The Slovak Republic had isolates generally associated with β-lactam resistance. The quinolones were not significant in the first or second component and hence no interpretations can be made regarding quinolone resistance and country for these two components. Note that this PLS model, with two components, explained only 30% of the information among the countries. The model points toward relationships between countries and antibiotics that are potentially of interest and worth further investigation.

The multivariate analysis clearly shows the broad patterns of antibiotic MICs and resistance for 1998 in *S. pneumoniae*. Based on PCA, β-lactam and macrolide resistance were responsible for the greatest variation among isolates, followed by quinolone resistance and resistance to chloramphenicol. At the MIC and antibiotype levels there were distinct clusters of isolates, which were largely determined by β-lactam and macrolide resistance. As the prevalence of quinolone resistance is still low, these agents were set apart from the other classes. At the antibiotype level, amoxicillin and amoxicillin/clavulanic acid were distinct from the population of other β-lactams, with less resistance to these two antibiotics.

4.2.1. Multidimensional scaling

It is often of interest to genotype the isolates. Isolates that are similar genetically can be grouped together, and likewise, isolates dissimilar genetically,

can be identified using a multitude of techniques. Multidimensional scaling (MDS) is one such technique, which is often applied to the analysis of genetic distances—see, for example, Agodi *et al*., 1999. For a detailed description of this approach, see texts on multivariate statistics, such as Morrison, 1990. In this section we show an example from a set of 193 isolates sampled and genotyped from the 1998 and 1999 Alexander Project. In this example, the housekeeping gene *gki* was genotyped. After the nucleotide sequences were determined, the genetic divergence matrix for pairwise divergence among the isolates was computed (see Section 4.3 for details).

Similar to PCA and PLS, MDS approximates the data matrix, in this case the matrix of genetic divergence, in a lower dimensional space. Hence the 193-dimensional divergence matrix will be represented in a few, interpretable dimensions. We applied the PCA program from SIMCA to the divergence matrix; the two-dimensional loadings plot for p[1] and p[2] is shown in Figure 7. From this figure, along the p[1] axis we see isolates which are genetically highly divergent. In particular, a lone isolate collected in Italy in 1999 at the p[1] = −0.05 level (circled in Figure 7a) was very different from, for example, any of the isolates in the p[1] = 0.06–0.08 range. This was confirmed by plotting pairwise distances for the isolate at p[1] = −0.05 and two isolates at the other extreme. As an example we picked two isolates near p[1] = 0.08 and p[2] = 0 (one from the United Kingdom 1998, the other from Portugal 1999), and plotted the set of 193 distance pairs for these two, and for the former against the p[1] = −0.05 Italian 1999 isolate (Figure 8). In Figure 8a, the two isolates that

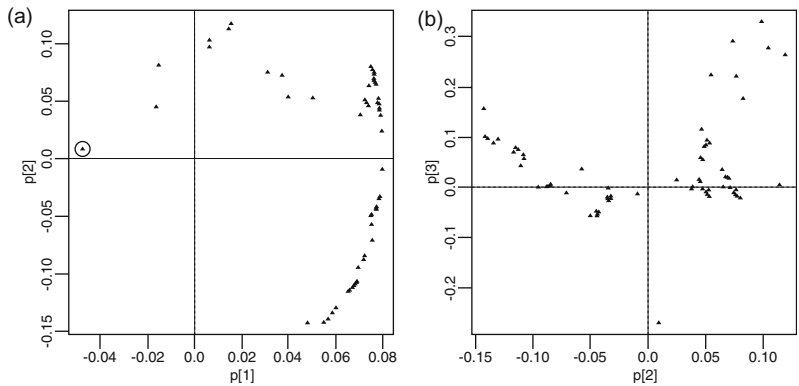

Figure 7. Multidimensional scaling of *gki* divergence matrix for 193 isolates from the Alexander project, 1998–9. (a) Two-dimensional plot for the first two components, identifying divergent isolates. The circled isolate is genetically divergent on p[1] and was collected in Italy in 1999; (b) Two-dimensional plot for the last two components, identifying divergent isolates and clusters of similar isolates.

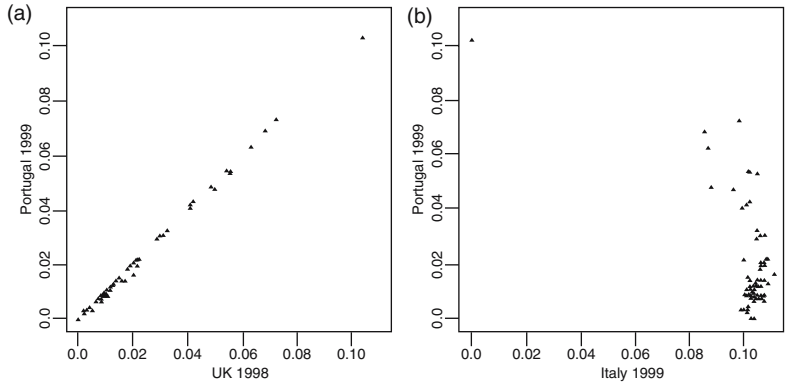

Figure 8. Comparison of pairwise distances for selected isolates. (a) Isolates close together in the first two components of the MDS model; (b) Isolates far apart in the first two components of the MDS model.

are close together on the plot are highly correlated; the Italy 1999 isolate is not correlated with these and the divergence from each of the other isolates is large (Figure 8b). Once genetically divergent isolates are identified, cross referencing with MIC and/or antibiotype data can be carried out, to determine the relationship of these divergent isolates to their MIC/resistance phenotype.

4.2.2. Summary

The projection methods described above can handle extremely large numbers of observations and variables. For example, the Alexander Project 10-year dataset (1992–2001) has well over 35,000 isolates; despite this, analysis of the entire dataset is well within the computational boundaries of these projection methods. This may be of particular interest, for example, when investigating time trends over the 1992–2001 period. The many variables available in the Alexander Project data can also be modeled, such as age, gender, isolate source, and country. It is also possible to assess genetic divergence for these very large sets of data. Many other data mining methods exist for evaluation of such large datasets, and each method has its merits. Projection methods are robust methods that can handle extremely large data sets of predictor (X) and/or response (Y) matrices, and user-friendly software is readily available.

4.3. Evolutionary genetic approaches

Technical developments associated with the polymerase chain reaction (PCR) and automated DNA sequencing technology, over the course of the last

decade or so, have made it very easy to obtain large amounts of comparative sequence data from virtually any organism. Alongside this progression in laboratory technology, the evolutionary analysis of molecular sequence data has also advanced. This concomitant development of molecular biology technology and analytical perspective has resulted in the burgeoning field of molecular phylogenetics, which is now influencing all areas of biology. Despite this, the application of modern principles and techniques of molecular phylogenetics to the analysis of antibiotic resistance development and spread has not been widely undertaken. Nonetheless, we feel the application of such a perspective is ideally suited to extracting important information from large antimicrobial susceptibility databases, such as the Alexander Project. This is at least partly due to the fact that molecular phylogenetics falls within the broader realm of comparative biology and the nature of such databases provides numerous comparative possibilities. Databases such as the Alexander Project have a temporal perspective allowing comparisons between years, a geographical component allowing comparisons between collecting centers, as well as susceptibility data for numerous antibiotics for each isolate. This permits the correlation of isolate genetics with resistance over time and across geographical regions. The purpose of this section is to outline a few modern molecular phylogenetic perspectives on typical questions of antibiotic resistance development and spread in *S. pneumoniae*.

4.3.1. *S. pneumoniae* and questions of clonal spread

Analyses performed over recent years suggest that a small number of genotypes are responsible for >85% of fully penicillin-resistant pneumococci in the United States (MICs ≥2 mg/L) (Corso *et al*., 1998; Gherardi *et al*., 2000; Richter *et al*., 2002). The majority of these studies used pulse field gel electrophoresis (PFGE), however, and in an organism such as *S. pneumoniae*, where nucleotide substitutions are relatively uncommon, PFGE is arguably too imprecise a method to gain an accurate picture of the species' genetic diversity. Furthermore, most of these studies examined only resistant isolates, whereas the inclusion of isolates with resistant as well as susceptible phenotypes is necessary to gain a more complete picture of the origins of resistance. An important additional level of specificity to isolate typing has been achieved with multiple locus sequence typing—MLST (Enright and Spratt, 1998; Maiden *et al*., 1998; McGee *et al*., 2001). However, the allele frequency data that arise from such studies do not lend themselves to forming an accurate picture of the cladistic history of the isolates.

Individuals from the same species diverge later than individuals from different species, which means that intraspecific molecular sequence data are typified by much lower levels of sequence variation. We have found this to be

particularly the case in an organism like *S. pneumoniae*. This species exhibits very low levels of genetic diversity in comparisons of housekeeping gene sequences in clinical isolates sampled from globally distributed locations. In contrast, in comparisons involving the same housekeeping genes for the same country and year in the Alexander Project collection, *H. influenzae* has much higher genetic diversity than *S. pneumoniae*. For example, in the Alexander Project during collection of isolates for the United Kingdom in 1998, indexes of diversity for three different housekeeping gene sequences are 1.8–5.4 times higher for *H. influenzae* than for *S. pneumoniae* (Table 2).

The relative lack of genetic variation in *S. pneumoniae* means that it can be more difficult to accurately reconstruct the evolutionary history of isolates using traditional molecular phylogenetic procedures (which require at least moderate levels of sequence divergence between entities), particularly when the number of isolates is quite large. Furthermore, it is very difficult to root an *S. pneumoniae* phylogeny using another species. This is because most of the taxa widely recognized to be different species are much too distant to provide anything but a long branch attraction problem (a phylogenetic artifact which results in some sequences being artificially "dragged" to the base of the tree due to homoplasy with the outgroup), and other taxa which are currently classified as different species, may have no evolutionary basis for such a classification (Stanhope, unpublished data).

Methods have been developed over the course of the last decade, and are coming into increasing use over the last number of years, that employ phylogenetic statistical procedures for dealing with this problem of low sequence variation in the reconstruction of evolutionary history. One such method, known as statistical parsimony, was developed by Templeton in a series of important papers in the early to mid-1990s (Templeton, 1995; Templeton and Sing, 1993; Templeton *et al*., 1987, 1992). In addition to being able to reconstruct accurate histories with low sequence divergence, the method also has a number of other important benefits: (1) the algorithm collapses the sequences into their

Table 2. Nei and Li's (1979) index of nucleotide diversity; based on total number of alleles, population frequency of each allele, and number of nucleotide substitutions per site between alleles; for *S. pneumoniae* ($N = 75$) and *H. influenzae* ($N = 73$) collected from the United Kingdom as part of the 1998 Alexander Project

Species	*gdh*	*gki*	*recP*
S. pneumoniae	0.01092	0.01479	0.00717
H. influenzae	0.02623	0.02633	0.03844

various haplotypes, estimates the maximum number of differences among these haplotypes, resulting from single substitutions, and then joins the haplotypes in a parsimony network with each step justified by a probability level of 95%; (2) the program TCS estimates haplotype outgroup probabilities for each haplotype, with the highest probability being judged the most likely ancestral haplotype for that set of sequences (Clement *et al*., 2000); (3) it can be combined with a nested analysis procedure to partition the resulting network into a series of nested clades which can in turn be used to statistically test associations between genotype and phenotype, where "phenotype" could represent anything, such as geographic location, clinical setting, or antibiotic resistance phenotype (Templeton and Sing, 1993; Templeton *et al*., 1987).

To illustrate this evolutionary approach using TCS, statistical parsimony networks were reconstructed from *gki* (840 bp) and *gdh* (1,245 bp) sequences for a set of isolates possessing a mixture of resistance phenotypes collected from Ohio, USA (Alexander Project collection, 2000) (Figure 9). For each of these two networks we tested the following null hypothesis: no association between genotype and isolates with penicillin (Pen) MICs >4 mg/L. For both these loci and networks the null hypothesis is rejected. Thus, we can conclude that, at least for this set of isolates, there was a significant correlation between genotype and isolates with penicillin (Pen) MIC >4 mg/L, indicating a significant degree of clonality to this type of resistance. By then examining the phenotypes of the constituent members of the various nests in the network, we can determine if there is a single clone or several clones that have convergently evolved this phenotype. In the present example, the vast majority of penicillin resistance of MIC >4 mg/L could be explained by two or three clones which have convergently evolved this phenotype. The fact that this test is not significant at the "0–step" clades (i.e., the individual haplotypes) indicates that both the relationships depicted in the network and the nested cladistical design associated with it were necessary to detect the genotype–phenotype association. In other words, any attempt to correlate haplotypes with resistance, while not taking into consideration the evolutionary history depicted in these networks, would have resulted in inaccurate conclusions.

Although in recent years there has been important work accomplished on the question of clonality and resistance in *S. pneumoniae*, it is also true that the issue can still benefit from the analytical procedures and perspective typical of modern molecular evolutionary biology. The statistical parsimony approach has various advantages, including the fact that it can be scaled up tremendously to include hundreds of isolate sequences. In contrast, traditional phylogenetic methods would be bogged down with such large numbers of highly similar sequences, as those typical of comparative data regarding *S. pneumoniae* housekeeping genes.

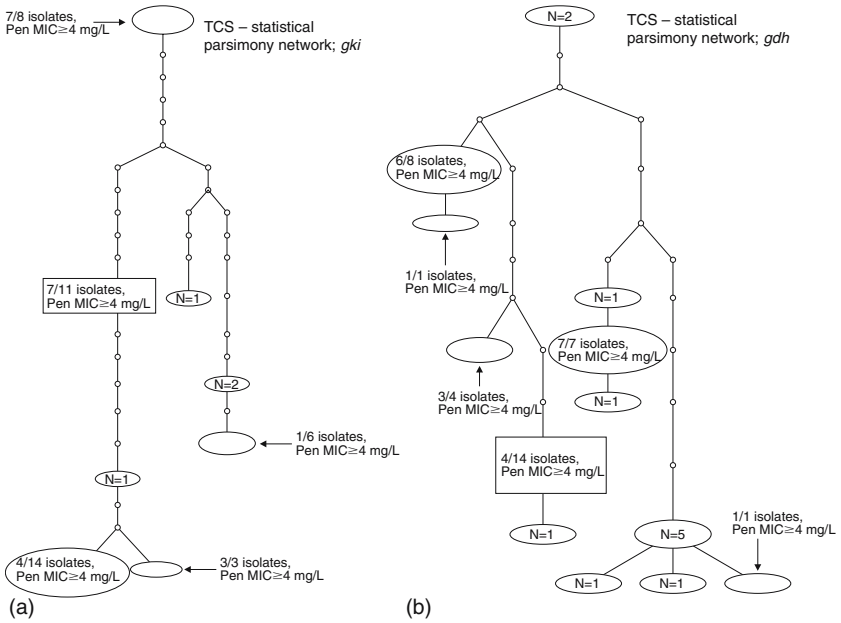

Figure 9. Statistical parsimony networks reconstructed from: (a) *gki* (840 bp) and (b) *gdh* (1,245 bp) sequences for a set of isolates ($N = 46$ for each locus) with a mixture of resistance phenotypes collected from Ohio, USA (Alexander Project, 2000). Small circles indicate nodes representing intermediate haplotype states not found in this sample of isolates. Each line represents a single mutational event. Larger ovals and two rectangles represent haplotypes that were sequenced in this sample of isolates, with the number of isolates of each indicated. The proportion of isolates of a particular haplotype possessing the penicillin MIC >4 mg/L phenotype are indicated (e.g., 4/14 isolates with penicillin MIC >4 mg/L). Haplotypes not represented by this resistance phenotype are simply labeled with the number of isolates with that sequence (e.g., $N = 5$).

4.3.2. Horizontal transfer of resistance loci in *S. pneumoniae*

It has been known for many years that one of the principal mechanisms for *S. pneumoniae* resistance to β-lactam antibiotics is through specifically mutated penicillin-binding protein (*pbp*) genes. Unlike the housekeeping genes of *S. pneumoniae, pbp*s tend to be highly variable between isolates with a mosaic pattern in their homology comparisons to other species of streptococci. This pattern has provided evidence that these hyper-variable, resistance-conferring *pbp*s in *S. pneumoniae* have their origins in lateral gene transfer events involving other species, followed with intraspecific recombination events to create the mosaics. The frequency with which such lateral resistance transfer events take place is not well understood.

Arguments regarding lateral gene transfer events, whether it be regarding antibiotic resistance or not, are often based on basic interpretations of percent identity, or quite simply, BLAST (basic local alignment search tool) alignment scores. However, understanding whether a gene has been laterally transferred from another lineage is an evolutionary biology problem. A high profile example of the risks of failing to use such a perspective concerns claims, by the Human Genome Sequencing Consortium, that hundreds of human genes likely resulted from independent horizontal transfer events from bacteria at various points in the diversification of vertebrates (IHGSC, 2001). This conclusion was based on BLASTP alignment scores, or in other words, basic interpretations of sequence homology. However, subsequent, detailed phylogenetic analysis of this claim, did not find any support for horizontal gene transfer from bacteria to vertebrates (Stanhope *et al*., 2001). Similarly, the best means to understand the history and dynamics of lateral transfer of *pbp*s in *S. pneumoniae* is to employ phylogenetic principles and techniques.

If two or more genes have the same evolutionary history then it is likely that there has been no lateral transfer involving those genes, and their history can be explained through a shared common ancestry and descent. In contrast, genes that have been laterally transferred will be discordant with the history depicted by genes that are not laterally transferred. Housekeeping genes are much less likely to be laterally transferred than are *pbp*s, and there is little *a priori* evidence to support their lateral transfer. Thus, the clearest phylogenetic evidence for lateral gene transfer of *pbp*s would be strongly supported by conflicting branching arrangements when comparing the phylogenies of *pbp* sequences and housekeeping genes for the same set of isolates. The phylogeny arising from the housekeeping genes can, therefore, be regarded as a "control phylogeny," or the best estimate of the true phylogeny. Nonetheless, in order to help verify their vertical inheritance, there are various methods available to detect the presence of recombinant sequences in any given sequence alignment (see, e.g., Posada, 2002). If one or more of such methods identify the presence of a recombinant, that particular sequence can be excluded from the set under analysis. These same methods can be used to distinguish between intragenic recombination of *pbp*s vs lateral transfer. In other words, it is possible that discordant phylogenies between housekeeping genes and *pbp*s could be the result of *pbp* lateral transfer, or intragenic recombination.

To illustrate this approach, we include an example of *pbp* sequences from a set of Alexander Project isolates from the United Kingdom collected in 1998. The control phylogeny in this case is based on an alignment of two concatenated housekeeping gene sequences, *gdh* and *gki* for a total of 2,085 bp; complete *gdh* and *gki* sequences were obtained for each isolate, these two sequences were joined together to make one long sequence for each isolate, and then the set of concatenated sequences were aligned. The low sequence

divergence typical of *S. pneumoniae* housekeeping genes means that a single gene often does not have sufficient phylogenetic signal to reconstruct robust phylogenies using traditional methods (i.e., in comparison with methods such as TCS). However, low sequence divergence is not typical of *pbp*s, and thus in our experience it is possible to easily reconstruct reliable histories of at least *pbp*1a, *pbp*2b, and *pbp*2x genes. Traditional phylogenetic methods are best suited for comparisons between evolutionary histories—there is a wealth of theory, and statistics, for comparing the branching arrangements of bifurcating phylogenies, but at present such methods for comparing networks, such as those obtained from TCS, lag behind. Programs for computing phylogenies from molecular sequence are numerous, but the two most common and comprehensive are PHYLIP (Felsenstein, 1993) and PAUP* (Swofford, 2002).

In the present example the resulting housekeeping gene sequences showed no evidence for intragenic recombination and there was no evidence of lateral transfer for either housekeeping gene (no strongly supported conflicting nodes in individuals trees for each gene). Consequently, both loci were combined to yield the unrooted maximum likelihood tree depicted in Figure 10. Similarly, there was no evidence for intragenic recombination in the *pbp*2x sequences. Note that the composition of the italicized clade of isolates in the housekeeping tree includes two phenotypes: susceptible (isolates labeled 0) and those resistant to penicillin, cephalosporins, and co-trimoxazole (isolates labeled 17). Note that in the *pbp*2x tree, the "17" phenotype is either identical or very closely related to several other isolates with multidrug resistant phenotypes (labeled 23, 20, and 24), to which it was unrelated in the housekeeping tree. Furthermore, the "0" isolates from the housekeeping 0/17 clade no longer group with the multidrug resistant 17 isolates. These results indicate that there was the lateral transfer of a particular *pbp*2x allele into one or another of the "0" isolates of the 0/17 clone, and that this lateral transfer was correlated with a major shift in phenotype from all susceptible to multidrug resistant.

In general, our present collection of data suggests that *pbp* genes that are judged by our approach to be laterally transferred are often highly similar between a selection of multidrug resistant isolates. As these isolates share no close relationship on the basis of housekeeping genes, the most parsimonious explanation is lateral transfer rather than intragenic recombination. Whether the explanation is intragenic recombination or lateral transfer, the knowledge of the evolutionary history of the isolates themselves, based on their housekeeping gene sequences, can lead to important conclusions regarding the shifts in phenotype that have occurred coincidentally with the acquisition or alteration of the *pbp*s in question.

This comparative approach can be scaled up to include larger numbers of isolates from different locations and years allowing examination of (1) the relative frequency of lateral transfer of the resistance loci and whether this differs

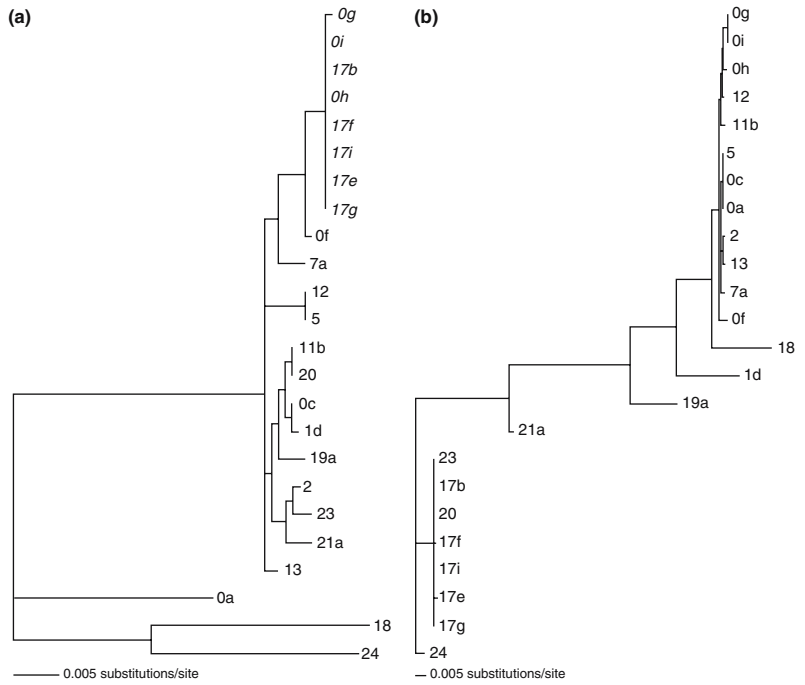

Figure 10. Unrooted maximum likelihood tree for *S. pneumoniae* (a) concatenated housekeeping genes *gdh* and *gki* and (b) *pbp*2x genes from isolates collected from the United Kingdom in 1998 (Alexander Project); branch lengths are drawn proportional to the amount of sequence change. The numbers and/or number–letter combinations refer to different isolates; the larger the number, the more antibiotics to which that isolate is resistant; 0 refers to isolates that are susceptible to all antibiotics tested for that particular year.

between years and countries; (2) which variants of these loci are associated with major shifts in resistance phenotype; (3) the sequence of lateral transfer events involving the resistance loci—that is, which resistance genes are acquired first, or all at once; (4) whether there are geographically specific *pbp*s that are being laterally transferred amongst a set of isolates within a given region.

5. CONCLUSIONS

Microbiologists are conditioned to approach a scientific subject with a hypothesis, a protocol outlining how to proceed, and a clear idea of what they are looking for. Data mining is a relatively new concept to microbiologists, and requires a change in mind set, as a key element of this approach is to apply methods developed for other fields like statistics and bioinformatics to a search

for novel facts within the accumulated data. The papers cited in Section 2 of this chapter are included here to encourage workers interested in the subject of antimicrobial resistance to approach their subject within a new paradigm. The three general methods presented in some detail, antibiotypes, multivariate analysis, and evolutionary genetics, are techniques designed to stimulate the investigator rather than present any one set approach to the subject. Interested parties are encouraged to take the first step, namely, search for colleagues with expertise in other fields to become familiar with the rich data generated by antimicrobial surveillance and other research programs from the viewpoint of their individual specialities and search for novel aspects that will shed light on a problem that will only become more critical over the coming years. Data mining is one approach that may offer novel insights into our understanding of resistance, and ultimately may result in providing potential solutions to slow the rate of resistance against organisms of medical and environmental importance.

REFERENCES

Agodi, A., Campanile, F., Basile, G., Viglianisi, F., and Stefani, S., 1999, Phylogenetic analysis of macrorestriction fragments as a measure of genetic relatedness in *Staphylococcus aureus*: The epidemiological impact of methicillin resistance. *Eur. J. Epidemiol*., **15**, 637–642.

Box, G. E. P. and Jenkins, G. M., 1976, *Time Series Analysis: Forecasting and Control*, 2nd edn. Holden Day, San Francisco, CA.

Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., and Moser, S. A., 1998, Association rules and data mining in hospital infection control and public health surveillance. *J. Am. Med. Inform. Assoc.*, **5**, 373–381.

Brossette, S. E., Sprague, A. P., Jones, W. T., and Moser, S. A., 2000, A data mining system for infection control surveillance. *Meth. Inf. Med.*, **39**, 303–310.

Brown, S. M., Benneyan, J. C., Theobald, D. A., Sands, K., Hahn, M. T., Potter-Bynoe, G. A. *et al*., 2002, Binary cumulative sums and moving averages in nosocomial infection cluster detection. *Emerg. Infect. Dis.*, **8**, 1426–1432.

Clement, M., Posada, D., and Crandall, K. A., 2000, TCS: A computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1659.

Corso, A., Severina, E. P., Petruk, V. F., Mauriz, Y. R., and Tomasz, A., 1998, Molecular characterization of penicillin-resistant *Streptococcus pneumoniae* isolates causing respiratory disease in the United States. *Microb. Drug Resis.*, **4**, 325–337.

Enright, M. C. and Spratt, B. G., 1998, A multilocus sequence typing scheme for *Streptococcus pneumoniae*: Identification of clones associated with serious invasive disease. *Microbiology*, **144**, 3049–3060.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S., 2001, *Multi- and Megavariate Data Analysis: Principles and Applications*. Umetrics Academy, Umea, Sweden.

Felsenstein, J., 1993, PHYLIP (Phylogeny Inference Package) version 3.6a2. Distributed by the author: http://evolution.genetics.washington.edu/phylip.html, Department of Genetics, University of Washington, Seattle.

Gherardi, G., Whitney, C. G., Facklam, R. R., and Beall, B., 2000, Major related sets of antibiotic-resistant pneumococci in the United States as determined by pulsed-field gel electrophoresis and pbp1a-pbp2b-pbp2x-dhf restriction profiles. *J. Infect. Dis.*, **181**, 216–229.

Hand, D. J., 1998, Data mining: Statistics and more? *Am. Statistician*, **52**, 112–118.

IHGSC, 2001, International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Janne, K., Pettersen, J., Lindberg, N.-O., and Lundstedt, T., 2001, Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration. *J. Chemometrics*, **15**, 203–213.

Kaslow, R. A. and Moser, S. A., 2000, Role of microbiology in epidemiology; before and beyond 2000. *Epidemiol. Rev.*, **22**, 131–135.

Lopez-Lazano, J. M., Monnet, D. L., Yague, A., Burgos, A., Gonzalo, N., Campillos, P. *et al*., 2000, Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: A time series analysis. *Int. J. Antimicrob. Agents*, **14**, 21–31.

MacGregor, J. F. and Kourti, T., 1995, Statistical process control of multivariate processes. *Control Eng. Practice*, **3**, 403–414.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R. *et al*., 1998, Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA*, **95**, 3140–3145.

McGee, L., McDougal, L., Zhou, J., Spratt, B. G., Tenover, F. C., George, R. *et al*., 2001, Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. *J. Clin. Microbiol.*, **39**, 2565–2571.

Monnet, D. L., Lopez-Lazano, J. M., Campillos, P., Burgos, A., Yague, A., and Gonzalo, N., 2001, Making sense of antimicrobial use and resistance surveillance data: Application of ARIMA and transfer function models. *Clin. Microbiol. Infect.*, **7**, 29–36.

Morrison, D. F., 1990, *Multivariate Statistical Methods*, 3rd edn. McGraw-Hill, Hightstown, NJ.

Moser, S. A., Jones, W. T., and Brossette, S. E., 1999, Application of data mining to intensive care unit microbiologic data. *Emerg. Infect. Dis.*, **5**, 454–457.

NCCLS, 2000, *Performance Standards for Antimicrobial Susceptibility Testing; Tenth Informational Supplement*. National Committee for Clinical Laboratory Standards, Wayne, PA.

Nei, M. and Li, W. H., 1979, Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*, **76**, 5269–5273.

Nguyen, D. V. and Rocke, D. M., 2002, Multi class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Peterson, L. R. and Brossette, S. E., 2002, Hunting health care-associated infections from the clinical microbiology laboratory: Passive, active and virtual surveillance. *J. Clin. Microbiol.*, **40**, 1–4.

Posada, D., 2002, Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Mol. Biol. Evol.*, **19**, 708–717.

Poupard, J., Brown, J., Gagnon, R., Stanhope, M. J., and Stewart, C., 2002, Methods for data mining from large multinational studies. *Antimicrob. Agents Chemother.*, **46**, 2409–2419.

Richter, S. S., Heilmann, K. P., Coffman, S. L., Huynh, H. K., Brueggemann, A. B., Pfaller, M. A. *et al*., 2002, The molecular epidemiology of penicillin-resistant *Streptococcus pneumoniae* in the United States, 1994–2000. *Clin. Infect. Dis.*, **34**, 330–339.

Sahm, D. F., Jones, M. E., Hickey, M. L., Diakun, D. R., Mani, S. V., and Thornsberry, C., 2000, Resistance surveillance of *Streptococcus pneumoniae, Haemophilus influenzae* and *Moraxella catarrhalis* isolated in Asia and Europe, 1997–1998. *J. Antimicrob. Chemother.*, **45**, 457–466.

SIMCA, 2000, *8.0. Umetrics AB.* Umea, Sweden.

Stanhope, M. J., Lupas, A., Italia, M. J., Koretke, K. K., Volker, C., and Brown, J. R., 2001, Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, **411**, 940–944.

Swofford, D. L., 2002, *PAUP* Version 4.0b10*. Sinauer Associates, Sunderland, MA.

Templeton, A. R., 1995, A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics*, **140**, 403–409.

Templeton, A. R., Boerwinkle, E., and Sing, C. F., 1987, A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, **117**, 343–351.

Templeton, A. R., Crandall, K. A., and Sing, C. F., 1992, A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.

Templeton, A. R. and Sing, C. F., 1993, A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Thornsberry, C., Ogilvie, P. T., Holley, H. P. Jr., and Sahm, D. F., 1999, Survey of susceptibilities of *Streptococcus pneumoniae, Haemophilus influenzae*, and *Moraxella catarrhalis* isolates to 26 antimicrobial agents: A prospective U.S. study. *Antimicrob. Agents Chemother.*, **43**, 2612–2623.