ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Aug. 2002, p. 2409–2419 Vol. 46, No. 8
0066-4804/02/$04.00+0 DOI: 10.1128/AAC.46.8.2409–2419.2002

Methods for Data Mining from Large Multinational Surveillance Studies

James Poupard,* James Brown, Robert Gagnon, Michael J. Stanhope, and Chad Stewart

GlaxoSmithKline, Collegeville, Pennsylvania 19426-0989

Received 19 April 2001/Returned for modification 27 January 2002/Accepted 24 April 2002

Traditionally, large surveillance studies have been analyzed by the use of the MICs at which 90% of isolates tested are inhibited (MIC_{90}s), MIC_{50}s, frequency distributions, and percent susceptibility. In the past, these approaches have proved satisfactory for the monitoring of resistance. From these traditional uses, one can readily detect an increase in MICs for organism and drug combinations. Now that large surveillance studies have been conducted for a number of years and databases have grown to include a large number of datum points, new approaches to the extraction of useful information from these studies are needed. The present study proposes approaches, including the use of antibiotypes, principal components analysis, phylogenetics, and population genetic analysis, to the evaluation of data from large multinational surveillance studies. Application of these types of analyses can be used to describe genetic diversity, analyze changes in susceptibility patterns over time, and possibly, shed light on the origins and evolution of antimicrobial resistance. As global surveillance studies become more common and new questions concerning the evolution of resistance are raised, innovative approaches to analysis of the data will increase in importance.

With the increase in the numbers of multidrug-resistant organisms and the need to monitor evolving patterns of resistance, large multinational surveillance studies have become more critical. In the past, these surveillance studies have been analyzed by using the MICs at which 90% of isolates tested are inhibited (MIC_{90}s), MIC_{50}s, frequency distributions, and percent susceptibility. The present paper proposes approaches that may be used to create novel evaluations of susceptibility data from local or international surveillance studies.

Antibiotypes. Antibiotyping involves the conversion of the antimicrobial susceptibility pattern of an isolate into a series of 0s and 1s (susceptible and nonsusceptible, respectively). This series or string of digits can then be used in a variety of meaningful ways including some of the novel approaches described in this study. For example, basic susceptibility patterns can be derived from the string of 0s and 1s, which is adaptable to all computer databases, and this binary code can be used to generate both routine and novel analyses. The binary code can also be translated into a number with a small number of digits. Every number within a group of three digits is assigned a value of 1, 2, or 4, respectively. The one-digit value assigned for that group of three digits represents the sum of the values for nonsusceptible results, for example, 101 = 5 and 001 = 4.

It is also possible to assign a two- or three-digit hyphenated code based on the number of nonsusceptible results (1s) in the binary string. The first digit in the hyphenated number represents the total number of nonsusceptible results seen in the string. Each unique string with the same number of nonsusceptible results would be given a new second digit, and the process would be continued until each antibiotype has a unique number designation. For the present analysis a single breakpoint based on National Committee for Clinical Laboratory Standards (NCCLS) interpretations for susceptibility was used for each organism-drug combination. For a more sophisticated analysis, the intermediate category could be incorporated or an artificial breakpoint for the detection of very low level resistance could be used.

PCA. International surveillance studies generate databases that are large and multivariate (or multidimensional). The databases that are generated include thousands of observations (individual isolates). Principal components analysis (PCA) is applied to represent multidimensional data in a reduced-dimensional space, usually two to four dimensions, so that one can obtain an overview of the data. Such an overview may reveal clusters among the observations, time trends, and interesting (outlying) observations. One may also observe relationships among the observations and variables and among the variables themselves. In this paper we show how the direct application of PCA to MIC results or antibiotypes in a database from a large global surveillance study reveals clustering of bacterial isolates in and between countries and across years. The PCA process may add a level of understanding to the data that extends beyond that obtained from the standard univariate approaches and provides a framework for additional analysis.

Population genetics and evolution of antibiotic resistance.
(i) Population genetics. Population genetics involves the study of how evolutionary forces such as mutation, migration, genetic drift, selection, and recombination change gene frequencies in populations. Asexual bacterial populations inevitably consist of at least some distinct clonal lineages; however, a substantial body of information now suggests that recombination in natural populations of bacteria is commonplace (2, 11). The relative contribution of recombination versus mutation as a mechanism of genetic change may vary between species and

* Corresponding author. Mailing address: Antimicrobial Profiling and Clinical Microbiology, GlaxoSmithKline, 1250 S. Collegeville Rd., Mail Code UP1340, Collegeville, PA 19426-0989. Phone: (610) 917-6284. Fax: (610) 917-4617. E-mail: James.A.Poupard@gsk.com.

TABLE 1. Antibiotype variability for *S. pneumoniae* from 1992 to 1998[a]

Country	1992		1993		1994		1995		1996		1997		1998	
	n	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a b*)	*n*	*N*, (*a*) *b*
United Kingdom	166	13, (1) 8	150	7, (1) 3	152	17, (1) 8	197	22, (1) 11	88	8, (1) 10	111	12, (1) 12	100	24, (1-2) 13
Germany	104	14, (1-3) 7	52	3, (1) 1	33	7, (1-2) 7	74	11, (1) 6	75	9, (1) 6	62	17, (2) 4	168	20, (1-2) 8
Italy	70	14, (1-2) 6	61	7, (1) 5	138	25, (2) 11	147	25, (2) 12	237	35, (3) 12	191	27, (2–7) 12	100	31, (6) 12
France	169	38, (1) 8	225	53, (1-4-5) 8	189	59, (5) 14	284	60, (5) 13	165	44, (6) 13	181	36, (5) 13	167	44, (1-2-6) 14
Spain	269	41, (7) 11	280	50, (9) 13	314	64, (4) 13	273	65, (5) 13	136	36, (1) 12	175	37, (9-10) 13	NT	
United States	125	21, (1) 11	185	25, (2) 10	147	24, (1-2) 12	81	19, (1-2)	79	20, (1-10) 13	124	34, (1) 14	1,456[b]	125, (2) 15

[a] Antibiotype variability represents the total number of unique antibiotypes with at least one nonsusceptible result. *n*, number of isolates tested; *N*, number of unique antibiotypes present; (*a*), predominant antibiotype; *b*, highest antibiotype reached; NT, not tested.
[b] Comparisons for this year are in question due to the large number of isolates tested.

possibly also between populations of the same species (21). In all likelihood, the genes of most bacterial species consist of both clonal and nonclonal elements. Clearly there is international movement of the hosts as well as selection in the pathogens. These factors taken collectively suggest that population genetic approaches have a place in the study of bacterial diversity, and indeed, this is a relatively new and flourishing area of study (21).

Diversity indices have been used in population biology and ecology as a comparative measure of genetic diversity across different populations. The haplotypic or nucleon diversity estimate (*h*) of Nei and Tajima (20) was initially developed to measure the heterozygosity of nuclear loci, but it has also been used to represent the diversity of clonal lineages, such as mitochondrial DNA genotypes (19). Since antibiotic susceptibility is genetically determined, *h* can also be used to estimate antibiotype diversity both within and between different countries and years. The low and high range of values for *h* are 0.0 and 1.0, respectively. Thus, instances in which all isolates in the population have the same antibiotype would have the lowest diversity ($h = 0.0$), while, conversely, if all isolates in a population had unique antibiotypes, the diversity would be the highest ($h = 1.0$). Since *h* is an aggregate estimate of diversity based on both the number of different antibiotypes and their frequencies of occurrence relative to those of other antibiotypes, *h* can be considered a comparative measure of the levels of multidrug resistance in a bacterial population.

In population genetics, relationships between populations are often constructed from samples of gene frequency and/or DNA haplotype frequency data, in which haplotype is equivalent to allele in classical genetics, except that it refers to any DNA segment (large or small), but not necessarily a gene proper. Antibiotic resistance is genetically determined, and thus, antibiotype (i.e., our scheme of 0s and 1s describing susceptibility and nonsusceptibility, respectively) can be considered analogous to a descriptor or at least representative of a haplotype. Thus, the frequencies of various haplotypes (antibiotypes) between populations (in different countries and years) can be used to assess relationships between the populations of species in the Alexander Project collection. From the frequency of each antibiotype for each country and year, a commonly used measure in population genetics known as genetic distance Neis (18) can be calculated. From a matrix of such genetic distances, for any particular species, phylogenies can be reconstructed depicting the relationships and relative changes between various countries and years for which surveillance data exist.

(ii) Evolution of antibiotic resistance. Since the inception of the field of cladistics approximately 40 years ago (9), morphological systematists have used the principle of maximum parsimony (MP) concomitant with matrices of 1s and 0s defining character presence and absence, respectively, to determine evolutionary relationships between the entities for which those characters were so coded. Our scheme of representing antibiotypes is analogous to such a coding scheme of the presence and the absence of characters. As such, if there is phylogenetic signal to the data, we should be able to use parsimony analysis of the coded antibiotic susceptibility to determine relationships between the antibiotypes and thus examine the evolution of multidrug resistance patterns.

The Alexander Project database. The Alexander Project database, a collection of susceptibility data for *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Staphylococcus aureus*, *Streptococcus* group A, and *Moraxella catarrhalis* isolates collected internationally from patients with community-acquired infections, was used for this study. Identification and susceptibility testing were performed at one of three central laboratories. Susceptibility data, first collected in 1992 and validated up to 1998 inclusively, have been published previously (3).

MATERIALS AND METHODS

Antibiotypes. For the purposes of this study, only the data for *S. pneumoniae* ($n = 5,644$) and *H. influenzae* ($n = 7,553$) from the years from 1992 to 1998 were analyzed, and only those antimicrobial agents that remained in use throughout the course of those years were used. The following drugs were included: penicillin (*S. pneumoniae* only), ampicillin (*H. influenzae* only), amoxicillin, amoxicillin-clavulanic acid, erythromycin, azithromycin, clarithromycin, ceftriaxone, cefaclor, cefixime, cefuroxime, ciprofloxacin, ofloxacin, chloramphenicol, doxycycline, and co-trimoxazole. The data from the following countries were considered: the United States, the United Kingdom, France, Spain (1992 to 1997 only), Italy, and Germany.

The 2000 NCCLS breakpoints were applied (17). Those antibiotics without NCCLS breakpoints were analyzed by using the NCCLS breakpoints for agents in the same class. It should be noted that these breakpoints were used for statistical purposes only. For this type of analysis, breakpoints can be assigned on the basis of various criteria, as long as the same criteria are used for the entire data set. For the analysis performed with *S. pneumoniae*, doxycycline and ciprofloxacin were analyzed by using the breakpoints for tetracycline and ofloxacin, respectively. For the analysis performed with *H. influenzae*, amoxicillin, erythromycin, and doxycycline were analyzed by using the breakpoints for ampicillin, azithromycin, and tetracycline, respectively. The percentage of isolates that were considered susceptible to all of the antimicrobials observed was determined for each year and country. In addition, strings of 0s and 1s were converted to

TABLE 2. Antibiotype variability for *H. influenzae* from 1994 to 1998[a]

Country	1994		1995		1996		1997		1998	
	n	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*	*n*	*N*, (*a*) *b*
United Kingdom	302	26, (3) 5	353	3, (2) 5	247	23, (4) 6	208	14, (1-2) 4	100	13, (4) 6
Germany	69	7, (1) 5	158	17, (2) 7	160	16, (3) 4	151	11, (1-3) 5	193	16, (2) 6
Italy	184	22, (3) 6	252	22, (4) 5	164	16, (3) 5	205	10, (1-2) 4	100	9, (1-3) 6
France	252	18, (3) 9	269	31, (2-3) 6	217	28, (3) 6	166	17, (3) 5	158	22, (2) 6
Spain	258	37, (3-4) 6	424	43, (5) 7	223	38, (3) 7	167	24, (2) 5	NT	
United States	372	3, (3) 6	386	26, (2) 6	230	25, (3) 5	215	27, (2-4) 7	1,370[b]	73, (3) 3

[a] Antibioty variability represents the total number of unique antibiotypes with at least one nonsusceptible result. *n*, number of isolates tested; *N*, number of unique antibiotypes present; (*a*), predominant antibiotype; *b*, highest antibiotype reached; NT, not tested.
[b] Comparisons for this year are in question due to the large number of isolates tested.

five-digit antibiotype numbers. However, this approach was found to be too cumbersome, and most analyses were performed with a two- or three-digit number based on the frequency of nonsusceptible designations. The frequencies of unique antibiotypes were determined by year and country.

PCA. Distributions of MIC data are generally not symmetric. For modeling purposes, transformation of the MIC data to achieve a close-to-symmetric distribution is common. For example, MIC distributions are often summarized by 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 an undue influence of high values (in this case, high MICs) on the model. For PCA, therefore, data were transformed to achieve a closer-to-symmetric distribution. A log transformation was used.

We examined the distribution of antimicrobial MICs for *S. pneumoniae* isolates in Spain from 1992 to 1997. PCA was carried out with the log MICs by using SIMCA software (version 8.0, 2000; Umetrics, AB, Umea, Sweden.). This software has no limitations in the allowable number of rows or columns of a data table and executes rapidly. The analyses conducted for the study described in this paper were run with SIMCA software in a few seconds. The mathematical complexities of PCA are beyond the scope of this paper; for introductory discussions, see the text of Morrison (16). Briefly, multidimensional data are projected into a lower-dimensional space, which allows a hidden structure to be revealed while at the same time conserving the variation in the data. By the methodology with SIMCA software, the principal components are summarized graphically by using score and loading plots. The coordinates of the observations in the lower-dimensional space are the scores. In order to interpret the score plot, it is necessary to know which variables are influential in the model and how they are correlated, and this information is provided by the loadings. Variables contributing similar information, for example, are correlated and are grouped together in the loading plot. Variables which are negatively correlated are located on opposite sides of the plot. The farther from the plot origin that a variable lies, the greater the impact that it has on the PCA model. By relating the score and loading plots, one is able to interpret the PCA model and gain understanding of the data set.

Population genetics. Antibiotype diversity was quantified by using the *h* value of Nei and Tajima (20), where

$$h = \frac{n\left(1 - \sum_{i=1}^{r} x_i^2\right)}{n-1}$$

and where x_i is the frequency of the *i*th antibiotype in a population of *n* antibiotypes and *r* is the number of antibiotypes. Possible values range from 0.0 to 1.0. These values represent the diversity range from lowest (all isolates in the population have the same antibiotype) to highest (each isolate has a unique antibiotype). From the frequency of each antibiotype, the genetic distance of Nei (18) was calculated with the PHYLIP program (4). A matrix of such genetic distances for both *H. influenzae* and *S. pneumoniae* populations for each country and year then served as input for phylogenetic reconstructions for each species by the method of Fitch and Margoliash (6), as implemented with the PHYLIP program. Fitch-Margoliash trees were reconstructed with global branch swapping in effect, by randomizing the input order 20 times, and with negative branch lengths disallowed (4). The Fitch-Margoliash method was chosen over the neighbor-joining method because of simulation studies that suggest its superior performance, as long as negative branch lengths are disallowed (13).

Evolution of antibiotic resistance. The principle of MP involves the identification of a topology that requires the smallest number of evolutionary changes to explain the observed differences among the entities under study. Parsimony can be conducted by a variety of methods. The simplest method and the one which imposes minimal constraints upon character state changes is known as the Wagner parsimony method (1, 7). This method allows free reversibility of characters. This means that the probability of a change from a 0 to a 1 is just as likely as the reverse. A consequence of this reversibility is that the length of the tree and the order of branching are independent of the position of the root. However, the assignment of the root (the ancestral branch) then provides an indication of the directionality of evolution for that particular data set. For the data in question, trees were rooted at the all-susceptible condition, i.e., the antibiotype represented by 0s for all antibiotics. This analysis therefore depicts the pattern of evolution of antibiotic resistance starting from an organism that is susceptible to all 15 antibiotics. We believe that this is reasonable, since the all-susceptible condition is the most frequent antibiotype for all countries and years, and thus, much of the evolution of multidrug resistance for a particular year and country must ultimately arise from this state. The data were divided into individual years

TABLE 3. Percentage of completely susceptible *S. pneumoniae* isolates[a]

Country	1994		1995		1996		1997		1998	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
France	189	29.1	284	26.8	165	38.2	181	38.1	167	31.7
Germany	33	48.5	74	81.1	75	78.7	62	61.3	168	71.4
Italy	138	50.7	147	46.9	237	44.3	191	47.6	100	28.0
Spain	214	28.7	273	29.7	136	33.8	175	29.1	NT	NT
United Kingdom	152	71.7	197	69.0	88	73.9	111	73.0	100	51.0
United States	147	60.5	81	60.5	79	63.3	124	51.6	1,456[b]	31.3

[a] Completely susceptible indicates an antibiotype with all 0s in the binary code. *n*, number of isolates tested; NT, not tested.
[b] Comparisons for this year are in question due to the large number of isolates tested.

TABLE 4. Percentage of completely susceptible *H. influenzae* isolates[a]

Country	1994		1995		1996		1997		1998	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
France	252	59.1	269	53.2	217	54.4	166	63.3	158	12.7
Germany	69	66.7	158	63.3	160	74.4	151	84.8	193	19.7
Italy	184	41.8	252	70.6	164	58.5	205	75.6	100	34.0
Spain	258	25.2	424	30.9	223	35.9	167	29.3	NT	NT
United Kingdom	302	56.0	353	48.4	247	63.6	208	82.7	100	25.0
United States	372	38.2	386	49.7	230	44.3	215	53.4	1,370[b]	24.7

[a] Completely susceptible indicates an antibiotype with all 0s in the binary code. *n*, number of isolates tested; NT, not tested.
[b] Comparisons for this year are in question due to the large number of isolates tested.

and countries because it is not possible to analyze together data for all years for any particular country. This is because the number of possible antibiotypes over the period from 1992 to 1998 vastly exceeds the number of phylogenetically informative positions, which results in millions of equally most parsimonious trees and eliminates any meaningful interpretation. We argue that dividing the analysis into country and year is logical because it represents a snapshot in time that certainly must capture a significant portion of the development of multidrug resistance for that period. Our purpose here is to illustrate the possibilities of such an approach by analyzing a few examples; data for the collection of *S. pneumoniae* and *H. influenzae* isolates from the United Kingdom in 1998 were chosen to illustrate the approach. A more detailed report involving all countries and years will be presented elsewhere.

Wagner parsimony was conducted with the program PAUP* (version 4.0b4a) by using a heuristic search with stepwise random additions of the antibiotypes and branch swapping via nearest-neighbor interchanges. Majority-rule consensus trees were computed from the set of most parsimonious trees with the LE50 option in effect; this option retains groups on less than 50% of the trees as long as such groups are compatible with those already on the tree. Antibiotic characters were subsequently traced on the framework of the most parsimonious tree by using McClade software (15). The presence of phylogenetic signal in these matrices of 0s and 1s was assessed by relative apparent synapomorphy analysis (14) with RASA (version 2.2) software (J. Lyons-Weiler, RASA version 2.2 for the Mac [http://loco.biology.unr.edu/archives/rasa/rasa.html], 1998).

RESULTS AND DISCUSSION

Antibiotypes. Once the antibiotype string (0s and 1s) was converted into the two- or three-digit antibiotype number, the frequency of appearance of nonsusceptible strains was determined by country and year. The number of *S. pneumoniae* isolates tested, the number of unique antibiotypes, the most predominant antibiotype, and the highest antibiotype number that appeared in each country from 1992 to 1998 are presented in Table 1. The same information for *H. influenzae* from 1994 to 1998 is presented in Table 2.

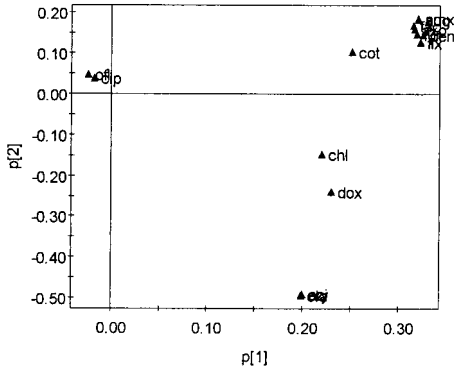

FIG. 1. PCA loadings plot of log MICs for isolates from Spain from 1992 to 1997. p[1], component 1; p[2], component 2; cot, co-trimoxazole; chl, chloramphenicol; dox, doxycycline; the cluster at the upper right consists of β-lactams; the cluster at the bottom consists of macrolides; the cluster at the upper left consists of the quinolones (ofl, ofloxacin; cip, ciprofloxacin).

For *S. pneumoniae* isolates (Table 1), variability (total number of unique antibiotypes) peaked in 1998 for the United Kingdom and Germany, 1995 for France and Spain, and 1996 for Italy. It is difficult to comment on the variability in the United States because a large number of isolates that may have biased the results were tested in 1998.

Variability peaked for the *H. influenzae* isolates obtained from the United Kingdom, Germany, Italy, France, and Spain in 1995 (Table 2). It is difficult to comment on the variability of the isolates in the United States because a large number of isolates that may have biased the results were tested in 1998. Except for the isolates obtained from Germany in 1998, variability among the *H. influenzae* isolates did correlate directly with the number of isolates tested, and no consistent trend toward an increase in a larger number of antibiotypes was observed.

One interesting observation that is not readily evident when surveillance data are presented in a traditional fashion (i.e., percent susceptibilities or MIC_{90}s) is the number of isolates that are completely susceptible (all 0s) to all of the antimicrobials tested. When the data are analyzed by the conversion of the susceptibility pattern into a binary code, the percentage of isolates susceptible to all antimicrobials tested becomes readily evident. The percentages of *S. pneumoniae* and *H. influenzae* isolates that were susceptible to all drugs tested are shown in Table 3 and Table 4, respectively, by year and country. The data presented show that for the *S. pneumoniae* isolates (Table 3) there was a slight trend toward an increase in the percentage of isolates in France and Germany that were susceptible to all antimicrobials tested. For the isolates in Italy, the United Kingdom, and the United States, there was a trend toward a decrease in the percentage of isolates that were susceptible to all antimicrobials tested.

For the *H. influenzae* isolates (Table 4), there was a trend from 1994 to 1997 toward an increase in the percentage of isolates from all of the countries that were susceptible to all of the antimicrobials tested when the results for 1994 and 1997 were compared. There was, however, a decrease in the percentage of isolates from all of the countries in 1998 susceptible to all antimicrobials tested.

PCA. For Spain for the years from 1992 to 1997 there are 15 antimicrobials, thus representing 15 dimensions or columns in

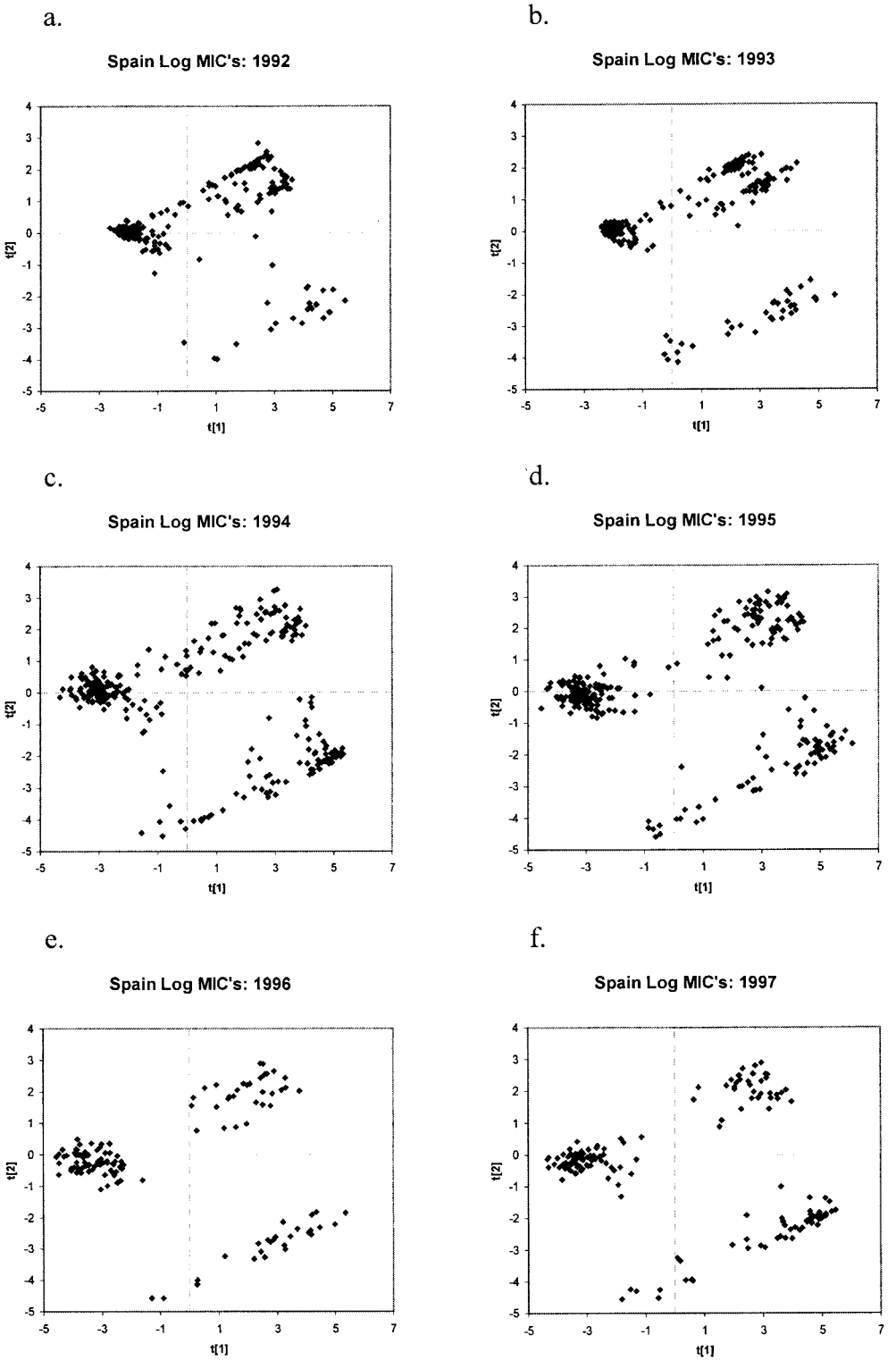

FIG. 2. PCA score plots of log MICs for isolates from Spain from 1992 to 1997 (a to f).

the data table and 1,447 observations (bacterial isolates). These 15 dimensions and 1,447 observations were represented by three principal components. These components describe 77% of the data; this is analogous to R^2 in regression analysis and represents the percentage of the total variation of MICs explained by the PCA model. Therefore, the data from the 15 dimensions are well summarized and interpretable in only 3 dimensions.

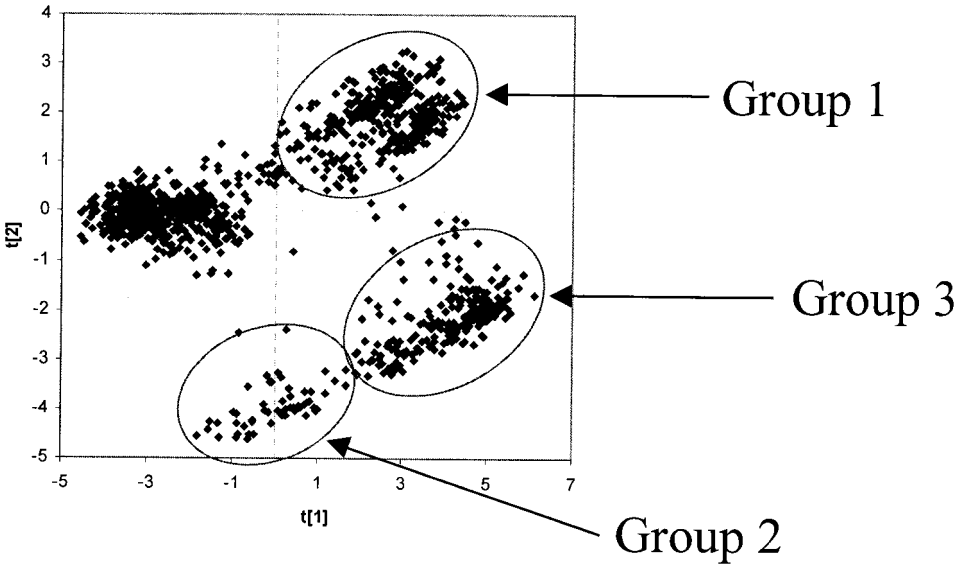

FIG. 3. PCA score plots showing log MICs for major clusters of isolates from Spain from 1992 to 1997.

The loadings plot for the first two principal components for *S. pneumoniae* in Spain from 1992 to 1997 is shown in Fig. 1. All β-lactams are clustered in the upper right-hand corner; thus, their MICs are strongly correlated. The co-trimoxazole MICs are closely correlated with the β-lactam MICs. The macrolide MICs are also highly correlated and are located in the lower part of the graph. The MICs of chloramphenicol and doxycycline are closely related to each other but are dissimilar from those of the other classes of antibiotics. The MICs of the quinolones are clearly unique and, being close to the origin, exert no influence on the first two components. From the loadings one can interpret that the first component separates low from high MICs and the second component separates isolates for which MICs are high by antibiotic class.

The PCA score plots for log MICs for isolates from Spain from 1992 to 1997 are shown in Fig. 2a to f. The score plots are interpreted by relating the position of observations in the plots (which represent individual bacterial isolates) to the positions of variables in the loadings plot. For example, the upper right quadrant represents isolates for which β-lactam and co-trimoxazole MICs are high. That is, the farther toward the upper right-hand corners of the plots, the higher MICs tend to be for β-lactams and co-trimoxazole. The macrolide MICs for these isolates are low and the doxycycline and chloramphenicol MICs are low to increased because these drugs are located in different regions in the loadings plot. Focusing on the isolates from 1992, the MICs of the full spectrum of antibiotics for the cluster of isolates just to the left of the origin are low. One can see a gradual increase in MICs in the direction of the β-lactams and co-trimoxazole until one finally reaches a fairly significant cluster of high MICs. Note that the cluster actually consists of two bands. These bands are separated by co-trimoxazole; for the isolates at the upper band the β-lactam MICs are high and co-trimoxazole MICs are moderate, whereas for the isolates at

TABLE 5. MIC groupings over time[a]

Yr	% of isolates		
	Group 1	Group 2	Group 3
1992	29.4	1.5	7.4
1993	29.6	2.9	10.1
1994	23.9	4.8	20.1
1995	27.5	5.5	20.1
1996	21.3	3.7	15.4
1997	20.6	6.3	25.7

[a] Group 1, isolates for which β-lactam MICs are high; group 2, isolates for which macrolide MICs are high; group 3, isolates for which the MICs of all antimicrobials are high.

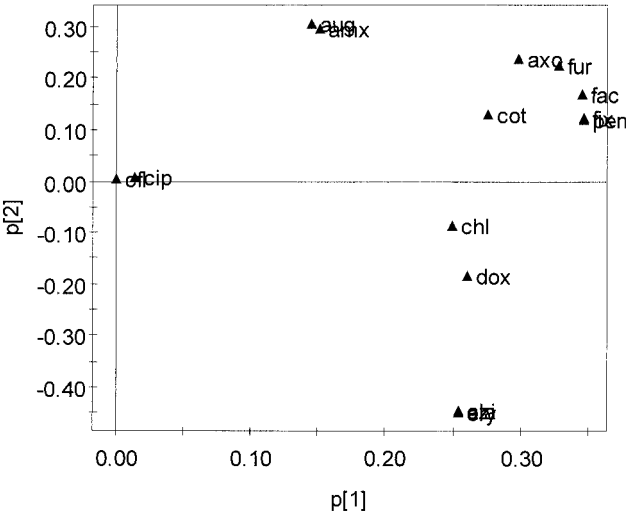

FIG. 4. PCA loadings plot for antibiotypes of isolates from Spain from 1992 to 1997. p[1], component 1; p[2], component 2; cot, co-trimoxazole; chl, chloramphenicol; ofl, ofloxacin; cip, ciprofloxacin; dox, doxycycline; amx, amoxicillin; aug, Augmentin (amoxicillin-clavulanic acid); axo, ceftriaxone; fur, cefuroxime; fac, cefaclor; pen, penicillin; fix, cefixime; the cluster at the bottom represents macrolides.

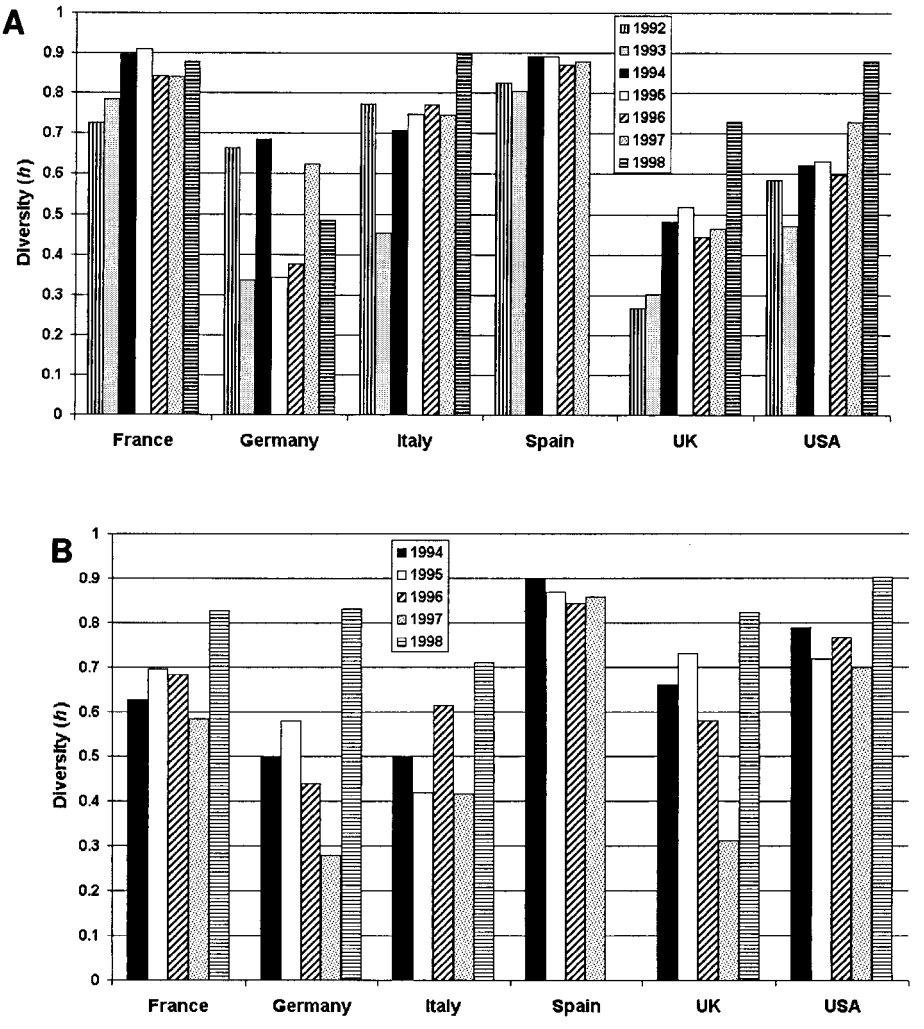

FIG. 5. Histograms of diversity estimates (h) for *S. pneumoniae* (A) and *H. influenzae* (B) antibiotypes. The method of Nei and Tajima (20) was used to calculate h, where 0 and 1 represent low and high levels of genetic diversity, respectively. For both species, no data were collected for Spain in 1998.

the lower band the MICs of both classes of drugs are high. Isolates for which macrolide MICs are high are at the very bottom of the plot. In 1992 macrolide MICs were high for only three or four isolates. Note the all-or-nothing effect; a lot of white space between the cluster of isolates for which macrolide MICs are low and those for which macrolide MICs are high indicates that the macrolide MICs for the isolates go from low to high with little in the way of intermediate MICs. As one moves counterclockwise from the bottom, a group of isolates is picked up and this group is characterized first by increased doxycycline and chloramphenicol MICs and eventually increased MICs of all compounds. Therefore, one can conclude that in 1992 the MIC elevations in Spain are characterized mainly by increases in the MICs of β-lactams and co-trimoxazole. There are very few isolates for which macrolide MICs are increased, and isolates for which MICs of all classes of drugs were increased are rare. The evolution of MICs in subsequent years is readily interpreted.

A few observations can be made. (i) The population of isolates for which macrolide MICs are high increases over the years, and this population is always associated with an all-or-nothing type of effect; (ii) the two bands in the β-lactam area, characterized by co-trimoxazole, dissipate; (iii) the all-or-nothing effect becomes evident for the β-lactam group by 1996-1997; and (iv) the group for which the MICs of multiple drugs are high builds up over time. The fourth point can be summarized roughly from a plot in which data for all years are plotted and three major groups are highlighted (Fig. 3). In Fig. 3, group 1 corresponds to the cluster for which β-lactam and co-trimoxazole MICs are increased, group 2 corresponds to the cluster for which macrolide MICs are increased, and group 3 corresponds to the cluster for which the MICs of all compounds are increased. The population of isolates in each group is quantitated in Table 5.

The third principal component represents the quinolones. For the sake of brevity an extensive discussion of this result is not included. However, evaluation of the third principal component illustrates a clear separation of quinolone activity and

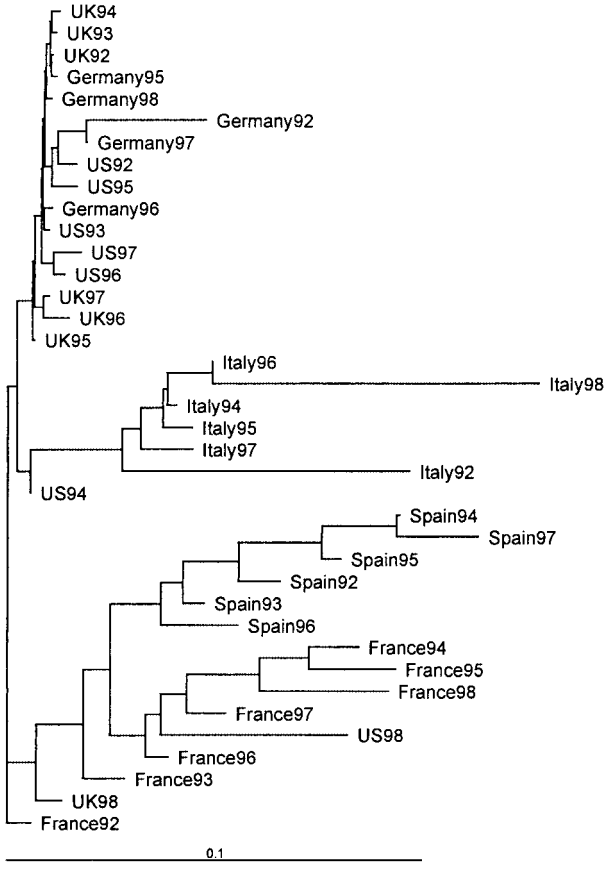

FIG. 6. Fitch-Margoliash phylogenetic tree determined from the *S. pneumoniae* susceptibility data; branch lengths are drawn proportional to the amount of genetic change.

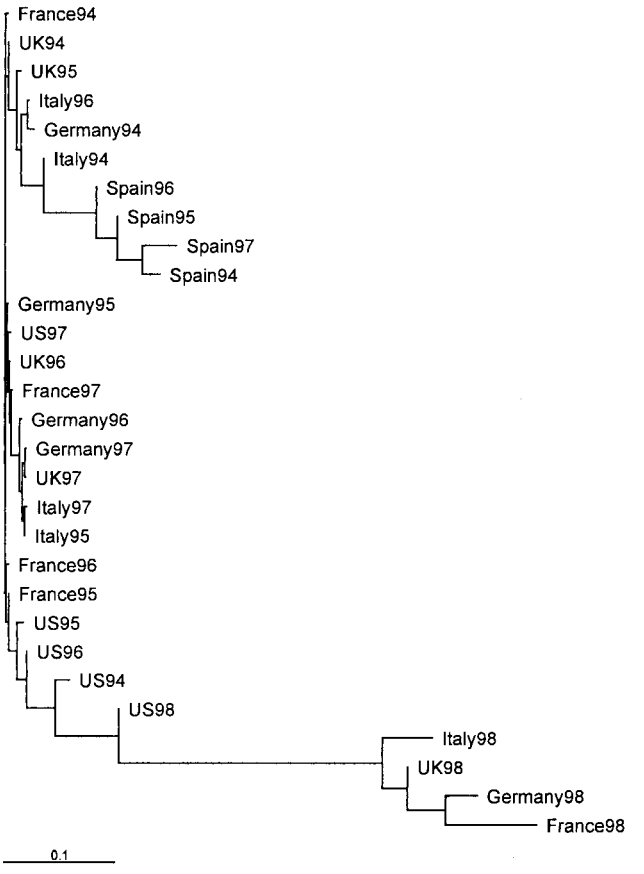

FIG. 7. Fitch-Margoliash phylogenetic tree determined from the *H. influenzae* susceptibility data; branch lengths are drawn proportional to the amount of genetic change.

enables the identification of trends in quinolone MICs as well as the identification of isolates for which quinolone MICs are extremely low and extremely high.

The methodology described above can also be applied to the antibiotypes, and expectedly, the results are similar, with some notable exceptions. First, the loadings plot for the first two components illustrates that the clusters for the antibiotic groups are similar to those obtained by evaluation of log MICs, with the exception that the results for amoxicillin and amoxicillin-clavulanic acid do not correlate as strongly with those for the rest of the β-lactam group (Fig. 4). Thus, at the antibiotype level, the results for amoxicillin and amoxicillin-clavulanic acid are distinct from those for the other β-lactams. This distinction is due to increased susceptibility to both drugs.

As described above, the third component describes quinolone resistance. In the case of the antibiotype data, a fourth component is required, and this component clearly separates the unique isolates resistant to amoxicillin and amoxicillin-clavulanic acid. The four components together explain over 80% of the variation in antibiotypes.

Population genetics. Estimates of haplotypic diversity (h) for *S. pneumoniae* ranged from 0.27 (United Kingdom, 1994) to 0.90 (France, 1995). The order of increasing diversity (mean ± standard deviation h) by country was as follows: United Kingdom ($h = 0.49 \pm 0.15$), Germany ($h = 0.50 \pm 0.15$), United States ($h = 0.64 \pm 0.13$), Italy ($h = 0.72 \pm 0.14$), France ($h = 0.84 \pm 0.06$), and Spain ($h = 0.86 \pm 0.04$). The order of increasing diversity (mean ± standard deviation h) by year is as follows: 1993 ($h = 0.53 \pm 0.21$), 1992 ($h = 0.64 \pm 0.20$), 1996 (0.65 ± 0.21), 1995 ($h = 0.67 \pm 0.22$), 1997 ($h = 0.71 \pm 0.15$), 1994 ($h = 0.71 \pm 0.16$), and 1998 ($h = 0.77 \pm 0.18$). Antibiotype diversity remained generally high among isolates from France, Italy, and Spain throughout the sampling period (Fig. 5A). A trend of annual increases in antibiotype diversity was clearly evident for isolates from the United Kingdom and the United States. Nearly all countries showed marked increases in antibiotype diversity in 1998.

H. influenzae antibiotype diversity values showed a trend markedly different from those for *S. pnuemoniae* by country. The order of increasing diversity (mean ± standard deviation h) by country was as follows: Germany ($h = 0.53 \pm 0.20$), Italy ($h = 0.53 \pm 0.13$), the United Kingdom ($h = 0.62 \pm 0.19$), France ($h = 0.68 \pm 0.09$), the United States ($h = 0.78 \pm 0.08$), and Spain ($h = 0.87 \pm 0.02$). The order of increasing diversity (mean ± standard deviation h) by year was as follows: 1997 ($h = 0.53 \pm 0.23$), 1996 ($h = 0.65 \pm 0.14$), 1994 (0.66 ± 0.16), 1995 ($h = 0.67 \pm 0.15$), and 1998 ($h = 0.82 \pm 0.07$). Annual trends in *H. influenzae* antibiotype diversity values were either constant or slightly multimodal in all years except 1998, in which diversity values clearly increased (Fig. 5B). However, in

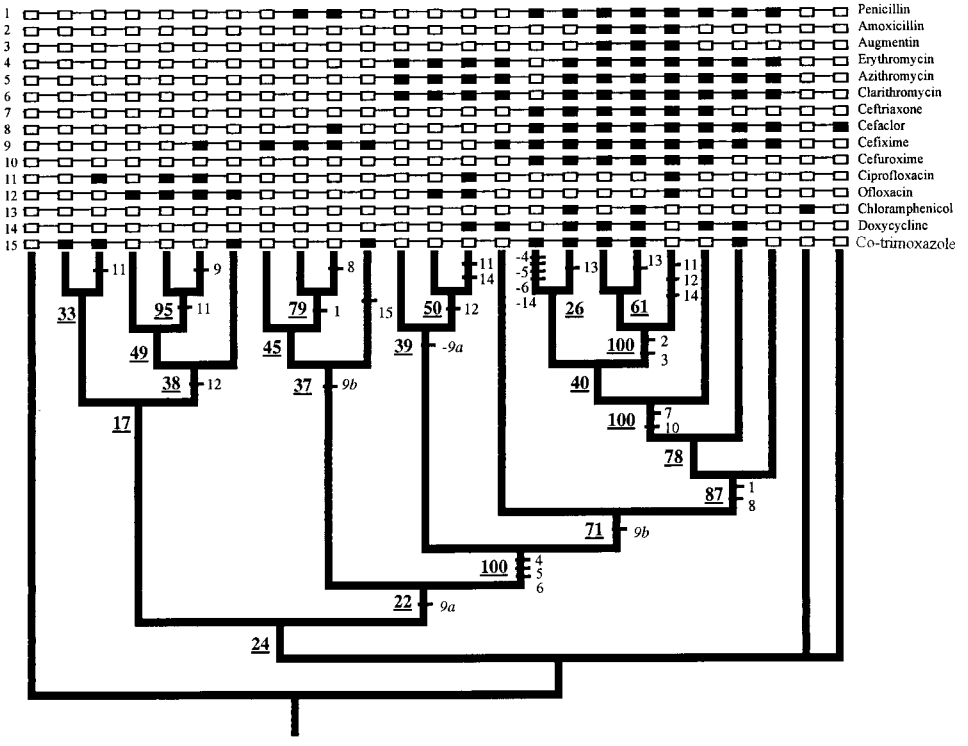

FIG. 8. Majority consensus of the set of most parsimonious trees determined from analysis of *S. pneumoniae* isolates from the United Kingdom from 1998. White boxes, susceptibility to the indicated antibiotic; black boxes, resistance to the indicated antibiotic. The numbers on the branches adjacent to the dashed markers refer to the origin or loss of characters, where the characters are the 15 antibiotics. A minus sign indicates the loss of a character, and a lack of a minus sign indicates the origin of that character. Underlined numbers in boldface type refer to the percentage of times that a particular node appears in the set of most parsimonious trees. Augmentin, amoxicillin-clavulanic acid.

nearly all years, Spain and the United States had the highest diversity values.

Population phylogenetic trees (Fig. 6 and 7) drawn from genetic distance values computed with the antibiotype frequency data resulted in certain clusters of years and countries. For example, in the *S. pneumoniae* data, Spain formed a monophyletic group composed of years 1992 to 1997 and was joined by a group composed of predominately French isolates of various years (Fig. 6). Isolates from Italy also formed a monophyletic group to the marked exclusion of isolates from all other countries. The other principal grouping in this *S. pneumoniae* population tree was a mixed clade of isolates from Germany, the United Kingdom, and the United States. Branch lengths for isolates from France, Spain, and Italy were much longer than those for isolates from the other countries, suggesting increased rates of antibiotic resistance evolution in those countries. The respective monophyletic clusters of the isolates from the various years from Spain and Italy suggest the possibility that *S. pneumoniae* isolates in these two countries comprise genetic races distinct from those in the rest of the world.

The tree for *H. influenzae* (Fig. 7) suggested the same distinction of isolates from Spain compared with those from the rest of the world; isolates from Spain from all years together formed a monophyletic group. Unlike the data for *S. pneumoniae*, *H. influenzae* isolates from France and Italy did not tend to form monophyletic groups. The topology of this tree suggests that the marked increase in the rate of antibiotic resistance in 1998 for isolates from all countries had its origin in the United States; the isolates from 1998 form a monophyletic group, with isolates from the United States at the base of this radiation. The long branch length separating United States isolates of 1998 and European isolates of 1998 compared to the branch lengths for isolates from within Europe suggests a large and/or rapid change in the development of antibiotic resistance patterns in the process of the transatlantic colonization, followed by more moderate amounts of change in the subsequent diversification throughout Europe.

Evolution of antibiotic resistance. An MP analysis of the antibiotype data for *S. pneumoniae* isolates from the United Kingdom from 1998 resulted in 11,632 MP trees each with a length of 39, with a consistency index (CI) of 0.3846 and a retention index (RI) of 0.7551. A majority consensus tree of the set of MP trees (Fig. 8) resulted in several clades that appeared in the entire set of MP trees and several further groupings that appeared in at least 70% of the set of MP trees. This obvious structure to the data concomitant with the moderately high and very high values for CI and RI, respectively (7), indicates that the data provide phylogenetic signal. Additionally, more statistics-based measures, such as relative apparent synapomorphy analysis (14) and g1 (skewness) statistics (10), also support the presence of phylogenetic signal in these data (tRASA = 1.951 [$P < 0.05$]; g1 = −0.406671 [$P < 0.01$]). Furthermore, the score of the MP tree (score, 39) is much less

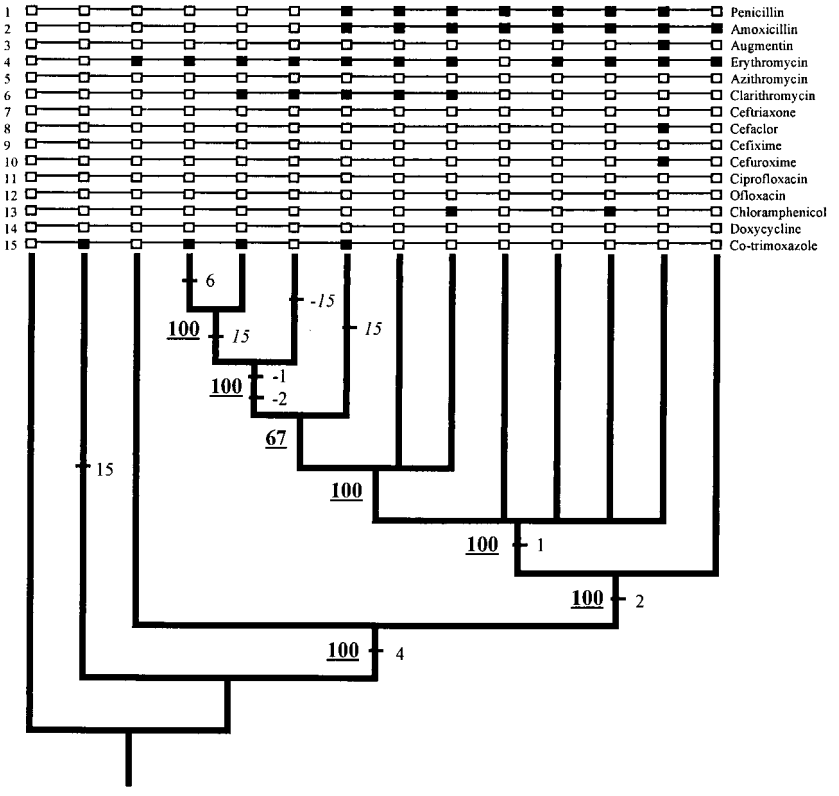

FIG. 9. Majority consensus of the set of most parsimonious trees determined from analysis of *H. influenzae* isolates from the United Kingdom from 1998. See the legend to Fig. 8 for an explanation of the various tree descriptors. Augmentin, amoxicillin-clavulanic acid.

than the mean score of a set of 100,000 random trees (score, 94.7), and there is a marked skewness to the distributions in tree lengths obtained with these data (the basis of the g1 statistic), which in turn is widely regarded as an indicator of a phylogenetic signal (5, 8, 10). This view is based primarily on the fact that such skewed distributions of tree length allow a much better discrimination among the near-optimal solutions.

The first type of antibiotic resistance to evolve from the state of susceptibility to all antimicrobials in the United Kingdom in 1998 is not entirely clear but most likely involves resistance to either cefaclor or chloramphenicol (Fig. 8). Penicillin resistance evolved much later on (character state 1 in Fig. 8) and was preceded by the development of resistance to macrolides. This particular tree in fact predicts that the development of multidrug resistance (i.e., resistance to at least three different antibiotics) is preceded or is perhaps even somehow driven by the development of macrolide resistance (node indicated with character states 4, 5, 6 in Fig. 8). The descendants from this node retain macrolide resistance, while they progressively acquire additional forms of antibiotic resistance. Stepping back one node, one of the two equally most parsimonious interpretations for the character state changes would place the development of cefixime resistance at the node indicated by *9a* in Fig. 8. Under this interpretation cefixime resistance would be retained in the main line of descent but would be lost on the branch indicated by *−9a*. The alternative, equally parsimonious interpretation would have cefixime resistance evolving on two independent occasions, both indicated as 9b in Fig. 8. Thus, the question of which type of antibiotic resistance might be precipitating the development of multidrug resistance depends, at least in this example, on where one draws the ancestral reference. Nonetheless, the tree does not support the view that penicillin resistance alone is somehow driving multidrug resistance, because cefaclor resistance (character state 8 in Fig. 8) evolves at the same node, and furthermore, resistance to both macrolides and cefixime precedes the development of resistance to penicillin.

MP analysis of *H. influenzae* isolates from the same country and the same year (United Kingdom, 1998) (Fig. 9) yielded three MP trees each with a length of 16, with a CI of 0.5625 and an RI of 0.6667. The small number of MP trees and the relatively high values for CI and RI support the presence of phylogenetic signal in these data. Relative apparent synapomorphy analysis and g1 statistics also support the presence of phylogenetic signal in this data set (tRASA= 4.34 [$P < 0.0001$]; g1 = −0.306106 [$P < 0.01$]). The mean score of a set of 100,000 random trees constructed from these data is 25.5 (which differs from the value for the MP tree by 9 steps), and the distribution of tree lengths is markedly skewed, providing yet further support for the presence of phylogenetic signal. In this case the first type of antibiotic resistance to evolve from the condition of susceptibility to all antimicrobials is resistance to co-trimoxazole. Other trends evident in the resulting tree include the evolution of amoxicillin resistance before ampicillin resistance, and the two of these were preceded in turn by the development of resistance to erythromycin.

In summary, the present paper proposes novel approaches to presenting and analyzing data from large multinational surveillance studies. By converting the antimicrobial susceptibility patterns either into a string of 0s and 1s or into a smaller two- or three-digit code, new analyses may be performed with these data. For example, when the susceptibility pattern is converted to a string of 0s and 1s, the percentage of isolates susceptible to all of the antimicrobials tested can easily be determined. In addition, the PCA and parsimony approaches may be useful for determining the patterns and origins of antimicrobial resistance. The PCA analysis performed with *S. pneumoniae* isolates from Spain from 1992 to 1997 appear to indicate that the MICs of β-lactams, not including amoxicillin, are increased and may be driving resistance in that population. Parsimony analysis performed with *S. pneumoniae* isolates from the United Kingdom from 1998 suggest that macrolide resistance may precede the development of penicillin resistance in that population. More detailed studies are necessary to confirm the preliminary findings that these new approaches appear to be highlighting.

REFERENCES

1. **Farris, J. S.** 1970. Methods for computing Wagner trees. Syst. Zool. **19:**83–92.
2. **Feil, E. J., E. C. Holmes, D. E. Bessen, M. S. Chan, N. P. J. Day, M. C. Enright, R. Goldstein, D. W. Hood, A. Kalia, D. E. Moore, J. Zhou, and B. G. Spratt.** 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98:**187.
3. **Felmingham, D., and J. Washington.** 1999. Trends in the antimicrobial susceptibility of bacterial respiratory tract pathogens—findings of the Alexander Project 1992–1996. J. Chemother. **11:**5–21.
4. **Felsenstein, J.** 1993. PHYLIP (phylogeny inference package), version 3.5c. Department of Genetics, University of Washington, Seattle.
5. **Fitch, W. M.** 1984. Cladistic and other methods: problems, pitfalls, and potentials, p. 221–252. *In* T. Duncan and T. F. Stuessy (ed.), Cladistics: perspectives on the reconstruction of evolutionary history. Columbia University Press, New York, N.Y.
6. **Fitch, W. M., and E. Margoliash.** 1967. Construction of phylogenetic trees. Science **155:**279–284.
7. **Forey, P. L., C. J. Humphries, I. L. Kitching, R. W. Scotland, D. J. Siebert, and D. M. Williams.** 1992. Cladistics: a practical course in systematics. Oxford University Press, New York, N.Y.
8. **Goodman, M., J. Czelusniak, and G. W. Moore.** 1979. Further remarks on the parameter of gene duplication and expression events in parsimony reconstructions. Syst. Zool. **28:**379–385.
9. **Hennig, W.** 1966. Phylogenetic systematics. University of Illinois Press, Urbana.
10. **Hillis, D. M., and J. P. Huelsenbeck.** 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Heredity **83:**189–195.
11. **Holmes, E. C., R. Urwin, and M. C. J. Maiden.** 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol. Biol. Evol. **16:**741–749.
12. **Kluge, A. G., and J. S. Farris.** 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. **18:**1–32.
13. **Kuhner, M. K., and J. Felsenstein.** 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11:**459–468.
14. **Lyons-Weiler, J., G. A. Hoelzer, and R. J. Tausch.** 1996. Relative apparent synapomorphy analysis (RASA). I. The statistical measurement of phylogenetic signal. Mol. Biol. Evol. **13:**749–757.
15. **Maddison, W. P., and D. R. Maddison.** 1992. MacClade, version 3.0. Sinauer, Sunderland, Mass.
16. **Morrison, D. F.** 1990. Multivariate statistical methods, 3rd ed. McGraw-Hill, Hightstown, N.J.
17. **National Committee for Clinical Laboratory Standards.** 2000. Performance standards for antimicrobial susceptibility testing, 10th informational supplement. Approved standard M100-S10. National Committee for Clinical Laboratory Standards, Wayne, Pa.
18. **Nei, M.** 1972. Genetic distance between populations. Am. Natur. **106:**283–292.
19. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.
20. **Nei, M., and F. Tajima.** 1981. DNA polymorphisms detectable by restriction endonucleases. Genetics **97:**145–163.
21. **Spratt, B. G., and M. C. J. Maiden.** 1999. Bacterial population genetics, evolution and epidemiology. Phil Trans. R. Soc. Lond. Ser. B **354:**701–710.