New Perspectives on Microbial Typing
Data management and Data Analysis

João André Carriço

Dissertation presented to obtain the Doutoramento (Ph.D.) degree in Biology at the Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa

Apoio financeiro da FCT e do FSE no âmbito do
Quadro Comunitário de Apoio,BD nº 3123/2000
New perspectives on Microbial Typing

ISBN: 989-20-0252-0
ISBN (From Jan 07): 978-989-20-0252-1

PhD Thesis Public Discussion in Biology
João André Carriço
16 June 2006 , at 14h30

President of the Jury

Prof. Dr. Peter Frank Lindley
President of the Scientific Council
Instituto de Tecnologia Química e Biológica (ITQB)
Universidade Nova de Lisboa, Portugal

Promoters

Prof. Dr. Jonas Silva Almeida
Coordinator of the Biomathematics Group at
Instituto de Tecnologia Quimica e Biologica (ITQB)
Oeiras, Portugal
and
Professor of Bioinformatics at the
Dept Biostatistics and Applied Mathematics
Univ. Texas MDAnderson Cancer Center
Houston, Texas, USA

Prof Doutora Herminia de Lencastre
Head of Laboratory of Molecular Genetics
Instituto de Tecnologia Quimica e Biologica (ITQB)
Oeiras, Portugal

Members of the Juri

Prof. Hajo Grundmann, MD, MSc, DTM&H
Project Leader, Scientific Coordinator for European Antimicrobial Resistance Surveillance System
National Institute of Public Health and Environment (RIVM)
Bilthoven, The Nederlands

Prof. Karl Ekdhal MD, PhD, MSc, DTM&H,
Assistant. Professor,Strategic Advisor to the Director
European Centre for Disease Prevention and Control (ECDC)
Stockholm, Sweden

Professor Doutor Jorge Carneiro
Head of Theoretical Immunology Group
Instituto Gulbenkian de Ciência (IGC)
Oeiras, Portugal

Professor Doutor Mário Ramirez
Head of Laboratory of Molecular Microbiology and Infection
Institute of Molecular Medicine, Faculty of Medicine of Lisbon
Lisbon, Portugal

From left to right: Peter Frank Lindley, Mario Ramirez, Jonas Almeida, Karl Ekdhal, João Carriço, Hajo Grundmann, Hermínia de Lencastre, Jorge Carneiro

“Crude classifications and false generalizations are the curse of organized life.”

George Bernard Shaw (1856 - 1950)

Acknowledgements

Writing acknowledgments is always a complicated matter since I start writing them with the feeling that I will forget someone important. So my first acknowledgment and a sincere apology should be made for the ones I will forget in the too few paragraphs that follow.

This thesis and all the work done over the last six years couldn't have been possible without the contribution and support of a series of extraordinary individuals and institutions.

I would like to start by thanking to my PhD supervisor, Prof. Jonas Almeida, for always pointing to me that making science is fun and never putting any barriers to theories and imagination in our discussions. Thanks to him, I now realize that science IS a fun thing to do. I'm sure that we will have much more fun discussions ahead! I thank him also for his trust, entrepreneurship, friendship and always making me feel at home when I was abroad.

Also I would like to thank to my co-supervisor, Prof. Hermínia de Lencastre, for the opportunity to work in a very productive and interesting scientific field that now is in my everyday thoughts. Thanks for all the trust and guidance provided during these last years, and for being so open-minded to the new methods and technologies.

I thank all my colleagues of the Biomathematics group: present and past ones. It has been "quite a ride" and they were (and still are and I hope that will be) the best partners I could have had in this "road trip" through Biomathematics, Computational Biology and Bioinformatics and science in general. To Dominick Beck, Andreas Bohn, Jaime Combadão, Helena Deus, António Maretzek, Sara Garcia, Rodrigo Gouveia-Oliveira, Isabel Oliveira, Francisco Pinto, João Quenino, Sara Silva, Susana Vinga and João Xavier my many thanks. A special thanks needs to be addressed to two of them: To Francisco Pinto, for always being the "guy with the cup half full" and always supporting my work with this positive attitude and knowledge; and to António Maretzek, for having the patience to teach me all about Linux systems management and general programming skills and showing me how always to do the right thing when it comes to computers. His enormous contribution only pales when compared to his friendship and support.

Since the work presented in this thesis is always based on data analysis, it couldn't have been made possible without the ones who ultimately spent hours on the lab bench to produce the valuable data. I profoundly thank the all the people of Laboratório de Genética Molecular (under the guidance of Prof. Hermínia de Lencastre) specially those I worked with more closely: Nelson Frazão, Maria Miragaia, Sónia Nunes, Carla

Rute Alves, Carla Simas, Natacha Sousa and Raquel Sá-Leão. I would also thank Prof. Ilda Santos-Sanches and Dr. Rosário Mato for all the help provided and interesting discussions.

More recently I also started to analyze the precious data from Unidade de Microbiologia Molecular and Infecção from Instituto de Medicina Molecular (under the guidance of Prof. Mario Ramirez and Prof. José Melo-Cristino). I thank all the people from the lab especially Ana Catarina Costa and Isa Serrano. Special thanks should be made to Prof. Mario Ramirez, for his contagious inquisitiveness and scientific curiosity he instilled in me and also all the friendship and support in the last years.

I would also like to thank to Prof. Alexander Tomasz for my stay at The Rockefeller University during the EURIS project PFGE workshop. It was a very productive time and Prof. Tomasz always demonstrated that how the simplest explanation of an event is the best place to start.

I would like to thank also to the Theoretical Immunology Group at Instituto Gulbenkian de Ciência, led by Prof. Jorge Carneiro, for all those interesting discussions and suggestions done during our shared Estudos Avançados de Oeiras (EAO) meetings. I thank Daniel Ferreira, Rui Gardner, Tiago Paixão, Nuno Sepúlveda and Prof. Jorge Carneiro for all the fun discussions we had.

I thank the people of Instituto de Tecnologia Química e Biologica (ITQB) and Instituto de Biologia Experimental e Tecnologica (IBET), from colleagues (too many to name) to workers of both Institutions who always helped me when I needed: Cristina Amaral, Lurdes Conceição, Isabel Ribeiro, Cristina Simão,Catarina Simões , Manuela Nogueira and Ana Maria Portocarrero.

I acknowledge the financial support from Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e do Ensino Superior, Portugal, through a PhD grant SFRH/BD/3123/2000 and the European Science Foundation for sponsoring the attendance of the ENEMTI (ESF Network for Exchange of Microbial Typing Information) workshops.

I also thank all my friends that always provided support and endured my complaints and frustrations in the last few years: Luis Reis, who always helped me with programming and showing me the new things in informatics since a long time ago; Eduardo Alves, Ricardo Araújo e Tiago Carita who were always there for me when I needed even when our private lives had lead us apart; Frederico Castelo, he has been a “brother in arms”, always supportive and full of energy; My paintball team, Trolls, António Frazão, António Ribeiro, António Santos, David Mourato, Ivo Carola, Jaime Menino, José Vieira, Pedro Botelho, Pedro Alexandre and Robert Mendes, for all the support and fun moments we had together over the last 4 years.

And finally, I would like to thank to the most important people for me (although it doesn't always show) and to whom I dedicate this thesis: my family. To my parents, Laura e José Carriço, for always supporting me and making me try better myself; my grandparents, Isabel e Manuel Nogueira for educating me in my early years; my aunts Manuela Nogueira, Maria de Lurdes Oliveira e Isabel Carriço, for always helping me when I needed; my brother, Pedro Carriço, who had to endure me as older brother with all the things that come with that "job"; my cats, Freddy, Mike and Fatma, as they are a part of my family too, for their calming presence and fun moments.

Finally I would like to make the most special thanks to Filipa Pereira, my companion for the last ten years, providing me all the love and friendship I could wish for.

Summary

A large number of methods are available to type microbial pathogens. These methods provide a phenotypic or genotypic characterization about the strains under study and may allow, together with collected epidemiological data, the inference of clonal relationships between isolates. This collected information about the strains makes possible studies on different subjects: bacterial population genetics, pathogenesis and natural history of infection, surveillance of infectious diseases and outbreak investigation and control.

Nowadays, databases of strains characterized with a plethora of typing methods are appearing all over the world, providing researchers with material to conduct the aforementioned studies. The novel challenge resides in the combined data analysis of such large numbers of typing and epidemiological data, since the conventional methods of analysis were developed for studies with fewer than one hundred strains.

In this thesis we present a series of articles in which we address this novel challenge of large scale epidemiological data storage and the development of new methods for the analysis of those data.

This thesis is organized in the following structure:

Chapter I – *Introduction* – Presents background information on microbial typing methods and related data analysis techniques.

Chapter II - *EURISWEB: Web-based epidemiological surveillance of antibiotic-resistant pneumococci in Day Care Centers* – this article presents an online database developed for the 5th Framework European project EURIS (European Resistance Intervention Study), demonstrating a multi-national and multi-centric database where strains are included with demographic information about their carriers as well as typing data. This database was constructed as a prototype for a fully-fledged Epidemiological Information System.

Chapter III – *New developments on EURISWEB* – in this chapter we present the latest developments on the EURISWEB online database and its expansion for the 6 th Framework project PREVIS (Pneumococcal Resistance Epidemicity and Virulence - An International Study).

Chapter IV - *Assessment of band-based similarity coefficients for automatic Type/Subtype classification of microbial isolates analyzed by Pulsed-Field Gel Electrophoresis* – In this article we present a methodology based on receiver operating characteristic (ROC) curves for assessment of commonly used band-based similarity

coefficients for type classification using an accepted criteria applied to a large Pulsed-Field Gel Electrophoresis band patterns collection.

Chapter V - *A common framework for relating multiple typing methods illustrated using macrolide- resistant Streptococcus pyogenes* – in this article, we demonstrated the usefulness of a framework of measures that quantitatively addresses two important questions namely, 1) could the results of a given typing method have been predicted from the results of another? and 2) how does a novel typing method relate to previously used typing schemes?

Chapter VI – *Final discussion* – In this final chapter we bring together the conclusions of the previous chapters and discuss further uses of the novel methodology presented here. Also future developments in the area are discussed.

This thesis presents work described in the following publications:

Silva, S., et al., *EURISWEB--Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers.* BMC Med Inform Decis Mak, 2003. 3(1): p. 9.

Carrico, J.A., et al., *Assessment of band-based similarity coefficients for automatic type and subtype classification of microbial isolates analyzed by pulsed-field gel electrophoresis.* J Clin Microbiol, 2005. 43(11): p. 5483-90.

Carrico, J.A, et al , *A common framework for relating multiple typing methods illustrated using macrolide- resistant Streptococcus pyogenes,* accepted for publication in J Clin Microbiol, *in press*.

Contents

Chapter I

1. Introduction

The Merriam-Webster online dictionary defines “classification” as the “systematic arrangement in groups or categories according to established criteria”. When associated with biology a definition commonly found is “the systematic grouping of organisms into categories on the basis of evolutionary or structural relationships between them”. The ability to classify microorganisms at strain level is as paramount for molecular epidemiology studies as it is for population genetics studies. Microbial typing methods are the tools that provide researchers with criteria to do that classification. In this chapter, we provide some definitions important to the field and describe briefly some of the most common typing and data analysis methods used to recognize a type.

1.1. From isolates to clones

Tenover *et al* (44), proposed a series of definitions for terms commonly used in epidemiological typing, that were recommended and adapted by the European Study Group on Epidemiological Markers (42). Of these definitions, four are of major importance for this report:

- **Isolate** - Population of microbial cells in pure culture derived from a single colony on an isolation plate and characterized by identification to the species level.
- **Strain** – Isolate or group of isolates exhibiting phenotypic and/or genotypic traits which are distinctive from those of other isolates of the same species.
- **Type** - A specific pattern, or set of marker scores, displayed by a strain on application of a particular typing system.
- **Clone** - An isolate or group of isolates presumed to descend from a common precursor strain by nonsexual reproduction exhibiting phenotypic or genotypic traits characterized by one or more strain-typing method to belong to the same group.(Adapted from (34) and (44))

The relationships derived from these definitions are represented in Figure 1. The definition of clone aggregates all the other definitions but it based on the assumption that an isolate characterized at strain level by one or more typing methods is capturing a phylogenetic signal, allowing the inference of clonality. So the definition of clone can vary depending on the typing method used and how the data analysis is performed.

This is especially true when the assignment to a given type depends on a decision of a cut-off at a given similarity level (such as the threshold selection for type definition in Pulsed-Field Gel Electrophoresis (PFGE) based on the number of band differences or on a dendrogram).

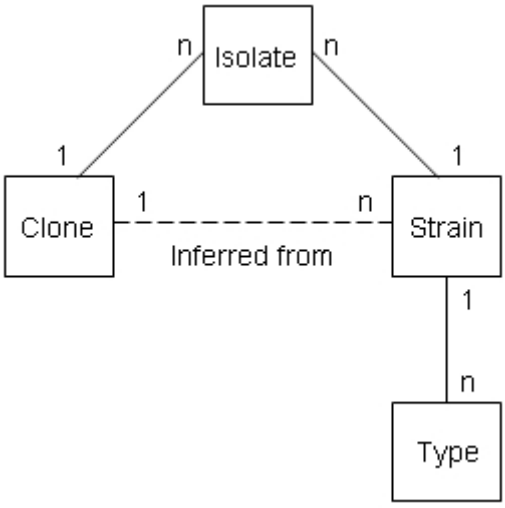


Figure 1- Data Model describing relationships between Isolate, Strain, Type and Clone definitions. The letters and numbers represent the cardinality of the relationships, i.e, a Strain can define a group of one or more Isolates (cardinality n) while each isolate can be assigned only to a single Strain (cardinality 1).

This flexibility for defining clone creates the need of standardization of the data analysis of microbial typing methods, if inter-study comparisons are to be performed.
Ideally, to achieve maximum ease of interpretation of results, there should be a direct equivalence from a clone to a single strain defined by a selection of typing methods.

1.2. Microbial typing methods

To achieve classification at strain level a plethora of microbial typing methods are available to the researchers. The ultimate goal when applying those methods is to discriminate epidemiological unrelated isolates and trying to quantify relatedness between those assumed to be clonally related.
Advances in the ability to discriminate and classify isolates at strain level, were always driven by technological developments. The first typing methods used were based on the phenotypic characteristics of the isolates, such as antimicrobial resistance, phage lysis of isolates (phage typing), biochemical tests (biotyping) or antigenic determinants (serotyping). Advances on molecular biology techniques, allowed new methodologies that probe for characteristics at the level of the bacterial chromosome. Methods like restriction fragment length polymorphisms (RFLP) (12) and pulsed-field gel electrophoresis (PFGE) (36) provided more discriminatory power than the classical

phenotypic methods, and become standards for epidemiological studies all over the world. With the increasing availability of affordable sequencing methods, another shift occurred towards the use of sequence based typing methods such as multilocus sequence typing (MLST) (25) and *emm* sequence typing (2), among others.

The sequence based methods have a large appeal since they provide unambiguous data and are intrinsically portable, allowing the creation of databases that, if publicly available through the internet, enable the comparison of local data with that of previous studies in different geographical locations.

Although complete description of all the typing methods in use is behind the scope of this report, we now present a description of the most currently used methods from phenotypic, genotypic and sequence based methods, which are also discussed in the next chapters.

1.2.1. Serotyping

Together with antimicrobial resistance profiles, serotyping is currently the most commonly used of the phenotypic typing methods. It is based on the fact that organisms belonging to the same species can express different antigenic determinants on the surface of the bacterial cell. These antigenic determinants include proteins, polysaccharides and lipopolysaccharides. The isolates are tested in an agglutination assay against a pool of known sera. The strain is given a serotype number following a key representing which combination of sera produced cell agglutination.

It remains an essential method for typing isolates of *Salmonella*, *Shigella* and pneumococci.

For the typing of *Streptococcus pneumoniae*, serotyping has demonstrated good discriminatory power (although significantly less than PFGE). Some serotypes were shown to be associated preferentially with invasive disease, and has been proposed that the capsule may have more importance that the genotype in the ability to cause invasive disease(4). Also, an association between certain *Salmonella* serotypes and food-borne disease has been demonstrated (6, 47). This association between serotypes and disease provide a fast way to detect possible outbreaks.

There are several limitations to this technique: different strains from the same species, or even strains for different species, may present cross-reacting antigens on their bacterial cell wall, yielding a false-positive result to more than one serum; some strains do not express antigens on the bacterial cell surface, being classified as non-typable and, maintaining a stock of sera.

1.2.2. Pulsed-field Gel Electrophoresis

The first described application of PFGE was the separation of yeast chromosomes by Schwartz in 1984 (36). The ability to separate DNA fragments of sizes from 10 to 800 Kb, allowed this technique to become the genotypic method of choice for many different bacterial species (e.g. *Streptococcus pneumoniae*(22), *Staphylococcus aureus* (31))

In this microbial typing method, total genomic DNA (ranging from 1.8 - 5 MB), is digested with a rare cutter endonuclease, generating typically 10 to 30 DNA fragments (depends on the endonuclease and on the microorganism). Since this technique is based on chromosomal DNA, it can be applied to all the species for which its isolation is possible. These fragments are resolved by a variation to conventional electrophoresis, where three pairs of electrodes form an hexagon around the gel. This allows periodically changing the orientation of the electric field across the gel, which causes the migration of the DNA to occur in three diferent directions, effectively increasing the distance migrated by each fragment, greatly improving the resolution of this technique. An example of an image of a PFGE gel is presented on Figure 2.

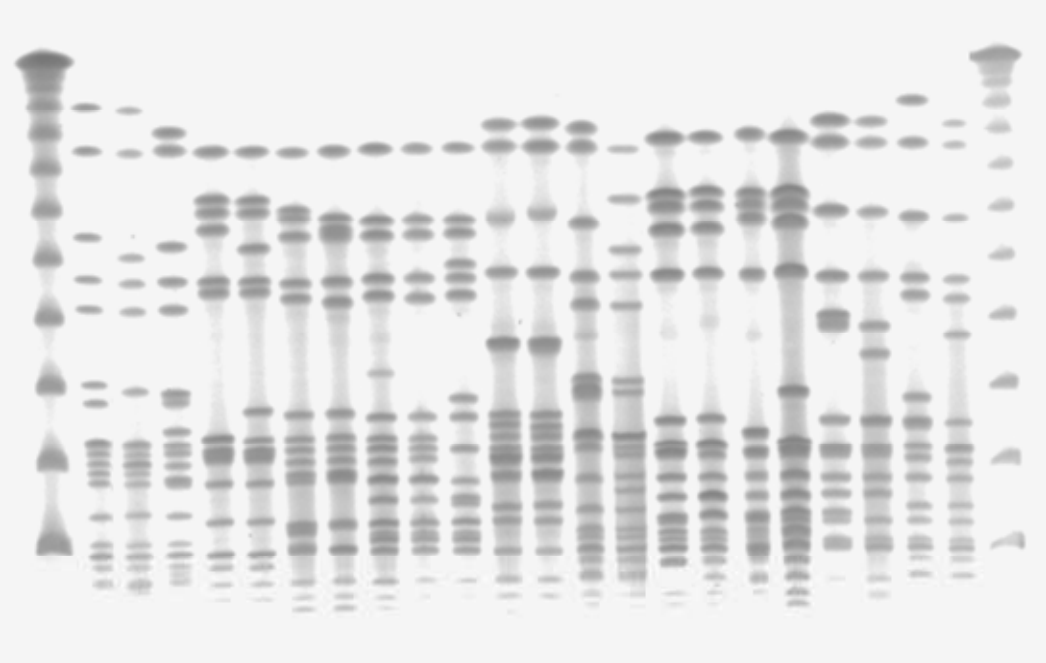


Figure 2 – PFGE gel image for isolates of *Staphylococcus epidermidis* (Adapted from Miragaia *et al* (30)

For each bacterial species, an enormous variety of band patterns has been found with type classification being achieved by the widely used criteria of counting the number of band differences between two lanes proposed by Tenover *et al* (44): if two strains differ up to 6 bands, counted in both lanes, they are considered in the same type and an

arbitrary name is assigned to that type. However, these authors pointed out that this method of classification should be used in outbreak studies only and should be backed up with other relevant typing data, such as antibiotic resistance or other epidemiological relevant data that supports the type assignment. This type classification usually corresponds to clusters generated at a cut-off value at 80% similarity in a Dice(13) /UPGMA(39) dendrogram[1] (19, 29, 37). The ability to measure a similarity for all the strains in a study can be use to classify them by degree of relatedness. In general, the small pattern changes that can be detected by PFGE reflect genetic events (recombination, insertion, mutation or deletion) that occur over a relatively short evolutionary time scale. This fact contributes for its high discriminatory power, and together with its high reproducibility when properly executed, makes this technique very appealing in epidemiological studies.

The limitations of PFGE are all of technical nature: it requires well trained personnel, specialized and expensive electrophoresis apparatus and incomplete restriction of chromosomal DNA can result in misclassification of band patterns. Although these technical problems can arise, high inter-laboratory reproducibility has been reported (7, 31), when standardization protocols are achieved.

1.2.3. Multi Locus Sequence Typing

Multi Locus Sequence Typing (40) is a microbial typing method based upon sequencing ~450-500 base pairs internal fragments of 7 housekeeping genes of a given strain and then assigning to each unique allele a number, by comparing the sequence to an online database (5). The seven number code obtained, designated Sequence Type (ST), is also compared with the online database to obtain the ST assignment .

Since the accurate determination of the sequence of the internal fragments can be reliably done with automated DNA sequencers, MLST has the inter-laboratory portability and accuracy, desired in typing methods used for tracking bacterial populations, while retaining discriminating power.

Because of its characteristics, MLST has become widely used in molecular epidemiology surveillance and microbial population studies. In those fields, the typing method must have the ability to determine the relationship between strains. For MLST, the eBURST algorithm (14) is commonly used, being preferred to dendrogram analysis,

[1] A dendrogram constructed using Dice coefficient for measuring similarity based on band differences and using Unweighted Pair-Group Mean Average (UPGMA) as the criteria creating the clusters.

which provide poor representations of clonal emergence and diversification. The first step of eBURST algorithm is the group creation. Every ST within an eBURST group has a user-defined minimum number of identical alleles (n) (typically n=6, creating the most exclusive group definition) in common with at least one other ST in the group. Group assignment of STs is mutually exclusive: a ST belongs only to a single group. Using this partition method, several groups are created and some have only one ST. These are called singletons, since they share only n-1 or less alleles with other ST in the data set. The second step in the algorithm is the primary group founder determination. The primary founder is predicted on the basis of parsimony as the ST that has the largest number of Single Locus Variants (SLVs: a single allele difference). In case two STs share the same number of SLVs, the one with more Double Locus Variants (DLVs) is considered the founder SLV. The next step in the algorithm is assigning a statistical significance for each of the group founders. This is performed using a bootstrap resampling procedure, where for each group, resampling with replacement is performed a user-defined number of times (typically 1000 times), and then primary founders are re-assigned for each group as previously described. A bootstrap value of 100% would be assigned to a ST considered group founder for all the resamplings.The typical results of eBURST can be visualized on Figure 3.

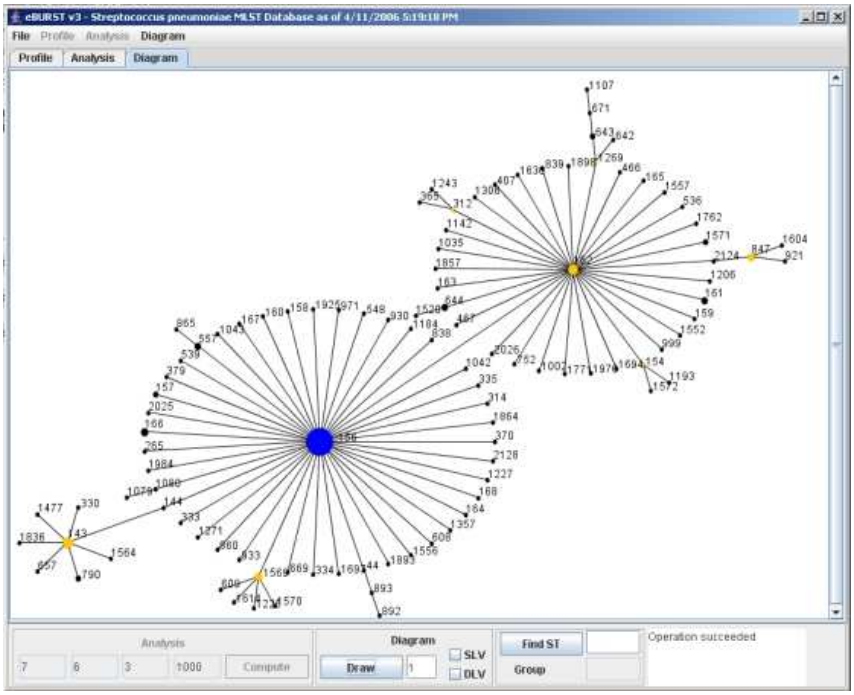


Figure 3 – eBURST diagram for *Streptococcus pneumoniae* clonal complex/group 1. Group founder is represented as a blue dot and subgroup founders as yellow dots. The lines represent a SLV link

These results produce a working hypothesis about the way each clonal complex may have diverged and diversified, but to validate the hypothesis further, phenotypic, genotypic and epidemiological data must be considered.
The relative drawbacks of MLST are the high cost, time, and expertise required for processing each sample. Although the sequencing costs are decreasing they are still often prohibitive for routine application to case isolates.

1.3. Typing methods comparison

Struelens *et al* (42) proposed a series of criteria for the evaluation of the typing methods, such as, discriminatory power (proposed by Hunter(21) as the average probability that a given typing method will assign a different type to two unrelated strains randomly sampled from the population), reproducibility (ability to assign the same type to a strain tested on independent and separated assays), typeability (proportion of the strains assigned to a type by the typing method), and epidemiological concordance (probability that epidemiologically related strains derived from a presumably single-clone outbreak are determined to be similar enough to be inferred to be the same clone).
A survey of these characteristics for several typing methods is presented on **Table** 1 , (adapted from Maslow *et al* (27) and vanBelkum *et a l*(45)). As described it shows that genotypic methods generally have better typeability, reproducibility and discriminatory power, while phenotypic methods have lower cost and better ease of performance and general availability. DNA sequencing methods still have the major drawback of the difficult performance but this greatly improved in recent years.
The suitability of a typing method for a given study depends also on other criteria such as the scale of the study and of the epidemiological markers to be studied. Studies involving a large number of strains to be typed will need large financial resources if genotypic typing methods are to be used. The epidemiological markers depend on the type of study the microbial typing method will be used and is discussed in the next section of this chapter.
Several molecular epidemiology studies of clinically relevant microorganisms provide a characterization of isolates based on different typing methods (9, 12, 26, 33). These studies focus on a comparison between the assigned types of different typing methods, from a qualitative point of view, i.e., indicating correspondences between the types of the different methods. Although this may be useful for the comparison of the genetic backgrounds of the particular set of isolates under study, it does not allow for a broader

view of how the results of the different typing methods are related. Chapter V presents a framework to address the global comparison of typing methods results.

Table 1 – Characteristics of several currently used microbial typing methods (Adapted from (27) and (45))

Typing method[a]	Typeability	Reproducibility	Discriminatory Power	Ease of Performance	Ease of Interpretation	General Availability	Cost
Phenotypic							
Antimicrobial susceptibility	Good	Good	Poor	Excellent	Excellent	Excellent	Low
Manual biotyping	Good	Poor	Poor	Excellent	Excellent	Excellent	Low
Automated biotyping	Good	Good	Poor	Good	Good	Variable	Medium
Serotyping	Variable	Good	Variable	Good	Good	Variable	Medium
Phage Typing	Variable	Fair	Variable	Poor	Poor	Excelent	Medium
MLEE	Excellent	Excellent	Good	Good	Excellent	Variable	High
Genotypic							
Chromosomal REA	Excellent	Variable	Variable	Good	Fair	Variable	Medium
Ribotyping	Excellent	Excellent	Good	Good	Good	Variable	High
PFGE	Excellent	Excellent	Excellent	Good	Good	Variable	High
PCR	Excellent	Fair	Excellent	Good	Fair	Good	Medium
AFLP	Excellent	Good	Excellent	Good	Fair	Low	High
DNA Sequencing	Optimal	Excellent	Excellent	Poor	Excellent	Low	High

[a]MLEE, multi locus enzyme electrophoresis; REA, restriction endonuclease analysis; PFGE, pulsed-field gel electrophoresis; PCR, polymerase chain reaction; AFLP, amplification fragment length polymorphism; DNA sequencing encompasses all the typing methods based on DNA sequencing such as MLST or *emm* typing (2)

1.4. Applications of microbial typing

The inference of clonal relationships between isolates through typing information is used for the study of bacterial population dynamics, from single hosts to entire ecosystems. These studies are divided in the following specific subjects. The frontiers between these subjects are sometimes very subtle as most of the studies aim for a combination of them to achieve their goals.

1.4.1. Bacterial population genetics

Large samples of isolates can be analyzed by typing methods in order to determine intraspecies population structure, and derive phylogenetic hypothesis from the determined structure (14, 41) together with theoretical models of bacterial evolution(17). Also studies of recombination and mutation rates for a species can be performed based on the new sequence based methods, such as MLST (15). Typing methods results can be calibrated with phylogenetic classification obtained from phenotypic markers or nucleic acid hybridization analysis (20) for assessing the validity of the phylogenetic information inferred. The results of these studies provide the knowledge for the definition of clones.

1.4.2. Pathogenesis and natural history of infection

Clinical studies typically use typing methods results for the identification of sources of transmission and patterns of colonization for carriage or invasive disease, (10, 11, 28, 35). These results have great impact in understanding the bacterial epidemiology and are used to design prevention strategies such as vaccines and standard operation protocols.

1.4.3. Surveillance of infectious diseases

Nowadays, the increasingly global nature of economic activity allows a speed of transportation of persons and goods all over the planet that comes with a correspondingly more awareness of the global nature of infectious diseases. The surveillance of infectious diseases can only be achieved by a series of processes, ranging from the initial data collection, its subsequent analysis and interpretation and its dissemination, in order to follow disease frequencies, and identify risk factors in the target population. Surveillance programs can be setup at various levels (regional, national or multi-national). These programs usually target the surveillance of markers associated with pathogenicity like the case of PulseNet, the molecular subtyping network for foodborne bacterial diseases in the United States (43), or drug resistance(3). This can lead to “early warning” systems for potential outbreak detection. These surveillance projects highlighted the need for the standardization of the methods of sample collection, sample processing, and typing methods protocols (1, 7, 31) as well as data storage and analysis (38, 43) .

1.4.4. Outbreak investigation and control.

Outbreaks are the occurrence of a large increase in the frequency of colonization by a given microorganism, over a short period of period of time, with or without increase in morbidity. This is usually caused by an increased rate of transmission of a given pathogen. Microbial typing systems are used in this setting to test a series of hypothesis, ranging from the identification of the clones and the sources of contamination causing the outbreak, to the evaluation of the control measures used to contain the spread of the epidemic clone. In outbreak detection, there is a need of microbial typing methods that provide rapid results, given the need for fast intervention to stop the spread of the outbreak. Further confirmation can be obtained by typing methods with greater discriminatory power if a refinement of results or additional confirmation is needed.

1.5. Recent developments in typing methods

The advances in DNA sequencing technology, led to the availability of complete genomes of several strains of microbial pathogens. Currently, at the National Center for Biotechnology Information (NCBI) website, three hundred and thirty three complete genomes are publicly available(32). With this information accessible, new sequence typing methods (or variations of current ones) that are specific for a single microorganism, such as *spa* typing (18) and *emm* typing (2), can be developed and tailored to the needs of specific study subjects. Also a new discipline, comparative genomics, is arising, that involves the whole-genome sequence comparison. Although the cost and time demands of sequencing whole genomes still makes this technique unsuitable for epidemiological studies on clinical settings, as more genomes are made available, it is becoming evident that some previous assumptions about phylogenetic relationships between strains were not accurate (16). In those cases, the choice of target sequences for sequence-typing methods were genes found to be recently acquired, therefore the inferences made on those results may not reflect the phylogenetic relationships of the strains under study.

Currently microarray technology (23, 24) is one of the tools of comparative genomics. Microarray chips, composed of thousands of DNA fragments of known sequences (probes), are hybridized against whole genomes, and the resulting hybridization profiles are analyzed. Correlations between gene sequences and the epidemiological data can provide new data and the patterns could even be used as epidemiological markers to successfully predict disease outcomes, like it as been used to predict the survival in breast cancer (46). One of the major drawbacks of this technique lies in the

reproducibility of results. The simultaneous analysis of thousands of DNA probes is subject to technical difficulty and results based on a single microarray are not usually considered reliable. So the study design and data analysis strategies of these techniques need to be carefully planned to avoid misleading conclusions.
Another important point is also raised when these new methods generate such an enormous amount of data: How can this data be stored for further analysis? With the availability of high speed internet connections, online databases are becoming widely available for typing data in a multitude of studies (5). This allows the multi-centric comparison of studies where data collection is standardized.

1.6. Thesis structure

As previously mentioned the technological advances in typing methods provided the researchers with the increased capability to generate data. This necessitates extended capacity to store and manage the data and new data analysis methodologies to deal with it. The work presented on this thesis is based on those new concerns.
In chapter II we present the WEBEURIS database originally developed as part of the graduate research on epidemiological information systems described in this thesis. In Chapter III, we report the extensions implemented on WEBEURIS, to further accommodate data from 6th framework project PREVIS (Pneumococcal Resistance Epidemicity and Virulence - An International Study), add some extended data query capabilities and to exchange data with the commercial software package Bionumerics(tm).
In terms of data analysis of large collections of isolates, with the goal of determining the best PFGE gel analysis parameters for type assignment, we propose a methodology based on receiver operating characteristic curves in Chapter IV, applied to a collection of *Streptococcus pneumoniae* but extensible to any other bacteria. Finally in Chapter V, we present a framework of methods for the quantitative comparison of different typing methods type assignments and, as the first step, for mapping type assignments equivalences between different typing methods.

1.7.References

1. **Aires-de-Sousa, M., K. Boye, H. de Lencastre, A. Deplano, M. C. Enright, J. Etienne, A. Friedrich, D. Harmsen, A. Holmes, X. W. Huijsdens, A. M. Kearns, A. Mellmann, H. Meugnier, J. K. Rasheed, E. Spalburg, B. Strommenger, M. J. Struelens, F. C. Tenover, J. Thomas, U. Vogel, H. Westh, J. Xu, and W. Witte.** 2006. High interlaboratory reproducibility of DNA sequence-based typing of bacteria in a multicenter study. J Clin Microbiol **44:**619-21.
2. **Beall, B., R. R. Facklam, J. A. Elliott, A. R. Franklin, T. Hoenes, D. Jackson, L. Laclaire, T. Thompson, and R. Viswanathan.** 1998. Streptococcal emm types associated with T-agglutination types and the use of conserved emm gene restriction fragment patterns for subtyping group A streptococci. J Med Microbiol **47:**893-8.
3. **Bronzwaer, S. L., W. Goettsch, B. Olsson-Liljequist, M. C. Wale, A. C. Vatopoulos, and M. J. Sprenger.** 1999. European Antimicrobial Resistance Surveillance System (EARSS): objectives and organisation. Euro Surveill **4:**41-44.
4. **Brueggemann, A. B., D. T. Griffiths, E. Meats, T. Peto, D. W. Crook, and B. G. Spratt.** 2003. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. J Infect Dis **187:**1424-32.
5. **Chan, M. S., M. C. Maiden, and B. G. Spratt.** 2001. Database-driven Multi Locus Sequence Typing (MLST) of bacterial pathogens. Bioinformatics **17:**1077-83.
6. **Chiu, C. H., L. H. Su, and C. Chu.** 2004. Salmonella enterica serotype Choleraesuis: epidemiology, pathogenesis, clinical disease, and treatment. Clin Microbiol Rev **17:**311-22.
7. **Chung, M., H. de Lencastre, P. Matthews, A. Tomasz, I. Adamsson, M. Aries de Sousa, T. Camou, C. Cocuzza, A. Corso, I. Couto, A. Dominguez, M. Gniadkowski, R. Goering, A. Gomes, K. Kikuchi, A. Marchese, R. Mato, O. Melter, D. Oliveira, R. Palacio, R. Sa-Leao, I. Santos Sanches, J. H. Song, P. T. Tassios, and P. Villari.** 2000. Molecular typing of methicillin-resistant Staphylococcus aureus by pulsed-field gel electrophoresis: comparison of results obtained in a multilaboratory effort using identical protocols and MRSA strains. Microb Drug Resist **6:**189-98.
8. **Clarke, S. C., M. A. Diggle, and G. F. Edwards.** 2001. Semiautomation of multilocus sequence typing for the characterization of clinical isolates of Neisseria meningitidis. J Clin Microbiol **39:**3066-71.
9. **Coenye, T., T. Spilker, A. Martin, and J. J. LiPuma.** 2002. Comparative assessment of genotyping methods for epidemiologic study of Burkholderia cepacia genomovar III. J Clin Microbiol **40:**3300-7.
10. **Dagan, R., R. Melamed, M. Muallem, L. Piglansky, D. Greenberg, O. Abramson, P. M. Mendelman, N. Bohidar, and P. Yagupsky.** 1996. Reduction of nasopharyngeal carriage of pneumococci during the second year of life by a heptavalent conjugate pneumococcal vaccine. J Infect Dis **174:**1271-8.
11. **Dagan, R., R. Melamed, M. Muallem, L. Piglansky, and P. Yagupsky.** 1996. Nasopharyngeal colonization in southern Israel with antibiotic- resistant pneumococci during the first 2 years of life: relation to serotypes likely to be included in pneumococcal conjugate vaccines. J Infect Dis **174:**1352-5.

12. **de Lencastre, H., I. Couto, I. Santos, J. Melo-Cristino, A. Torres-Pereira, and A. Tomasz.** 1994. Methicillin-resistant Staphylococcus aureus disease in a Portuguese hospital: characterization of clonal types by a combination of DNA typing methods. Eur J Clin Microbiol Infect Dis **13:**64-73.
13. **Dice, L. R.** 1945. Measures of the amount of ecological association between species. Ecology **26:**297-302.
14. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol **186:**1518-30.
15. **Feil, E. J., J. M. Smith, M. C. Enright, and B. G. Spratt.** 2000. Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. Genetics **154:**1439-50.
16. **Fitzgerald, J. R., and J. M. Musser.** 2001. Evolutionary genomics of pathogenic bacteria. Trends Microbiol **9:**547-53.
17. **Fraser, C., W. P. Hanage, and B. G. Spratt.** 2005. Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci U S A **102:**1968-73.
18. **Frenay, H. M., A. E. Bunschoten, L. M. Schouls, W. J. van Leeuwen, C. M. Vandenbroucke-Grauls, J. Verhoef, and F. R. Mooi.** 1996. Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. Eur J Clin Microbiol Infect Dis **15:**60-4.
19. **Gertz, R. E., Jr., M. C. McEllistrem, D. J. Boxrud, Z. Li, V. Sakota, T. A. Thompson, R. R. Facklam, J. M. Besser, L. H. Harrison, C. G. Whitney, and B. Beall.** 2003. Clonal distribution of invasive pneumococcal isolates from children and selected adults in the United States prior to 7-valent conjugate vaccine introduction. J Clin Microbiol **41:**4194-216.
20. **Grothues, D., and B. Tummler.** 1991. New approaches in genome analysis by pulsed-field gel electrophoresis: application to the analysis of Pseudomonas species. Mol Microbiol **5:**2763-76.
21. **Hunter, P. R., and M. A. Gaston.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol **26:**2465-6.
22. **Lefevre, J. C., G. Faucon, A. M. Sicard, and A. M. Gasc.** 1993. DNA fingerprinting of Streptococcus pneumoniae strains by pulsed-field gel electrophoresis. J Clin Microbiol **31:**2724-8.
23. **Lipshutz, R. J., S. P. Fodor, T. R. Gingeras, and D. J. Lockhart.** 1999. High density synthetic oligonucleotide arrays. Nat Genet **21:**20-4.
24. **Lockhart, D. J., and E. A. Winzeler.** 2000. Genomics, gene expression and DNA arrays. Nature **405:**827-36.
25. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140-5.
26. **Malachowa, N., A. Sabat, M. Gniadkowski, J. Krzyszton-Russjan, J. Empel, J. Miedzobrodzki, K. Kosowska-Shick, P. C. Appelbaum, and W. Hryniewicz.** 2005. Comparison of multiple-locus variable-number tandem-repeat analysis with pulsed-field gel electrophoresis, spa typing, and multilocus sequence typing for clonal characterization of Staphylococcus aureus isolates. J Clin Microbiol **43:**3095-100.
27. **Maslow, J. N., M. E. Mulligan, and R. D. Arbeit.** 1993. Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. Clin Infect Dis **17:**153-62; quiz 163-4.
28. **Mato, R., I. S. Sanches, C. Simas, S. Nunes, J. A. Carrico, N. G. Sousa, N. Frazao, J. Saldanha, A. Brito-Avo, J. S. Almeida, and H. D. Lencastre.**

2005. Natural history of drug-resistant clones of Streptococcus pneumoniae colonizing healthy children in Portugal. Microb Drug Resist **11:**309-22.

29. **McDougal, L. K., C. D. Steward, G. E. Killgore, J. M. Chaitram, S. K. McAllister, and F. C. Tenover.** 2003. Pulsed-field gel electrophoresis typing of oxacillin-resistant Staphylococcus aureus isolates from the United States: establishing a national database. J Clin Microbiol **41:**5113-20.
30. **Miragaia, M., I. Couto, S. F. Pereira, K. G. Kristinsson, H. Westh, J. O. Jarlov, J. Carrico, J. Almeida, I. Santos-Sanches, and H. de Lencastre.** 2002. Molecular characterization of methicillin-resistant Staphylococcus epidermidis clones: evidence of geographic dissemination. J Clin Microbiol **40:**430-8.
31. **Murchan, S., M. E. Kaufmann, A. Deplano, R. de Ryck, M. Struelens, C. E. Zinn, V. Fussing, S. Salmenlinna, J. Vuopio-Varkila, N. El Solh, C. Cuny, W. Witte, P. T. Tassios, N. Legakis, W. van Leeuwen, A. van Belkum, A. Vindel, I. Laconcha, J. Garaizar, S. Haeggman, B. Olsson-Liljequist, U. Ransjo, G. Coombes, and B. Cookson.** 2003. Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant Staphylococcus aureus: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. J Clin Microbiol **41:**1574-85.
32. **NCBI** 2006, posting date. Complete Microbial Genomes: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. National Center for Biotechnology Information. [Online.]
33. **Nemoy, L. L., M. Kotetishvili, J. Tigno, A. Keefer-Norris, A. D. Harris, E. N. Perencevich, J. A. Johnson, D. Torpey, A. Sulakvelidze, J. G. Morris, Jr., and O. C. Stine.** 2005. Multilocus sequence typing versus pulsed-field gel electrophoresis for characterization of extended-spectrum beta-lactamase-producing Escherichia coli isolates. J Clin Microbiol **43:**1776-81.
34. **Riley, L. W.** 2004. Molecular epidemiology of infectious diseases: principles and practices, 1st ed, vol. 1. ASM Press.
35. **Sa-Leao, R., A. Tomasz, I. S. Sanches, S. Nunes, C. R. Alves, A. B. Avo, J. Saldanha, K. G. Kristinsson, and H. de Lencastre.** 2000. Genetic diversity and clonal patterns among antibiotic-susceptible and - resistant Streptococcus pneumoniae colonizing children: day care centers as autonomous epidemiological units. J Clin Microbiol **38:**4137-44.
36. **Schwartz, D. C., and C. R. Cantor.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:**67-75.
37. **Serrano, I., J. Melo-Cristino, J. A. Carrico, and M. Ramirez.** 2005. Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal. J Clin Microbiol **43:**1706-15.
38. **Silva, S., R. Gouveia-Oliveira, A. Maretzek, J. Carrico, T. Gudnason, K. G. Kristinsson, K. Ekdahl, A. Brito-Avo, A. Tomasz, I. S. Sanches, H. de Lencastre, and J. Almeida.** 2003. EURISWEB--Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers. BMC Med Inform Decis Mak **3:**9.
39. **Sneath, P. H., and R. R. Solkal.** 1973. Numerical Taxonomy, San Francisco.
40. **Spratt, B. G.** 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Curr Opin Microbiol **2:**312-6.
41. **Spratt, B. G., W. P. Hanage, B. Li, D. M. Aanensen, and E. J. Feil.** 2004. Displaying the relatedness among isolates of bacterial species -- the eBURST approach. FEMS Microbiol Lett **241:**129-34.
42. **Struelens, M. J.** 1996. Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. Clin Microbiol Infect **2:**2-11.

43. **Swaminathan, B., T. J. Barrett, S. B. Hunter, and R. V. Tauxe.** 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis **7:**382-9.
44. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol **33:**2233-9.
45. **van Belkum, A., M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc.** 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin Microbiol Rev **14:**547-60.
46. **van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards.** 2002. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med **347:**1999-2009.
47. **Zhang, S., R. A. Kingsley, R. L. Santos, H. Andrews-Polymenis, M. Raffatellu, J. Figueiredo, J. Nunes, R. M. Tsolis, L. G. Adams, and A. J. Baumler.** 2003. Molecular pathogenesis of Salmonella enterica serotype typhimurium-induced diarrhea. Infect Immun **71:**1-12.

Chapter II

2. EURISWEB: Web-based epidemiological surveillance of antibiotic-resistant pneumococci in Day Care Centers

Published in Silva, S., et al., *EURISWEB--Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers.* BMC Med Inform Decis Mak, 2003. 3(1): p. 9.

BMC Medical Informatics and Decision Making

BioMed Central

Research article

Open Access

EURISWEB – Web-based epidemiological surveillance of antibiotic-resistant pneumococci in Day Care Centers

Sara Silva[1], Rodrigo Gouveia-Oliveira[1], António Maretzek[1], João Carriço[1], Thorolfur Gudnason[2], Karl G Kristinsson[2], Karl Ekdahl[3], António Brito-Avô[4], Alexander Tomasz[5], Ilda Santos Sanches[1,6], Hermínia de Lencastre[1,5] and Jonas Almeida*[1,7]

Address: [1]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. República (EAN), PO Box 127, 2781-901 Oeiras, Portugal, [2]Department of Pediatrics and Microbiology, Landspitali University Hospital, Reykjavik, Iceland, [3]Swedish Institute for Infectious Diseases Control, Department of Epidemiology, Se-171 82 Solna, Sweden, [4]Centro de Saúde de Oeiras, Av. Salvador Allende, 2780-163 Oeiras, Portugal, [5]Laboratory of Microbiology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA, [6]Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Monte de Caparica, 2829-516 Caparica, Portugal and [7]Dept Biometry & Epidemiology, Medical Univ South Carolina, 135 Cannon Street, Suite 303, PO Box 250835, Charleston SC 29425, USA

Email: Sara Silva - sara@itqb.unl.pt; Rodrigo Gouveia-Oliveira - rodrigo@itqb.unl.pt; António Maretzek - antonio.maretzek@microcortex.com; João Carriço - jcarrico@itqb.unl.pt; Thorolfur Gudnason - thorgud@landspitali.is; Karl G Kristinsson - karl@landspitali.is; Karl Ekdahl - karl.ekdahl@mtc.ki.se; António Brito-Avô - abritoavo@netcabo.pt; Alexander Tomasz - tomasz@mail.rockefeller.edu; Ilda Santos Sanches - isanches@itqb.unl.pt; Hermínia de Lencastre - lencash@mail.rockefeller.edu; Jonas Almeida* - almeidaj@musc.edu

* Corresponding author

Published: 08 July 2003

BMC Medical Informatics and Decision Making 2003, **3**:9

Received: 28 February 2003
Accepted: 08 July 2003

This article is available from: http://www.biomedcentral.com/1472-6947/3/9

Abstract

Background: EURIS (European Resistance Intervention Study) was launched as a multinational study in September of 2000 to identify the multitude of complex risk factors that contribute to the high carriage rate of drug resistant *Streptococcus pneumoniae* strains in children attending Day Care Centers in several European countries. Access to the very large number of data required the development of a web-based infrastructure – EURISWEB – that includes a relational online database, coupled with a query system for data retrieval, and allows integrative storage of demographic, clinical and molecular biology data generated in EURIS.

Methods: All components of the system were developed using open source programming tools: data storage management was supported by PostgreSQL, and the hypertext preprocessor to generate the web pages was implemented using PHP. The query system is based on a software agent running in the background specifically developed for EURIS.

Results: The website currently contains data related to 13,500 nasopharyngeal samples and over one million measures taken from 5,250 individual children, as well as over one thousand pre-made and user-made queries aggregated into several reports, approximately. It is presently in use by participating researchers from three countries (Iceland, Portugal and Sweden).

Conclusion: An operational model centered on a PHP engine builds the interface between the user and the database automatically, allowing an easy maintenance of the system. The query system is also sufficiently adaptable to allow the integration of several advanced data analysis procedures

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

far more demanding than simple queries, eventually including artificial intelligence predictive models.

Background

Social forces that produced Day Care Centers (DCCs) for preschool age children in many developed countries have – ironically – also created in these structures one of the major ecological reservoirs of drug resistant strains of *Streptococcus pneumoniae*, which spread globally and began to create serious complications in the chemotherapy of diseases caused by this dangerous pathogen [1–3]. Day Care Centers recruit in close physical proximity children of an age group that is characterized by high rate of carriage of *S. pneumoniae*, an immature immune system and frequent viral and bacterial respiratory tract infections leading to extensive use of antimicrobial agents which provide a powerful selective milieu for the emergence of resistant strains [4–7]. The best evidence that such strains can cause both pediatric and adult disease came from molecular epidemiological studies, which demonstrated that resistant clones of *S. pneumoniae* most frequently identified in disease [8,9] were also the ones frequently carried in the nasopharynx of healthy children in DCCs [10–12].

If DCCs are ecological reservoirs of resistant *S. pneumoniae* then reduction in the rate of carriage of such strains in DCCs should also impact on the frequency of infections caused by resistant pneumococci. Testing the efficacy of such a novel strategy was the purpose of the multinational initiative EURIS (European Resistance Intervention Study – Reducing Resistance in Respiratory Tract Pathogens in Children) [13] launched by the European Community in September of 2000 until 2003. Investigators from four countries (France, Iceland, Portugal and Sweden) supported by scientists from Germany and the USA joined forces to test the effect of a variety of different interventions methods (e.g. reduction in drug prescriptions; changing antibiotic dosing; improving hygienic conditions in DCCs etc.) on the frequency of nasopharyngeal colonization by resistant pneumococci – in carefully controlled studies.

The structure of EURIS is composed of four centers where strain collections and interventions are carried out: Portugal – Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa; Iceland-Landspitali University Hospital; Sweden – Swedish Institute for Infectious Diseases Control; France – Institut National de la Santé et de la Recherche Médicale. However EURISWEB represents data generated in only three of the four collection centers: Portugal, Iceland and Sweden, the three countries in which the timing, the age groups of the children and the methods used were fully harmonized. The French initiative, while addressing the same issues, was not directly comparable, as it involved different age groups, different mode of sampling, and schools rather than Day Care Centers. Therefore the French data were deposited in a different database. Four additional collaborating units assist as reference centers for the harmonization of methods in clinical microbiology (Iceland: Landspitali University Hospital); molecular epidemiology of antibiotic resistant genes and clones (USA: Laboratory of Microbiology, The Rockefeller University; Germany: University of Kaiserslautern); data management and mathematical modeling of epidemiological aspects of EURIS (Portugal: Instituto de Biologia Experimental e Tecnológica).

The risk factors, *i.e.* the nature and number of the factors that influence the rate of carriage of drug resistant *S. pneumoniae* in preschool age children and their quantitative contribution to the degree of colonization, are not well understood. Furthermore, major risk factors for nasopharyngeal colonization may differ significantly from one setting to another [14–16], which makes analysis of data generated by a multinational study like EURIS, more complex. The evaluation and comparison of such massive amounts of surveillance data necessitated the construction of a computerized infrastructure organized in such a manner that it would assure not only data storage and retrieval but also an eventual bioinformatics analysis. The purpose of this communication is to describe such a web-based infrastructure specially designed to fit the purposes of EURIS – the EURISWEB.

Several potential conflicting attributes had to be accommodated in the design of such an infrastructure. On the one hand, it was to provide full integration of data from different countries in a common normalized repository, fully accessible to all EURIS participants. On the other hand, it was also supposed to exhibit the properties of a local database with full separation between the countries involved. Finally, it was anticipated that, eventually, EURISWEB would be made available for wider usage for research and public health management at a later stage, with steep requirements of stability, scalability, security, user-friendly access, low cost portability, and transparent implementation for subsequent independent development. In the design of EURISWEB we took into account the multiple goals of such a web-based infrastructure which now includes a relational online database coupled with data retrieval and analysis tools, where registered users can access data and tools by using a personal login

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

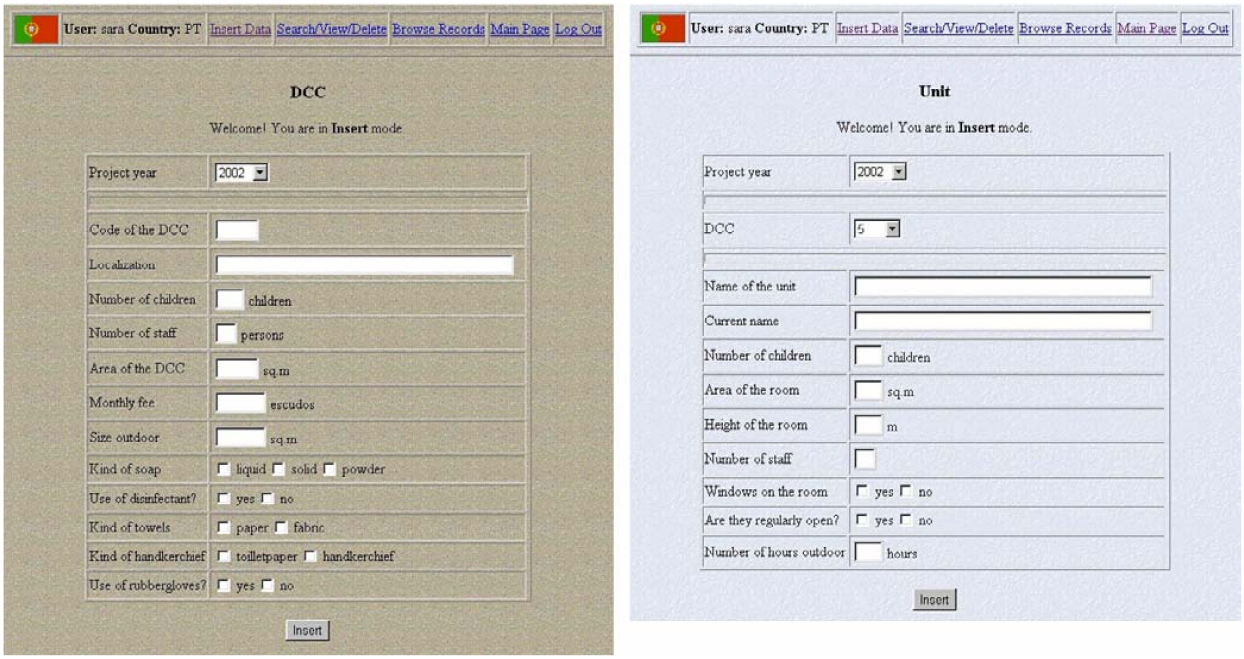


Figure 1
Data entry forms Data entry forms for the DCC and room (unit within the DCC) questionnaires. Upon pressing the Insert button, the validation procedure checks the data and either inserts it in the database or informs the user about errors in the data.

name and password through a standard web browser. Ultimately three of the four centers (Iceland, Portugal and Sweden), in which the nature of the pediatric population and mode of sampling were most comparable, chose to combine all data for deposition in EURISWEB, which now covers a large number of participant institutions: 16 DCCs in Portugal, 30 DCCs in Iceland and 25 DCCs in Sweden; a wide variety of sources of data, including demographic, socio-economic factors, clinical data, patterns and types of drug use and drug prescription; microbiological data on the antibiotypes and serotypes as well as molecular types of the pneumococcal isolates and DNA fingerprints of resistant genes.

Availability

A demo version of EURISWEB is available to the general public [17], accessible with username euris and password welcome. For those who intend to receive the e-mails sent by the query agent (see User-Friendly Query System), please request a personal account to the authors. As any modifications applied to the current implementation of the infrastructure will automatically be reflected on the demo version, new features may be already apparent when compared with illustrations and examples used in this manuscript.

Methods

Data and data acquisition

The diverse surveillance data (demographic, clinical, microbiological etc.) generated in project EURIS are used to fill five different types of Questionnaires which serve as the source of information to be introduced into the EURISWEB database. The relationship between the five questionnaires is described and illustrated later in this report (see Database structure and Database tables versus online forms). Typically each site will update surveillance information at least once per year. Questionnaires 1 and 2 are provided by the staff of each participating DCC.

• Questionnaire 1 contains information regarding physical features of the DCC (address, number of rooms and windows, area inside and outside the facility, number of children and staff, hygiene protocols and practice) – see Figure 1.

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

• Questionnaires 2 provide the same type of information for each room (also referred to as "unit") in the particular DCC – see Figure 1.

• Questionnaires 3 are filled at least once every year by the parents of the children. They contain demographic information on the household and environment where the child lives, including number and age of siblings, shared bedrooms, and specific conditions such as smoking in the house.

• Questionnaires 4 are filled by the parents just prior to each strain collection. They provide information on antibiotic consumption prior to sampling (type of antibiotic, taken when and for how long). Also provided are data on illness and hospitalizations of the child.

• Questionnaires 5 are filled by the participating microbiology and molecular biology laboratories. They contain characterization of each *S. pneumoniae* isolate for serotype; antibiotype (susceptibility to oxacillin, chloramphenicol, erythromycin, clindamycin, tetracycline, sulfamethoxazole-trimethoprim, and levofloxacin); MIC values for penicillin and ceftriaxone; molecular type by PFGE (Pulsed-Field Gel Electrophoresis) and MLST (Multilocus Sequence Typing) (for selected isolates); DNA probes for antibiotic resistant factors; and RFLP (Restriction Fragments Length Polymorphisms) for *pbp* (penicillin binding protein) genes of selected penicillin resistant isolates. All data in questionnaire 5 are obtained by common harmonized methods.

Database conception

Although there was an effort towards the normalization of data acquisition taking place in the different participant countries, the questionnaires delivered to the DCCs and to the parents contain various questions that reflect realities specific to the country involved. Accordingly, some questions only appear in the questionnaires of some countries. Also, the frequency with which updated information is collected differs between countries. Since discarding data was to be avoided at all costs in order not to confront local practices, the normalization process had to be extended to database conception itself. Instead of designing an optimal database structure for each country, an iterative consulting process was followed for nearly a year to produce a normalized database structure that fits the reality presented by all the countries involved. The final structure of the EURISWEB database accommodates both country specificity and common European health management practices.

Data retrieval

Besides providing comprehensive data storage, the web-based data management infrastructure must also allow the easy querying and retrieval of the data it contains. Since some of the retrieval requests may generate large amounts of data, or may require intensive computation, the requests are processed as background processes managed by a software agent that send e-mails to the user with information on the execution state of each request and, finally, a link to the completed report. User-friendliness was the primary concern in building the interface available to make these requests, with current version reflecting extensive user feedback.

Software and Hardware

All the software used to implement this infrastructure is Open Source and is provided under public license. The scripts that generate the HTML (HyperText Markup Language) interfaces were written in PHP (PHP: Hypertext Preprocessor) [18] 4.x. The Database Management System is PostgreSQL [19] 7.1.x. The scripts for the agents that handle the data retrieval requests were written in shell (bash 2.04) and Perl [20] (5.x). The server is a PC, CPU 2x PIII (coppermine) @ 800 MHz, with 512 MB SDRAM, running Linux [21] (based on Slackware [22] 7.1, kernel 2.2.x), SSL (Secure Socket Layer) [23] enabled Apache Web-Server [24] (Apache/1.3.x Ben-SSL/1.x). Software versions described above were updated regularly throughout the course of the EURIS project (2000–2003) with no negative impact on its performance.

Results

Database structure

The basic internal structure of the database consists of 9 tables with an average of 14 fields per table. Figure 2 shows a simple model of this structure, where boxes represent tables and lines represent the relations between them. There is only one type of relation in this structure, which is "one-to-many", the "one" side being represented by the single line and the "many" side being represented by the forked line. A one-to-many relation between two tables means that one record from one table can be associated to several records from the other table. For example, one DCC can be associated to several rooms (units) in the same DCC; one unit can be associated with several children; and one child can be associated with several siblings.

The description illustrated in Figure 2 is country specific. A separate set of tables was defined for each of the three participant countries – Portugal, Iceland and Sweden, all inside the same database, but not formally connected to each other. Although the questionnaires for the different countries have significant differences, as some countries may lack many fields or even whole tables of this structure, the critical feature is that all the common fields can be found in exactly the same location in each country-specific structure. Equally critical, the key fields are

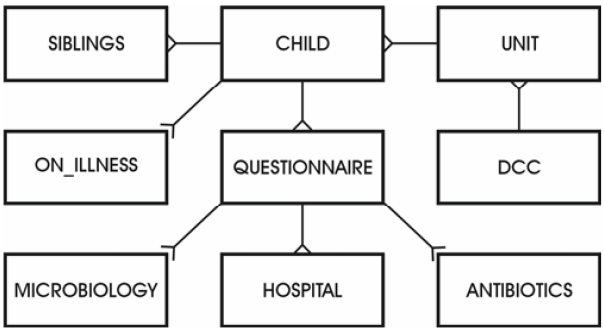


Figure 2
Database structure Boxes represent tables and lines represent the relations between them. There is a similar set of tables for each country. There is only one type of relation in this structure, which is "one-to-many". The single line represents the "one" side and the split line represents the "many" side. A one-to-many relation between two tables signifies that one record from one table can be associated to several records from the other table.

obligatorily shared by all countries, a feature that can only be easily achieved if a common host infrastructure is in place, which is the case in EURISWEB. Because the frequency with which updated information is collected differs between countries (see Data and data acquisition, and Database conception), many of the key fields are related to the specification of the sampling periods, playing an important role as temporal normalization features. Once the access restrictions are lowered, the conservation of ontology and structure enables intersection between country-specific structures to produce comprehensive data sets jointly describing epidemiological data, which are valid for all the participating countries. Furthermore, because all countries also share the same data retrieval system (see User-Friendly Query System), queries already built by different countries produce compatible results that can be promptly joined after removing the country-specific fields.

Online interface

The interface between the database and the users is made of standard HTML pages (no external applications, "plug-ins", needed on the client side). Data entering is performed through five online forms that mimic the original paper questionnaires, to facilitate the insertion task (see examples of two forms in Figure 1). All data entered in the forms is submitted to online validation procedures before entering the database, thus avoiding some of the most common user errors that may cause integrity or consistency violations in the database. Upon pressing the Insert button for submitting data, the user is promptly informed of all its mistakes and given a chance of resolving them on the same page (example in Figure 3). Only after passing all the checks is the data effectively inserted in the database, and fitted into the respective internal data structure (see Database structure). Searching and visualizing data can be done on a record-by-record basis, using the same five forms format, or by browsing as a table that shows several records at the same time (example in Figure 4). Some simple statistics can also be requested online. For convenience, most of the tables presented can be directly viewed or saved in Excel format.

Data retrieval requests can also be made by filling a simple online form in which the amount of typing required is kept to a minimum (see User-Friendly Query System). The results can be viewed and downloaded in delimited text format, also readily importable into Excel.

Database tables versus online forms

The relationship between the internal database structure and the set of online forms available to the user is not a one-to-one association. Behind each form there can be more than one table, as shown in Figure 5. Although the mimicking of the original questionnaires by the online forms is meant to facilitate the user's adaptation to the data insertion and visualization, that is not the optimal data organization in a relational database. For example, the repeated set of questions about each antibiotic taken prior to sampling (see Data and data acquisition) should not result in a repeated set of fields in the same database table (table QUESTIONNAIRE, see Figure 2). Instead, each set of questions constitutes a row of fields in a different table (table ANTIBIOTICS, same figure).

Operational model

The operational model of the database interface is depicted in Figure 6, where the arrows represent flow of information between the various entities. The five online forms for record insertion and visualization, available to the user, are all built with the same general procedure (PHP engine). This program, written in PHP, reads files that contain all the information regarding the forms layout (layout files), designs the forms and manages all the interactions between the users and the database. Each layout file describes a form (for all countries) and consists of a few lines written in a subset of the PHP language, which indicate each field's properties, such as whether it is a numeric or Boolean field, a date or time field, and what are the range and type of values allowed. This program and the subset of PHP used to define the layout files are the core of the surveillance system reported here. Accordingly, to alter an existing form, or generate a new one, all the database manager has to do is update or build a layout file.

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

Figure 3
Validation checks Example of data entry form for the child records with error warnings to the user. The user must correct all the errors before being able to insert the record.

The layout files also include the description of the connection between the form fields and the actual database fields. This information must be in accordance to the internal database structure, which is managed by SQL (Structured Query Language) code also stored in files (structure files). Therefore, the database manager will need to keep them consistent with any changes in the structure files required by modifications in the online interface. These two simple tasks ensure both the automatic construction of personalized forms – together with online validation check procedures – and a smooth linkage between them and the database internal structure.

User-Friendly Query System

Although SQL is the standard way to access data stored in a database, using it requires some prior knowledge and experience from the user. The User-Friendly Query System, available to all the EURISWEB users, is an interface that facilitates query construction in order to make the wide range of possibilities offered by SQL amenable to the untrained user. The users are presented with a series of selection boxes where they can select the fields they want to see, the restrictions they want to apply to the records returned, and how the returned records are to be grouped (Figure 7). The chosen options are then transformed into actual SQL formatted statements that are sent to the query management agent, through the PHP engine, as shown in Figure 8. The arrows in the figure represent flow of information between the various entities (see Figure 9 for the whole operational model).

The query agent manages all the requests and runs them exclusively in background, so that high usage rates and complex requests do not interfere with the normal usage of the database interface. The agent interacts with the database and informs the users, by e-mail, of when their requests start being processed and when they finish, including the information of whether the query was

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

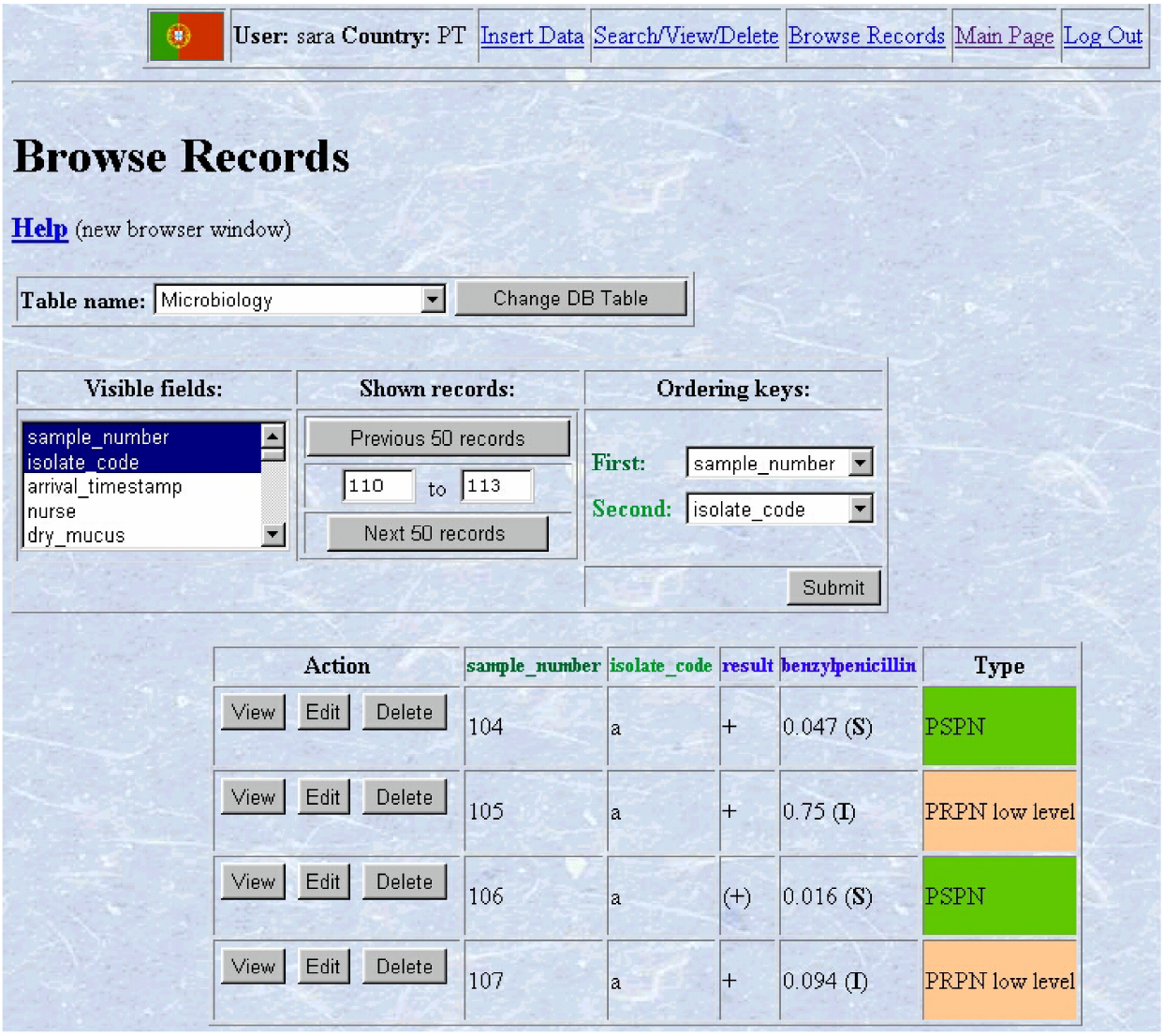


Figure 4
Browsing records Browsing page for microbiology records. Here the user can browse through several records of the same table. Selection of visible fields and ordering by one or two fields is possible. Direct access to individual records is provided by the View, Edit, and Delete buttons.

successfully answered (the interface gives users enough freedom to request impossible things) or not, in which case the results presented are an empty text page. Due to security reasons, the results of queries are never sent by e-mail – they can only be downloaded from the server via an SSL connection.

Users can rerun, edit, or delete saved queries. They can also group queries into reports, so that a single request will yield all the results from the several queries of that report. Furthermore, users can save restrictions used often, and apply them to other queries. To minimize the time and effort required of the users, we have provided several pre-made queries, already aggregated into several logical reports. This feature may prove particularly useful if standard reporting formats become a regulatory requirement.

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

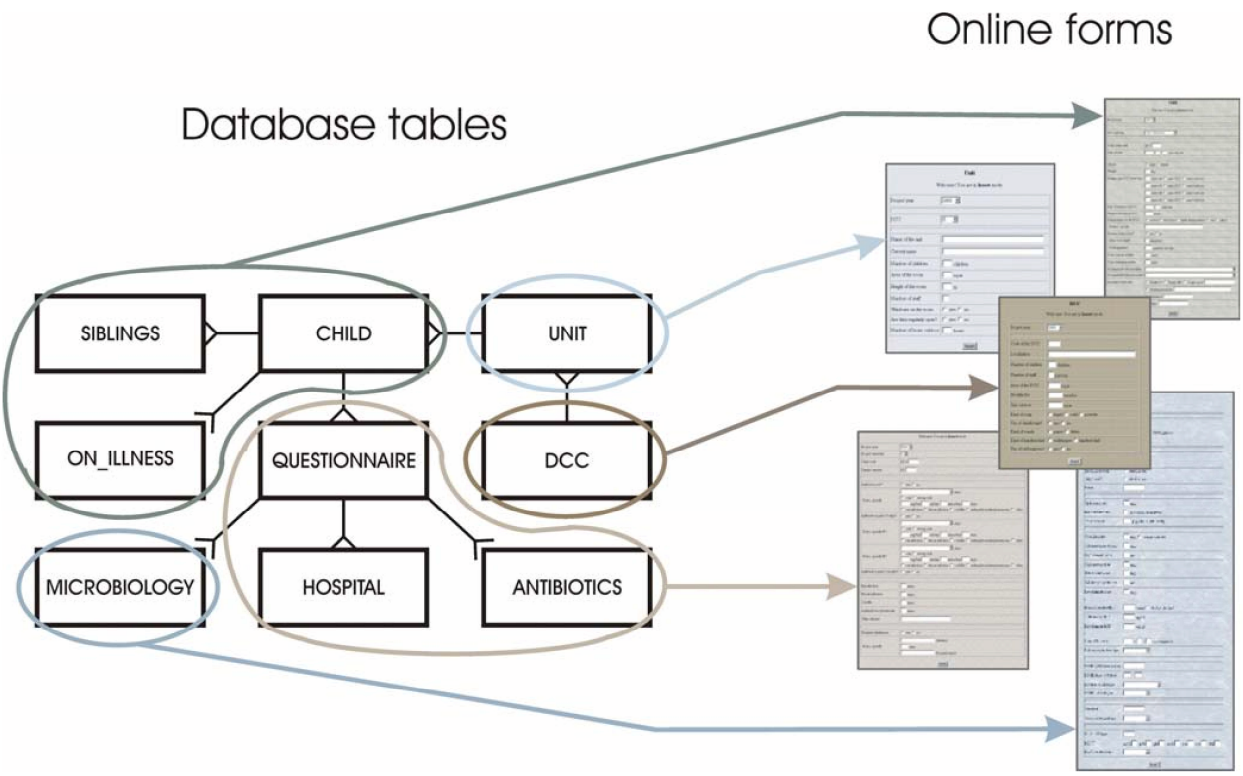


Figure 5
Tables *versus* forms Relationship between database tables and online forms. The optimal data organization in a relational database may not be agreeable to the user. There may be several database tables behind each online form.

Usage

The EURIS online database has been adopted as the data storage standard by three of the EURIS participant countries – Portugal, Iceland, and Sweden. Growing steadily since its birth, February 2001, it now has 24 registered users and contains a total of 213 DCC records, 720 unit records, 10991 children records, 13207 questionnaire records, and 13504 microbiology records, totaling more than 25 megabytes of data. The User-Friendly Query System, available since April 2002, now contains 400 premade and 786 user-made queries, aggregated into several reports.

Discussion and Conclusions

Privacy and security precautions

In EURISWEB, each user registry includes not only its login name and password, but also its country identification, which completely blocks access to data belonging to other countries. In fact, by using different table sets for each country, the central database can behave like several different local databases, and the user is never aware that its access is restricted to only a subset of the complete system. In a near future, the fact that all countries use, after all, the same normalized structure, will allow simple queries and complex data analysis to be performed in the common data, as if we were dealing with a single country.

The user registration process also includes a level access number that defines restrictions for each type of user. Although the system was built anticipating this need, other precautions proved sufficient to monitor and to recover from possible destructive actions. All user inputs are scanned for invalid characters to prevent SQL code injection, and a record of all actions performed in the database is kept, including who did what, and when. Any accidentally deleted record can be promptly restored by the database manager; all the updates a record has undergone since its insertion can be tracked; and many database usage statistics can be easily performed.

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

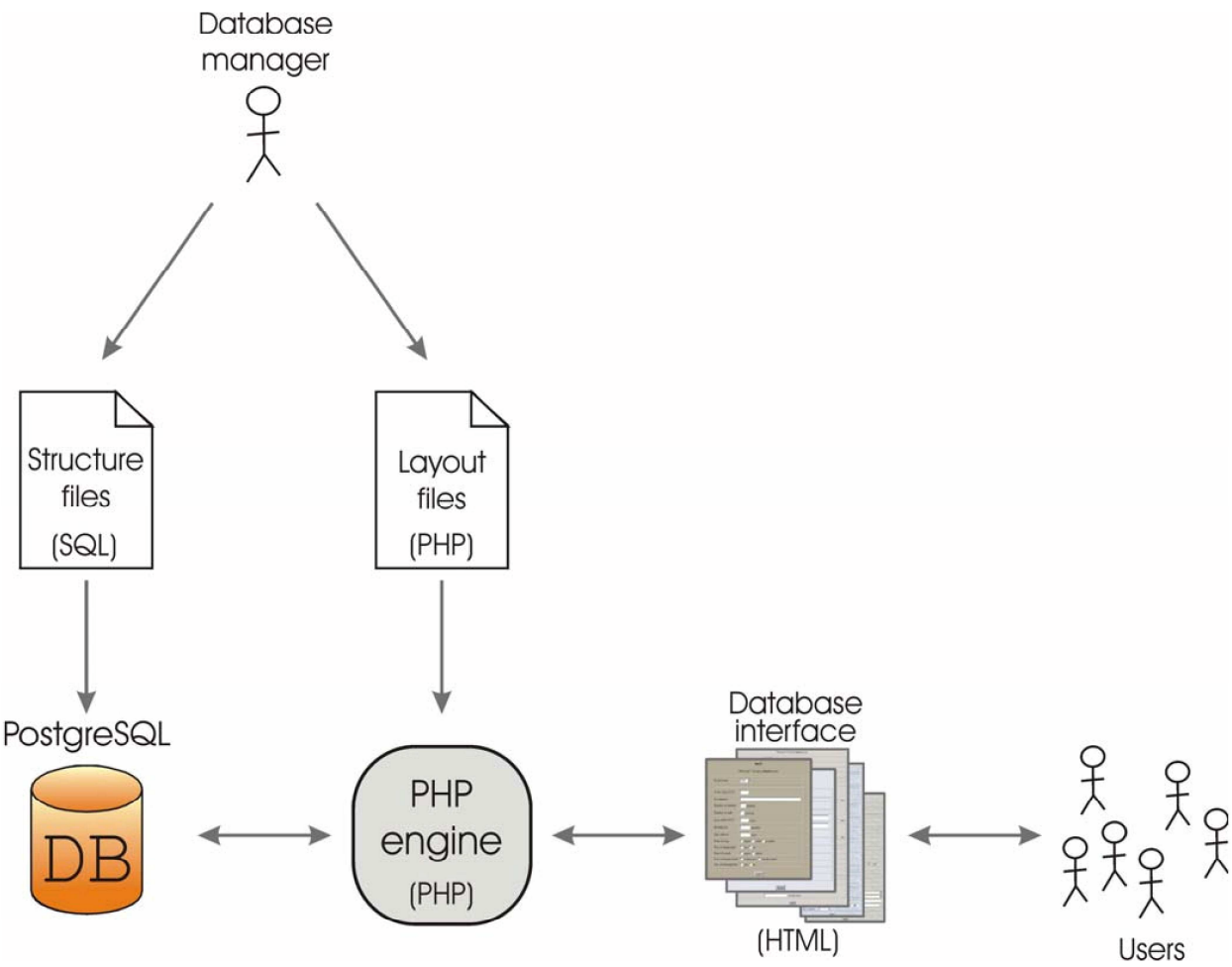


Figure 6
Operational model – database Operational model of the EURISWEB database. The arrows represent flow of information between the various entities. The database manager builds the layout files, used by the PHP engine to build the online forms, and the structure files, used to define the structure of tables and fields in the database. Layout and structure must be in accordance with each other. The PHP engine manages all the interactions between the database and the users.

Intrusion by unauthorized parties (hackers) is repelled by the need to log on with login name and password, and subsequent identification of the user with cookies protected by SSL, without which no page is ever shown and no query is run. A brute force attack is also limited by a delay introduced in the password checking cycle, and resources consumption at the server. Additionally, page accesses are monitored on a daily basis. Repeated login attempts would therefore be promptly detected before a sufficiently high number of probes take place. Furthermore, a firewall protects the server from being accessed on other ports apart from the HTTPS port, and the server software (Apache, PHP, kernel, etc.) is promptly updated if any security breach is detected in the current versions.

In all cases, names of children and DCCs are not kept in the database, instead being replaced by codes and acronyms manually assigned prior to insertion.

Scalability

The operational model described in the Results section is the basis for easy improvements and extensions to the whole infrastructure of EURISWEB. As a consequence of the design described in that section, database management can be fully dealt with by manipulation of the

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

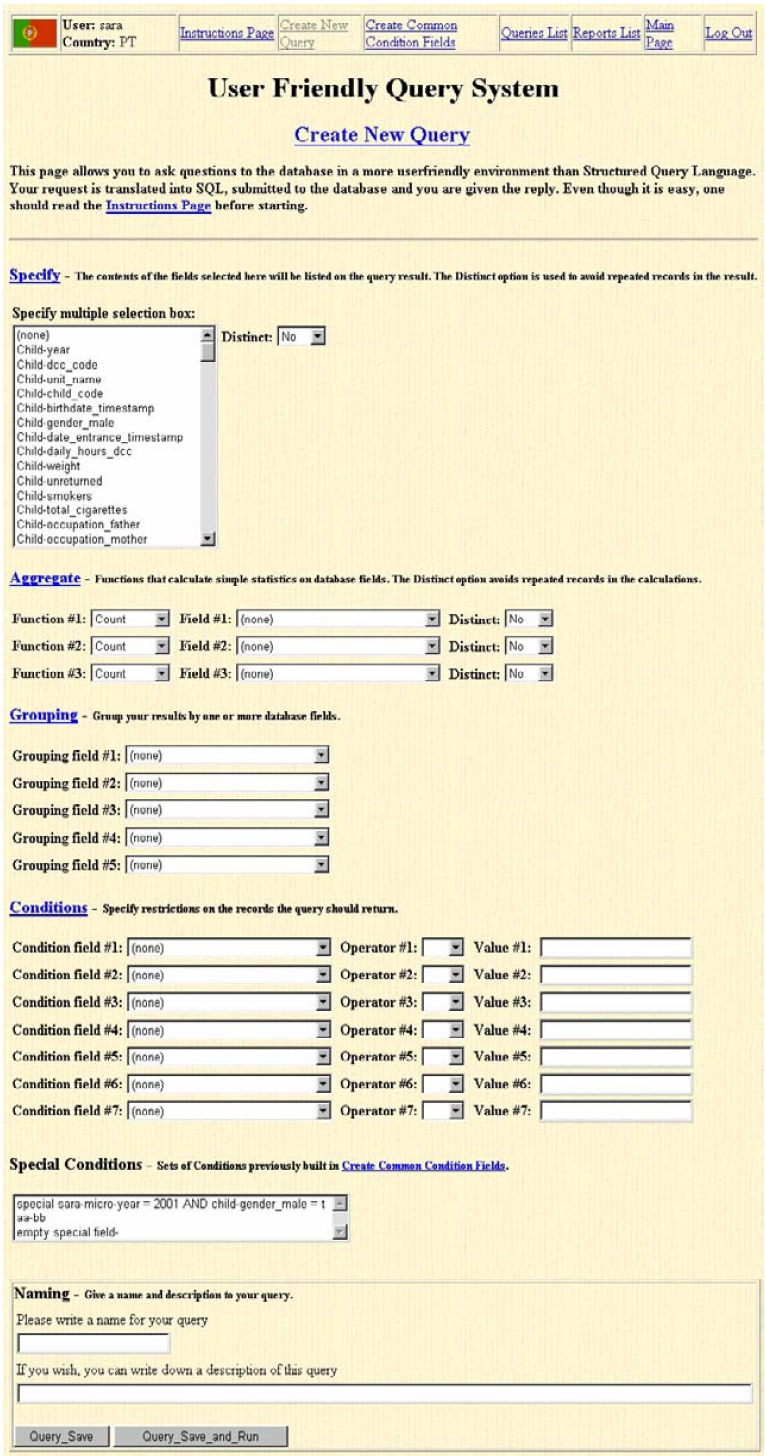


Figure 7
User-Friendly Query System interface Interface where the users build their queries, which are then translated into SQL. Saved queries can be rerun, edited, deleted, or grouped into reports. Commonly used sets of restrictions can also be saved and later applied to new queries.

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

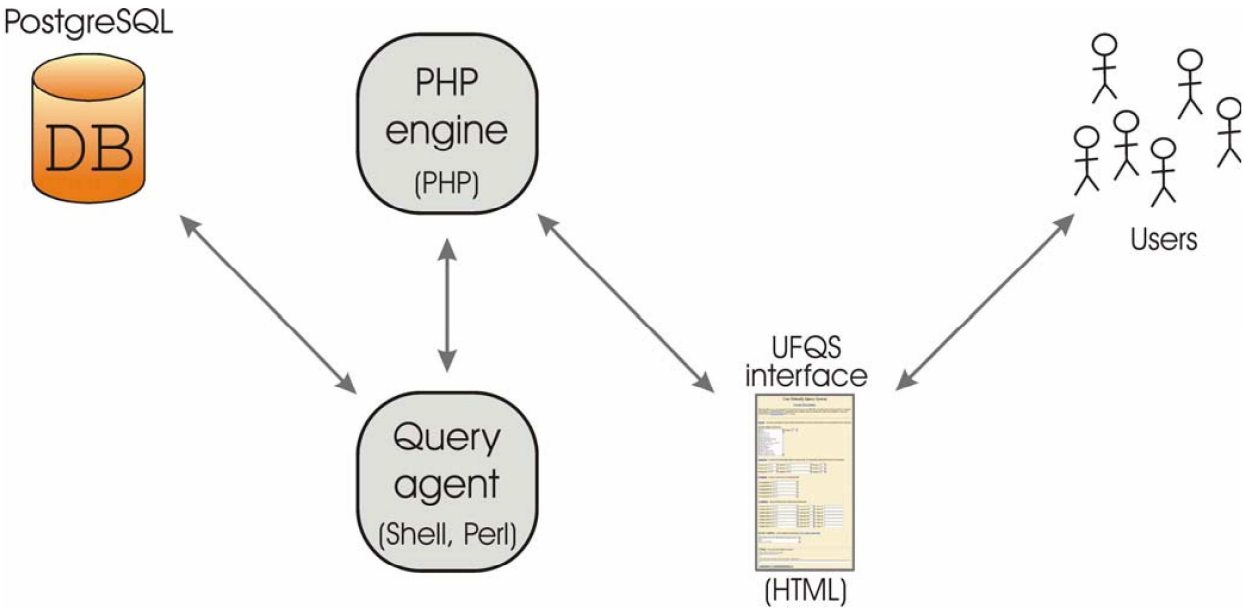


Figure 8
Operational model – query system Operational model of the EURISWEB query system. The arrows represent flow of information between the various entities. The user makes a request through the interface, which is translated into SQL by the PHP engine and sent to the query management agent. The query agent interacts with the database in background and informs the users, by e-mail, about the state of their requests.

layout and structure files. The core element of this operational functionality is supported by the PHP engine described in the Results section. As a result both maintenance and development scale well with increasing usage, particularly since availability of high performance hardware and Internet access have ceased to be an issue. It is noteworthy that the layout and structure files are particularly suited for extensions to the current model, including having new types of data integrated into the set already stored; having new countries and new country-specificities accommodated, while retaining previous accessibility and privacy. The demo version of the database shell and query system, made publicly available (see Availability for URL and login directions), was built by configuring a fictitious new country structure, which was achieved by performing minor modifications in the layout files.

This comes to illustrate that a possible useful extension to this system would be to allow selected users to manage their own tables and forms by providing a web-based interface to the layout and structure files. These users with management access permissions would not need to know the PHP or SQL languages, but simply interact with online forms with the same level of complexity as the regular database query forms. In our experience, most of the tasks requested to the database manager are simple and pose no risk whatsoever to the data, like adding an item to a drop-down selection field, resizing a text edit field, or even adding a new field to an already existing table. Given the wide geographic distribution of the EURISWEB users, describing what needs to be done to the database manager is as time consuming as specifying it in such idealized management forms. The database manager could then be left with only the more complex and "dangerous" tasks. Figure 9 shows how the operational model implemented, including the query system, could be configured to greatly remove the need for low level data management.

Future directions

The extension of data management to include data analysis is particularly suited for web-based implementations – such as EURISWEB – since all computation takes place on the server side. This approach enables a bioinformatics approach to establish itself alongside data storage. As a consequence, advanced data mining tools such as multivariate statistical analysis or the identification of artificial intelligence predictive models using neural networks [25] and rule extraction by genetic algorithms can be made available alongside the data itself. This is mutually beneficial for usage and development and, on the other

BMC Medical Informatics and Decision Making 2003, **3** http://www.biomedcentral.com/1472-6947/3/9

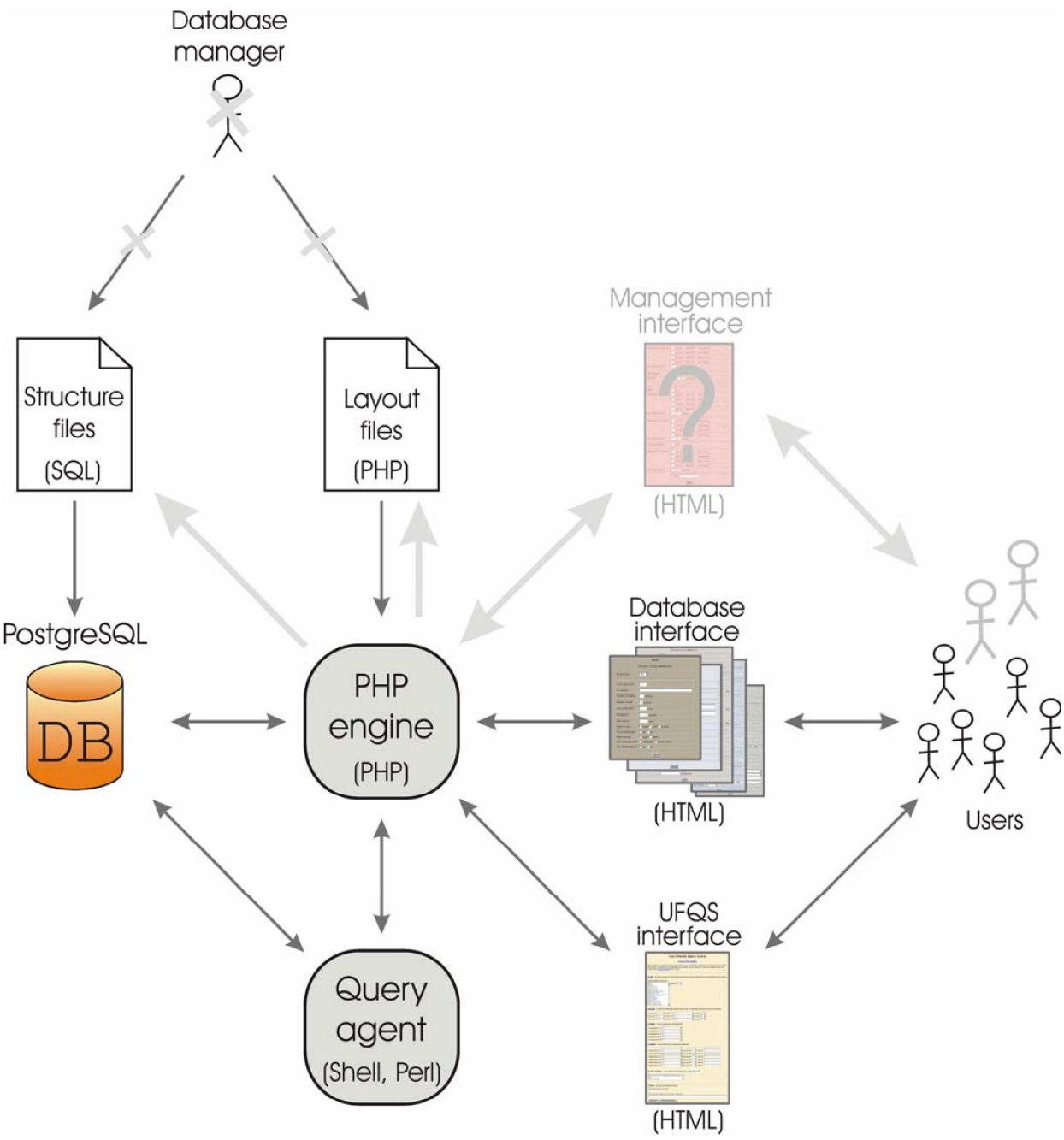


Figure 9
Operational model – the whole picture and hypothetical scenario Operational model of the whole EURISWEB infrastructure, by joining Figures 6 and 8, plus the hypothetical scenario where the database manager could be replaced by an extension to the PHP engine that would allow privileged users to manage both the layout and structure files through web pages.

hand, bioinformatics tool development has, in return, ready access to extensive datasets for validation as well as, even more important, facilitated interaction with domain experts that provide the context for its interpretation. This bi-directional integration enabled by a web-based development environment undoubtedly offers the best conditions for practical implementation of a full-fledged epidemiological information system.

Competing interests

None declared.

Authors' contributions

SS designed and manages the database, and prepared the manuscript draft. AM designed the query system and does the software and hardware maintenance. SS, RGO, AM and JC implemented the database and query system. TG, KGK, KE, AT, ISS, HL and JSA conceived of the EURIS Project, and participated in its design and coordination. ABA participated in the design of the study in Portugal. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grant QLK2-CT-2000-01020 (EURIS) from the European Commission. JC was supported by grant SFRH/BD/3123/2000, and AM by grant POCTI/1999/BSE/34794 (SAPIENS), both by Fundação para a Ciência e a Tecnologia, Ministério da Ciência e do Ensino Superior, Portugal. We also thankfully acknowledge the EURIS Portuguese team (C. Simas, R. Mato, S. Nunes, N. Sousa, N. Frazão) and the Icelandic Team for early advice that was essential for the EURISWEB prototype design, as well as J. Saldanha (Portugal) for help in designing the questionnaires.

References

1. Gray BM, Turner ME and Dillon HC Jr: **Epidemiologic studies of Streptococcus pneumoniae in infants. The effects of season and age on pneumococcal acquisition and carriage in the first 24 months of life** *Am J Epidemiol* 1982, **116:**692-703.
2. Yagupsky P, Porat N, Fraser D, Prajgrod F, Merires M, McGee L, Klugman KP and Dagan R: **Acquisition, carriage, and transmission of pneumococci with decreased antibiotic susceptibility in young children attending a day care facility in southern Israel** *J Infect Dis* 1998, **177:**1003-1012.
3. Butler JC, Dowell SF and Breiman RF: **Epidemiology of emerging pneumococcal drug resistance: implications for treatment and prevention** *Vaccine* 1998, **16:**1693-1697.
4. Reichler MR, Allphin AA, Breiman RF, Schreiber JR, Arnold JE, McDougal LK, Facklam RR, Boxerbaum B, May D, Walton RO and Jacobs MR: **The spread of multiply resistant Streptococcus pneumoniae at a day care center in Ohio** *J Infect Dis* 1992, **166:**1346-1353.
5. Arason VA, Kristinsson KG, Sigurdsson JA, Stefansdottir G, Molstad S and Gudmundsson S: **Do antimicrobials increase the carriage rate of penicillin resistant pneumococci in children? Cross sectional prevalence study** *BMJ* 1996, **313:**387-391.
6. Forssell G, Hakansson A and Mansson NO: **Risk factors for respiratory tract infections in children aged 2–5 years** *Scand J Prim Health Care* 2001, **19:**122-125.
7. Holmes SJ, Morrow AL and Pickering LK: **Child-care practices: effects of social change on the epidemiology of infectious diseases and antibiotic resistance** *Epidemiol Rev* 1996, **18:**10-28.
8. McGee L, Klugman KP and Tomasz A: **Serotypes and clones of antibiotic-resistant pneumococci** In *Streptococcus pneumoniae. Molecular Biology & Mechanisms of Disease* Edited by: *Tomasz A. New York: Mary Ann Liebert*; 1996:375-379.
9. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, Hakenbeck R, Hryniewicz W, Lefevre JC, Tomasz A and Klugman KP: **Nomenclature of major antimicrobial-resistant clones of Streptococcus pneumoniae defined by the pneumococcal molecular epidemiology network** *J Clin Microbiol* 2001, **39:**2565-2571.
10. Sá-Leão R, Tomasz A, Sanches IS, Brito-Avô A, Vilhelmsson SE, Kristinsson KG and de Lencastre H: **Carriage of internationally spread clones of Streptococcus pneumoniae with unusual drug resistance patterns in children attending day care centers in Lisbon, Portugal** *J Infect Dis* 2000, **182:**1153-1160.
11. Sá-Leão R, Tomasz A, Sanches IS, Nunes S, Alves CR, Brito-Avô A, Saldanha J, Kristinsson KG and de Lencastre H: **Genetic diversity and clonal patterns among antibiotic-susceptible and -resistant Streptococcus pneumoniae colonizing children: day care centers as autonomous epidemiological units** *J Clin Microbiol* 2000, **38:**4137-4144.
12. de Lencastre H and Tomasz A: **From ecological reservoir to disease: the nasopharynx, day care centres and drug resistant clones of Streptococcus pneumoniae** *J Antimicrob Chemother* 2002, **50(Suppl C):**75-81.
13. **The EURIS project** [http://www.itqb.unl.pt:1111/euris/]
14. Kristinsson KG: **Effect of antimicrobial use and other risk factors on antimicrobial resistance in pneumococci** *Microb Drug Resist* 1997, **3:**117-123.
15. Boken DJ, Chartrand SA, Moland ES and Goering RV: **Colonization with penicillin-nonsusceptible Streptococcus pneumoniae in urban and rural child-care centers** *Pediatr Infect Dis J* 1996, **15:**667-672.
16. Muhlemann K, Matter HC, Tauber MG and Bodmer T: **Nationwide surveillance of nasopharyngeal Streptococcus pneumoniae isolates from children with respiratory infection, Switzerland 1998–1999** *J Infect Dis* 2003, **187:**589-596.
17. **EURIS Login** [https://www.itqb.unl.pt:1122/]
18. PHP: **Hypertext Preprocessor** [http://www.php.net/].
19. **PostgreSQL** [http://www.postgresql.org/]
20. Perl.com: **The Source for Perl – perl development, perl conferences** [http://www.perl.com/].
21. **The Linux Home Page at Linux Online** [http://www.linux.org/]
22. **The Slackware Linux Project** [http://www.slackware.com/]
23. **Apache-SSL** [http://www.apache-ssl.org/]
24. **Welcome! – The Apache Software Foundation** [http://www.apache.org/]
25. Almeida JS: **Predictive non-linear modeling of complex data by artificial neural networks** *Curr Opin Biotechnol* 2002, **13:**72-76.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6947/3/9/prepub

Chapter III

3. New developments on EURISWEB

3.1. Summary

The aim of this chapter is to present the new functionalities implemented in EURISWEB, in order to receive data from participants of VIth Framework PREVIS project (Pneumococcal Resistance Epidemicity and Virulence - An International Study), which goal is the study of the molecular mechanisms of resistance, virulence and epidemicity in *Streptococcus pneumoniae*. The new developments presented in this chapter are the first step for implementing data analysis algorithms that will allow automated data mining and reporting. Our ultimate goal is to provide the complete infrastructure as an Open Source Epidemiological Information System, which can be easily adapted for different epidemiological studies.

3.2. Introduction

The EURISWEB database (5) is an web-based information system that allow data storage and advanced querying capabilities, such as automatic statistics and report building. Its modular structure was built bearing in mind the necessity to quickly adapt it to new data types that inherently became available as a study progresses and new methodologies and techniques are applied. It was built to allow multi-national data storage and analysis for EURIS project, preserving data security and privacy for the different partners.

PREVIS project (8), also focusing on the study of *Streptococcus pneumoniae*, provided new challenges for EURISWEB, since although the data collection and microbiological testing were similar in many aspects, new questionnaires needed to be implemented and data integration with the previously collected data in EURIS was considered fundamental for the study. EURISWEB had to evolve to a multi-project database while retaining all the data querying and facilities.

The requirements of stability, scalability, security, user-friendly access, low cost portability, and transparent implementation for subsequent independent development, which EURISWEB had, also had to be maintained for this new version.

Presently, the online database is being used by the Laboratory of Molecular Genetics of Instituto de Tecnologia Química and Biológica (ITQB), for the study of the

transmission of antibiotic resistant and susceptible pneumococcal clones among nasopharyngeal carriers(3).

After the data analysis algorithms are implemented and tested using the *Streptococcus pneumoniae* data collected from EURIS and PREVIS projects, we aim to provide the database and data analysis infrastructure as a Open Source Epidemiological Information System (EIS), capable of being adapted by third-parties for different epidemiological studies carried on various infectious agents. Becoming an Open Source project can attract different contributors willing to improve different aspects of the EIS and/or developing new algorithms for data analysis.

3.3. Materials and Methods

3.3.1. Hardware and Software

PREVIS Database setup installed on a a Dual CPU P4 Xeon @ 3.2 Ghz (512 kb of cache), with 4 Gb of SDRAM.

In the migration process, Operative System and software environment majority were also upgraded from the old EURIS infrastructure: Linux server based on Slackware 10, kernel version 2.4.28; Apache Web-Server (version 1.3.33 stable) and PHP 5.1.1 installed as scripting language to generate de HTML code, allowing for improved security as well as the capability for using new features, like simple XML import/export, statistical analysis, or image processing.

These improvements and the way that some new version of PHP functions work, required us to fine tune and sometimes completely remake about 50% of the existing code.

SSL (Secure Socket Layer) 1.55 is also installed, to encrypt and secure web based communications.

The Database Management System infrastructure is now based on PostgreSQL 8.0.3., allowing us to take advantage of some performance tuning, and increased array of features. Shell Scripts that handle the Data Retrieval for the Query System are essentially the same.

The XML data exchange was tested with Bionumerics version 4.5 from Applied Maths, Gent, Belgium, using the Bionumerics scripting language scripts available online from Applied Maths website.

3.4. Results

3.4.1. Database and Interface design

The new data available from PREVIS questionnaires did not require the expansion of the main data model of EURISWEB. The new fields were added to the existing tables on the database.

A new Layout file was created for inserting data on and displaying PREVIS questionnaires. This layout file included the new fields for the PREVIS data not present in EURIS forms. The decision on which layout is used is made on which sampling period is being used to insert data.

Interface design was improved by the use of Cascading Style Sheets (CSS 1.0), to facilitate future design changes required by different projects. Applying different webpage designs to different projects allows quick and accurate identification on which project the data is being inserted.

3.4.2. New Querying capabilities: Crosstab Queries

Since the new data types were inserted in the already built data model of EURISWEB, the use of the User Friendly Query System (UFQS) was maintained and allowed simultaneous querying capability for both EURIS and PREVIS data.

Furthermore, we have implemented in PHP a system of performing queries that emulates the OLAP (OnLine Analytical Processing) (term coined in white paper by Codd *et al*(1)) capabilities of a series of commercial relational database management systems such as MS SQL Server, Oracle or SAP. The relational structure of a database can impose a difficulty for building multidimensional reports, e.g., displaying the serotype distribution over the years, in a table. The Crosstab Query System (CQS) was created to allow the user the creation of such reports. It translates the query built with the interface displayed on Figure 1, to SQL commands that are sent to the database through some shellscripts as previously reported for EURISWEB and produces reports like the ones in Figure 2. Similarly to the UFQS Queries, users can save their queries for posterior use. This results format greatly speeds data analysis, avoiding the time-consuming process of rearranging the data.

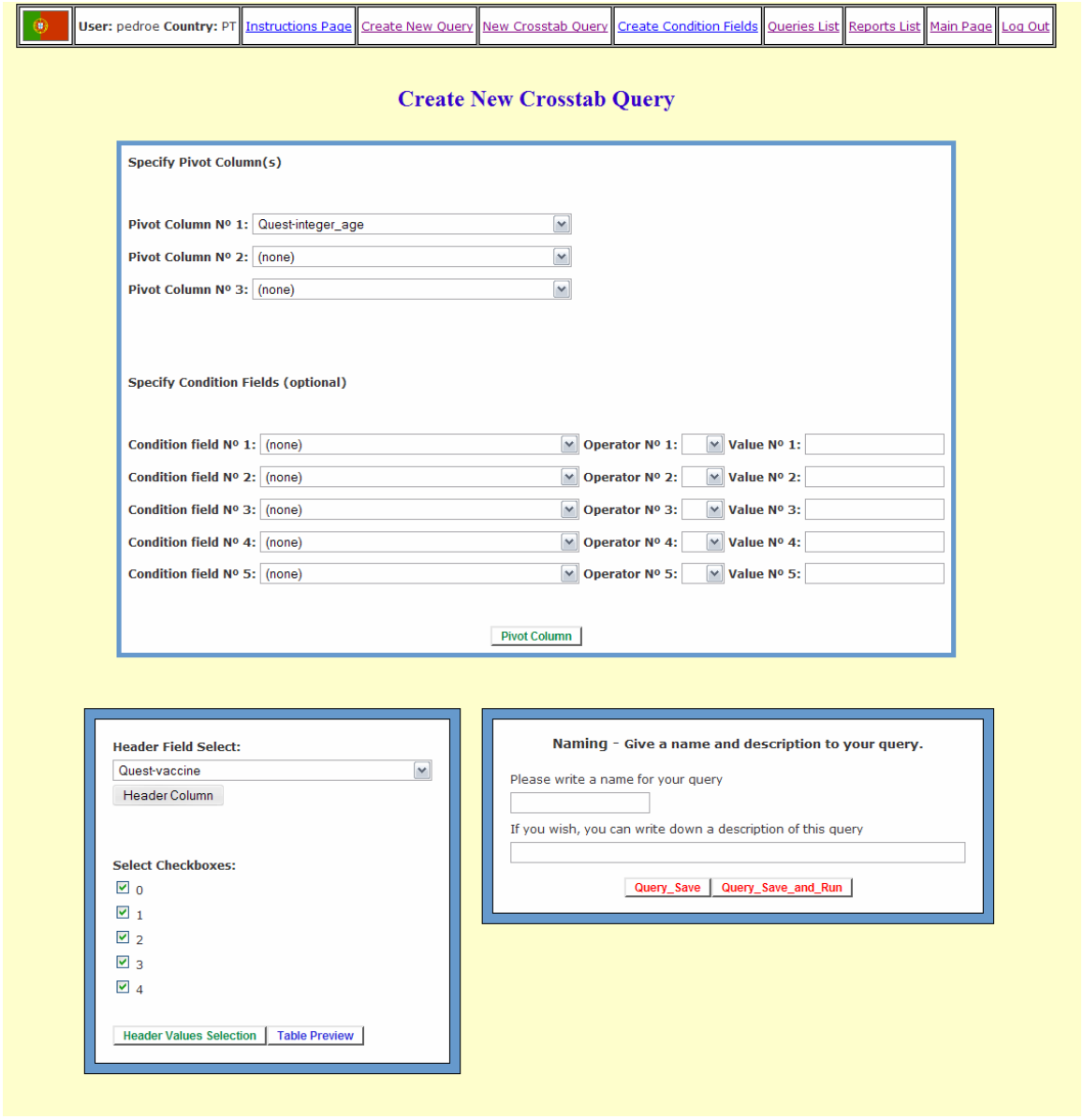
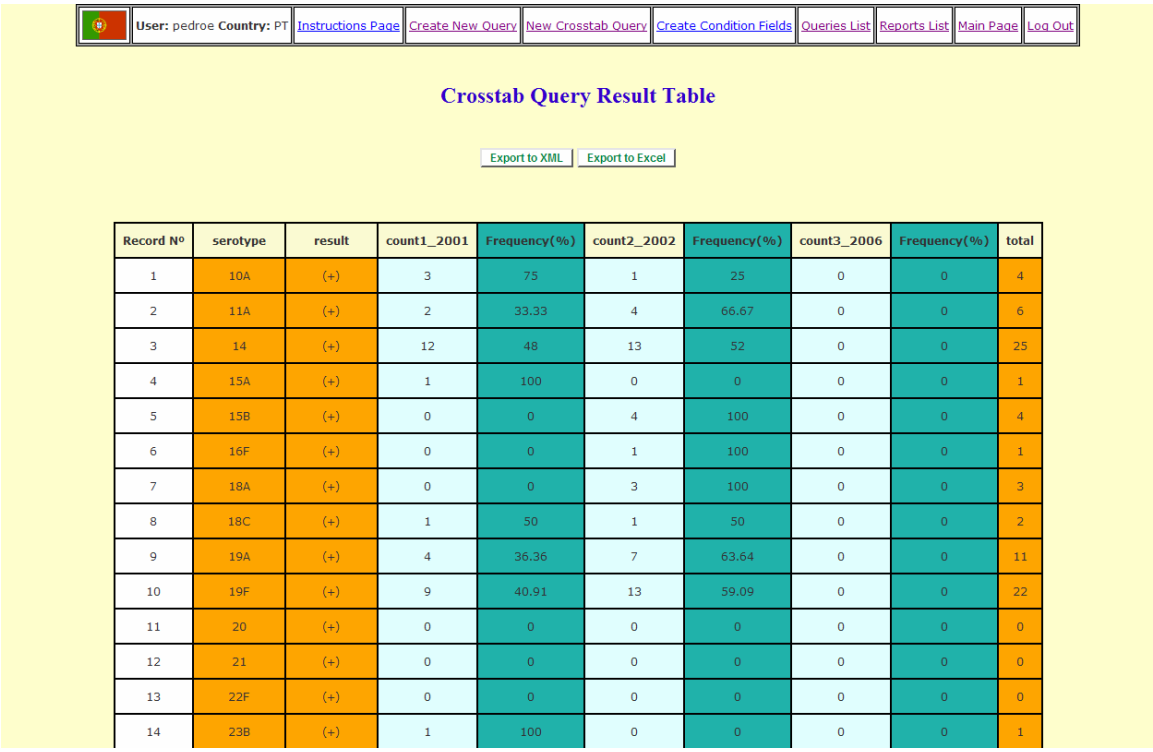


Figure 1 – Crosstab Query System Interface. Here is exemplified the design of a query that gives a report of the number of vaccinated children per age group.



User: pedroe Country: PT | Instructions Page | Create New Query | New Crosstab Query | Create Condition Fields | Queries List | Reports List | Main Page | Log Out

Crosstab Query Result Table

Export to XML | Export to Excel

Record Nº	serotype	result	count1_2001	Frequency(%)	count2_2002	Frequency(%)	count3_2006	Frequency(%)	total
1	10A	(+)	3	75	1	25	0	0	4
2	11A	(+)	2	33.33	4	66.67	0	0	6
3	14	(+)	12	48	13	52	0	0	25
4	15A	(+)	1	100	0	0	0	0	1
5	15B	(+)	0	0	4	100	0	0	4
6	16F	(+)	0	0	1	100	0	0	1
7	18A	(+)	0	0	3	100	0	0	3
8	18C	(+)	1	50	1	50	0	0	2
9	19A	(+)	4	36.36	7	63.64	0	0	11
10	19F	(+)	9	40.91	13	59.09	0	0	22
11	20	(+)	0	0	0	0	0	0	0
12	21	(+)	0	0	0	0	0	0	0
13	22F	(+)	0	0	0	0	0	0	0
14	23B	(+)	1	100	0	0	0	0	1

Figure 2 – Crosstab Query results together with relative frequency statistics. The results can also be exported in XML format and CSV (comma separated values), which can be easily read by a majority of available software such as MS Excel ™ and Bionumerics™. In this example we present the Serotype distribution over the years of study

3.4.3. Data exchange

Although data analysis algorithms are being implemented in EURISWEB, they wouldn't be able to cover all the range of possible data analysis, so the capability of exporting data in a universal format that could be read for the majority of software was implemented. The query system was adapted to export the query results in CSV (comma separated values) and XML (eXtensible Markup Language), that can be imported and manipulated by the majority of the software available.

Also data importing capabilities were implemented in this new version of EURISWEB. Bionumerics (BN) from Applied Maths is the image analysis software of choice of many studies ((2, 4, 6, 7). BN database fields and experiments such as Pulsed-Field Gel Electrophoresis (PFGE) lane images and band allocation can be exported in XML format using the BN scripts. By adding an extra table and using the XML manipulation of PHP, the EURISWEB database is able to accommodate the BN data submitted by a user and link it to the original microbiology sample allowing the visualization of the gel lane and band pattern identified in BN (Figure 3). This new data could be exported back to BN XML format allowing using EURISWEB as a backup for BN data.

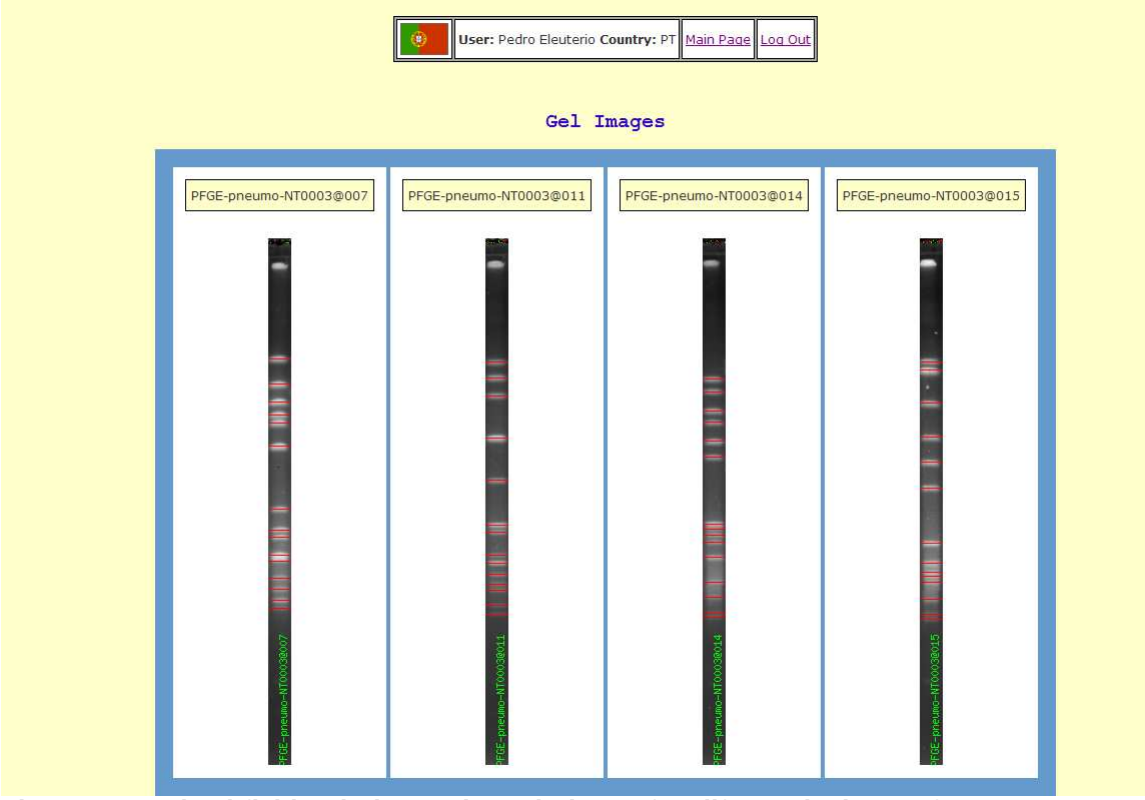


Figure 3 – Pulsed-field gel electrophoresis lanes for diferent isolates of *Streptococcus pneumoniae*, imported from BN XML format to EURISWEB

3.5. Conclusion and future work

EURISWEB design flexibility proved being able to cope with the changes needed to accommodate new data types and new features for data analysis. With the implementation of the Crosstab Query System, the first step was taken for implementing simple measure of association statistics such as Chi-square tests, directly on the database interface: the user could execute a query and automatically a p-value for a chi-square test of independence between the fields chosen to do the crosstab query could be displayed. Also automatic search for correlations between database fields is to be implemented and corresponding reports generated and automatically sent to user.

This ability to integrate data and data analysis is fundamental in the present studies as the technological advances allow researchers to generate much larger quantities of data than ever before. The capacity to quickly provide insights on the data as soon as it is collected and catalogued on the database can free the researcher to pursuit more complex hypothesis or to redesign experiments as they are made.

After these new features are implemented, the whole infrastructure will be made available as an Open Source project, available to anyone who which to adapt it or further develop it for their one studies.

3.6. Acknowledgments

The database work presented here in this paper was done by Pedro Eleutério and João Carriço, under the supervision of Jonas Almeida and Hermínia de Lencastre. Interface security, backups and shellscripts are maintained by António Maretzek. Testing and validation of the new infrastructure was done by Nelson Frazão, Sónia Nunes, Alexandra Simões and Raquel Sá-Leão.

3.7. References

1. **Codd, E. F., C. S.B., and S. C.T.** 1993. Providing OLAP to User-Analysts:: An IT Mandate : http://dev.hyperion.com/resource_library/white_papers/providing_olap_to_user_analysts.pdf. white paper.
2. **Grundmann, H., S. Hori, M. C. Enright, C. Webster, A. Tami, E. J. Feil, and T. Pitt.** 2002. Determining the genetic structure of the natural population of Staphylococcus aureus: a comparison of multilocus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing. J Clin Microbiol **40:**4544-6.
3. **Sa-Leao, R., A. S. Simoes, S. Nunes, N. G. Sousa, N. Frazao, and H. de Lencastre.** 2006. Identification, prevalence and population structure of non-typable Streptococcus pneumoniae in carriage samples isolated from preschoolers attending day-care centres. Microbiology **152:**367-76.
4. **Serrano, I., J. Melo-Cristino, J. A. Carrico, and M. Ramirez.** 2005. Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal. J Clin Microbiol **43:**1706-15.
5. **Silva, S., R. Gouveia-Oliveira, A. Maretzek, J. Carrico, T. Gudnason, K. G. Kristinsson, K. Ekdahl, A. Brito-Avo, A. Tomasz, I. S. Sanches, H. de Lencastre, and J. Almeida.** 2003. EURISWEB--Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers. BMC Med Inform Decis Mak **3:**9.
6. **Sousa, N. G., R. Sa-Leao, M. I. Crisostomo, C. Simas, S. Nunes, N. Frazao, J. A. Carrico, R. Mato, I. Santos-Sanches, and H. de Lencastre.** 2005. Properties of novel international drug-resistant pneumococcal clones identified in day-care centers of Lisbon, Portugal. J Clin Microbiol **43:**4696-703.
7. **Swaminathan, B., T. J. Barrett, S. B. Hunter, and R. V. Tauxe.** 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis **7:**382-9.
8. **URL:, P. p.** 2006, posting date. http://previs.itqb.unl.pt. [Online.]

Chapter IV

4. Assessment of band-based similarity coefficients for automatic Type/Subtype classification of microbial isolates analyzed by Pulsed-Field Gel Electrophoresis

Published in: Carriço, J.A..,F. R. Pinto, C. Simas , S. Nunes, N. G. Sousa; N. Frazão; H. de Lencastre and J. S. Almeida,. J Clin Microbiol, 2005. **43**(11): p. 5483-90.

JOURNAL OF CLINICAL MICROBIOLOGY, Nov. 2005, p. 5483–5490 Vol. 43, No. 11
0095-1137/05/$08.00+0 doi:10.1128/JCM.43.11.5483–5490.2005

Assessment of Band-Based Similarity Coefficients for Automatic Type and Subtype Classification of Microbial Isolates Analyzed by Pulsed-Field Gel Electrophoresis

J. A. Carriço,[1]* F. R. Pinto,[1] C. Simas,[2] S. Nunes,[2] N. G. Sousa,[2] N. Frazão,[2] H. de Lencastre,[2,3] and J. S. Almeida[1,4]

Biomathematics Group[1] and Laboratory of Molecular Genetics,[2] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal; Laboratory of Microbiology, The Rockefeller University, New York, New York[3]; and Department Biostatistics, Bioinformatics, and Epidemiology, Medical University South Carolina, Charleston, South Carolina[4]

Received 5 May 2005/Returned for modification 24 June 2005/Accepted 9 August 2005

Pulsed-field gel electrophoresis (PFGE) has been the typing method of choice for strain identification in epidemiological studies of several bacterial species of medical importance. The usual procedure for the comparison of strains and assignment of strain type and subtype relies on visual assessment of band difference number, followed by an incremental assignment to the group hosting the most similar type previously seen. Band-based similarity coefficients, such as the Dice or the Jaccard coefficient, are then used for dendrogram construction, which provides a quantitative assessment of strain similarity. PFGE type assignment is based on the definition of a threshold linkage value, below which strains are assigned to the same group. This is typically performed empirically by inspecting the hierarchical cluster analysis dendrogram containing the strains of interest. This approach has the problem that the threshold value selected is dependent on the linkage method used for dendrogram construction. Furthermore, the use of a linkage method skews the original similarity values between strains. In this paper we assess the goodness of classification of several band-based similarity coefficients by comparing it with the band difference number for PFGE type and subtype classification using receiver operating characteristic curves. The procedure described was applied to a collection of PFGE results for 1,798 isolates of *Streptococcus pneumoniae*, which documented 96 types and 396 subtypes. The band-based similarity coefficients were found to perform equally well for type classification, but with different proportions of false-positive and false-negative classifications in their minimal false discovery rate when they were used for subtype classification.

Several national and international surveillance studies have been collecting data on the antimicrobial resistance of several bacterial species, namely, *Staphylococcus aureus* and *Streptococcus pneumoniae* (3, 4, 13, 14, 17, 21, 25). In the majority of these studies, pulsed-field gel electrophoresis (PFGE) (20) has been the typing method of choice for clonal type and subtype identification. These large data collection studies provide an excellent resource for the identification of the emergence and the subsequent spread of new clones, which is of particular importance for the tracking of outbreaks as well as obtaining an understanding of the propagation of particular traits, such as resistance to antibiotics. PFGE is also widely used for exchanging clonal identification data between different laboratories, because it has a high interlaboratory reproducibility (6, 17). Its high discriminatory power (24) and relative cost-effectiveness also justify why PFGE is often considered favorably in comparison with complementary typing methods, such as multilocus sequence typing (12).

An enormous variety of band patterns have been found for each bacterial species, with the type and the subtype classification being achieved by the widely used criteria of counting the number of band differences between two lanes proposed by Tenover et al. (23): if two strains differ by up to six bands, counted in both lanes, they are considered the same type. However, these authors pointed out that this method of classification should be used in outbreak studies only and should be backed up with other relevant typing data, such as antibiotic resistance.

Nevertheless, in the majority of longitudinal studies, the use of this criterion (22) yields good discrimination results, particularly when a small number of strains with distinct patterns are being compared (5, 15, 19). This is usually confirmed by visually inspecting the cluster tree to find the cutoff linkage value that agglomerates the band patterns, in accordance with the criteria of Tenover et al. (23).

However, as the number of strains to be clustered increases, this procedure will eventually fail to work because the same difference will span different groupings. This observation is a reflection of the fact that type definitions are arbitrary, in the sense that they reflect the process of strain identification gradually filling a domain of possible band patterns. The loss of a clear distinction between groups produced by hierarchical clustering algorithms (22) will also cause the membership in existing clusters (types) to be rearranged at the previously used cutoff value when a new strain is added to the collection.

A possible solution to the classification instability would be to use a large collection of classified patterns and determine

* Corresponding author. Mailing address: Biomathematics Group, Universidade Nova de Lisboa, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal. Phone: 351 21 446 98 55. Fax: 351 21 442 87 66. E-mail: jcarrico@itqb.unl.pt.

what similarity value produces the best classification results. New strains would be classified by calculating the band similarity to all the entries in the existing catalog and using the highest similarity value determined to recognize membership in the same type. Such a solution would also require the determination of which band similarity coefficient best reproduces the reference classification. Accordingly, in this paper we evaluate the commonly used band-based similarity coefficients—the Dice, Jaccard, Jeffrey's X, Ochiai, Cosine, and Pearson's correlation coefficients—for use for the automatic classification of both type and subtype. The comparison is performed with reference to a collection of 1,798 isolates of *Streptococcus pneumoniae* visually classified, using the criteria of Tenover et al. (23), into 96 types and 396 subtypes. The assessment of goodness of classification of the different similarity coefficients is performed by using receiver operating characteristic (ROC) curves (8) to determine the ability of the different similarity measures to discriminate the visually recognized groups. The method described in this paper highlights the critical value of large visually classified strain collections as the foundation for the computerized automation of their classification. However, once the most effective similarity measure is found, the prospect is raised that the classification itself may be worth redefinition to adjust it to the natural granularity of the microbial population.

MATERIALS AND METHODS

Pulsed-field gel electrophoresis. In this study we analyzed a collection of *Streptococcus pneumoniae* strains collected from children attending day care centers in Lisbon, Portugal, between 2000 and 2004 (9, 21). Chromosomal DNA preparation, restriction with SmaI endonuclease, and PFGE were done as described previously (19).

Visual similarity group (VSG) assignments. PFGE patterns were assigned to types and subtypes by visual inspection of the macrorestriction profiles by using currently accepted criteria (23). Two strains are considered of the same subtype if they have an exact match of band patterns and are considered of the same type if they have up to six band differences on both lanes. In the rare case that a strain could have less than six differences from two types, the type assignment was done by comparison of the strain to all the strains of the two types, and the strain was then assigned to the type with a fewer overall number of band differences. In these cases the type assignment was also supported by other epidemiological information, such as antibiotic resistance patterns and, more recently, multilocus sequence typing information.

The type and subtype names were assigned one or more capital letters. The first pattern identified for a subtype in a type was assigned only a capital letter (e.g., A), and the remaining subtypes were named with capital letters and numbers (e.g., A2 and A3).

Gel analysis. A database of the PFGE patterns was created with Bionumerics software (version 3.0, Applied Maths, Ghent, Belgium). The gel photos were scanned and imported into a Bionumerics database as inverted 8-bit gray-scale TIF images. For each image, spectral analysis included in the software was used to determine the disk size that should be used in "rolling disk" background subtraction (background scale) and the cutoff threshold for least-squares filtering (Wiener cutoff scale). Furthermore, a median filter was used in the image to further smooth the densitometric curves.

After this image preprocessing, intergel and intragel normalizations of the PFGE runs were done with the *S. pneumoniae* R6 strain as a molecular marker. All the gels had three markers: one in the second lane, one lane in the middle, and in the lane before the last. Fifteen bands from 16,320 bp to 340,914 bp were used. The existence of these bands was verified, and their sizes were calculated by virtual digestion of the gel by using a perl script to recognize the restriction sequence of SmaI (CCC^GGG) in the GenBank file of the complete sequence (10). A cubic spline curve was used for the normalization and calibration of each gel. Strain R6 was obtained from the Rockefeller University culture collection.

On all gel images, band assignment was manually curated after automatic band detection. This step is of paramount importance, since there are band intensity variations from gel to gel, which cause errors in the automatic band assignment. Bands ranging from 14 kbp to 400 kbp were considered in this study.

TABLE 1. Band-based similarity coefficients between any two gel band patterns, *i* and *j*

Coefficient	Formula[a]
Dice	$S_{i,j} = \frac{2n_{i,j}}{2n_{i,j} + n_i + n_j}$
Jaccard	$S_{i,j} = \frac{n_{i,j}}{n_{i,j} + n_i + n_j}$
Jeffrey's X	$S_{i,j} = \frac{n_{i,j}}{N_i} + \frac{n_{i,j}}{N_j}$
Ochiai	$S_{i,j} = \frac{n_{i,j}}{\sqrt{(n_{i,j} + n_i)\cdot(n_{i,j} + n_j)}}$

[a] Similarity ($S_{i,j}$) is calculated as described, where n, is the number of bands occurring only in pattern i,n_{ij} is the number of bands shared between the two patterns, and N_i is the total number of bands in pattern i.

The software was then used to calculate the alternative band pattern similarity coefficients. For the 1,798 isolates used in this study, a comparison was created and the corresponding similarity matrices were exported by using the four different band-based similarity coefficients (the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients) and two curve-based correlation coefficients (the Pearson and Cosine coefficients). For the comparative evaluation of the different band-based coefficients, the optimization parameter was evaluated with a range of band position tolerances of from 0% to 8%.

Band-based similarity coefficients. The four most popular band-based similarity coefficients were considered in this study for quantification of the similarities between PFGE band patterns: the Dice (7, 22), Jaccard (22), Jeffrey's X (18a), and Ochiai (18) coefficients (Table 1). All these coefficients exclude negative band matches, which is a necessary compromise, since all possible band positions are unknown.

Also, the Pearson and Cosine correlation coefficients were considered for reference. Generally, these two methods yield lower similarity values than band-based methods, since they take into account all the densitometric curve values, which causes them to be more sensitive to small variations. This makes them the methods of choice for the comparison of strain similarity by typing methods in which the intensities of bands are to be considered, such as AFLP (16, 26), but they can produce erroneous conclusions when only the presence or the absence of a band is important and the band intensity varies among the strains being compared.

ROC curves. ROC curves were used to assess the classification by use of the different similarity coefficients. This method, created in signal detection theory, is frequently used in classification problems and is widely applied in medical diagnosis and psychometric analysis (8). This method is commonly employed for the binary classification of continuous data, usually categorized as positive and negative cases. In our study, the correct classification was considered the VSG assignment; for each coefficient, the VSG assignment thus classified each case at each threshold as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The classification accuracy of each coefficient was then measured by plotting for the different threshold values the ratio of the number of true-positive classifications over the total number of positive classifications, also named the sensitivity or the true-positive rate, versus the false-positive rate, or 1 − specificity (Table 2), which is the ROC curve.

The area under a ROC curve (AUC) is the parameter employed to quantify the goodness of classification of the classifier being tested, since it is a threshold-independent performance measure. For a perfect classifier the AUC is 1, and for a random classifier the AUC is 0.5. Additional results and a comprehensive discussion of the AUC measure are provided elsewhere (1, 2).

RESULTS

The PFGE patterns of 1,798 distinct strains of *S. pneumoniae* were visually classified into 96 types and 396 subtypes, with the assistance of Bionumerics software, at the Laboratory of Molecular Genetics of Instituto de Tecnologia Química e

TABLE 2. ROC curve parameters

Parameter	Formula[a]
Sensitivity, or true-positive rate	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
1 − specificity, or false-positive rate	$\frac{FP}{TN + FP}$

[a] TP, number of samples with a true-positive classification; TN, number of samples with a true-negative classification; FP, number of samples with a false-positive classification; FN, number of samples with a false-negative classification.

Biológica. This manually curated repository was analyzed for objective assessment of alternative measures of similarity for automation of PFGE-based classification of new isolates. Figure 1 displays the visual classification of strains into types and subtypes along with the normalized patterns. The strains are sorted alphabetically according to the naming protocol detailed in the Materials and Methods section. The 10 most represented types are delimited by vertical lines.

The VSG classification of the PFGE band patterns into types and subtypes was used as a reference to assess the goodness of classification of the similarity coefficients typically considered in comparisons of PFGE band patterns: the Dice, Jaccard, Jeffrey's X, Ochiai, Pearson, and Cosine coefficients (Table 1). This assessment was first performed by using ROC curves, which measure the classification by plotting, for different similarity coefficient threshold values, the ratio of TP matches over the total number of positive samples versus the ratio of FP matches over the total number of negative samples. The AUC was then used as the threshold independent measure of goodness of the classification (see Materials and Methods). Second, for each band-based similarity coefficient, different band position tolerance settings were compared to determine the optimal parameter settings, i.e., band position tolerance values. In this study, this is the most important parameter for accurate band matching between two different lanes.

For example, in Fig. 2, ROC curves are plotted for the comparison of the visual type assignments of the band and Dice coefficient values for different band position tolerance settings. This illustrates how best the tolerance value for the Dice coefficient can be determined. The table inset in Fig. 2 provides the corresponding AUC values. The Pearson correlation coefficient AUC value is also included to illustrate the relative inefficient classification of correlation similarity coefficients (AUC of 0.901 versus an AUC up to 0.984 for the Dice coefficient). Even worse performance was found for the Cosine correlation coefficient, with an AUC value of 0.882 (not plotted).

The goodness of classification, as assessed by the AUC, for the alternative similarity coefficients considered in this study is represented in Fig. 3. All the band-based similarity coefficients (the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients) behave remarkably similarly in the classification of types and subtypes. As was to be expected, when properly selected band position tolerance values are used, band-based similarity coefficients have higher AUC values than correlation coefficients.

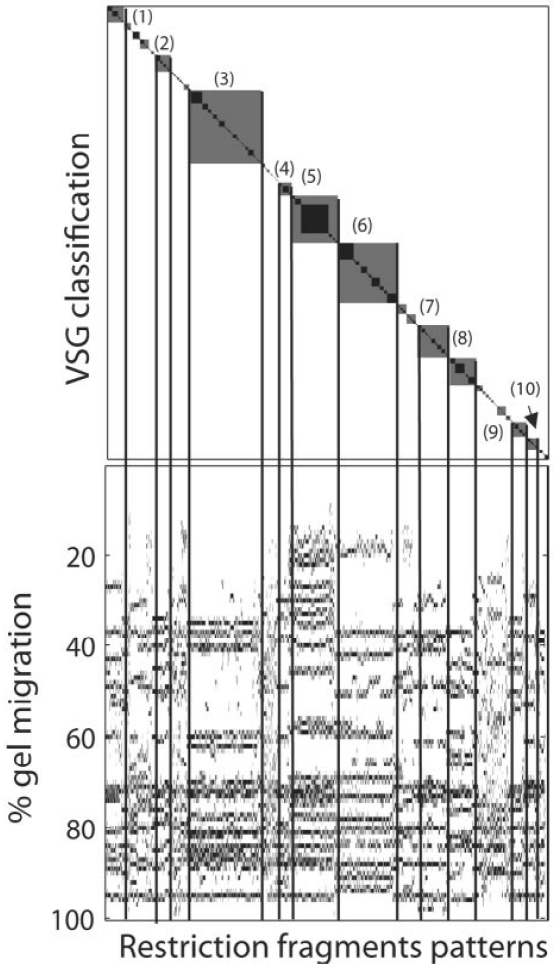


FIG. 1. Representation of VSG classification and band patterns for the 1,798 strains of *S. pneumoniae*. In the upper part (visual similarity group classification matrix), the black areas represent PFGE subtypes and the gray areas represent PFGE types. The most represented groups (PFGE types) are (point 1) A (67 isolates), (point 2) AO (65 isolates), (point 3) B (292 isolates), (point 4) DDD (51 isolates), (point 5) E (187 isolates), (point 6) FF (238 isolates), (point 7) M (131 isolates), (point 8) MM (107 isolates), (point 9) R (57 isolates), and (point 10) SI (47 isolates). The lower part of the figure includes the corresponding PFGE band patterns. The lines were drawn to help the reader isolate the PFGE patterns visually.

As shown in Fig. 2 (and also in Fig. 3B), for type classification the optimal band position tolerance was found to be 1.7% for all band-based similarity coefficients, with an AUC of 0.984. For subtype classification (Fig. 3A), the optimal settings were found for higher band position tolerance values, 2.5%, which also correspond to a higher AUC of 0.995, which is the same for all band-based similarity coefficients Again, correlation coefficients yielded a lower AUC of 0.906 for the Pearson correlation coefficient and 0.898 for the Cosine coefficient.

Although the different band-based similarity coefficients are surprisingly equivalent regarding the goodness of classification, the proportions of true-positive and false-positive subtype clas-

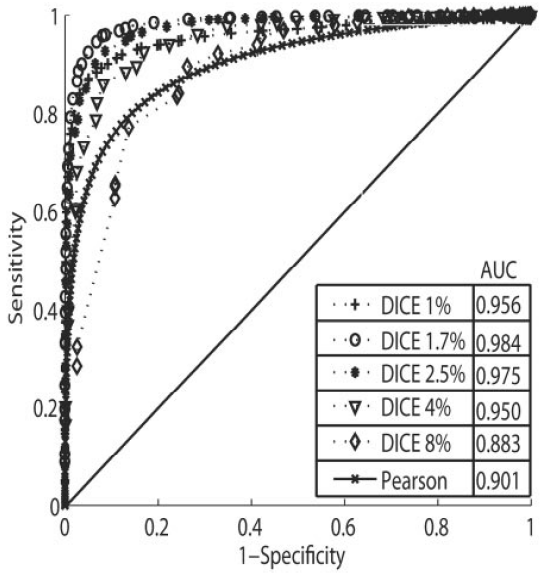


FIG. 2. ROC curves for several band position tolerances of the Dice coefficient in type classification. The maximum AUC value, 0.984, was found for a band position tolerance of 1.7%. The random classification (straight diagonal; AUC, 0.5) and the underperforming Pearson's correlation coefficient (AUC, 0.901) are plotted for reference.

sifications differ. Figure 3C and D represents the contribution of false-positive or false-negative classifications on the total classification error.

For each band position tolerance, the point where the similarity coefficient threshold had a minimum absolute classification error (a minimum of false-positive plus false-negative classifications) was plotted. For example, by using the Dice coefficient for type classification, the similarity threshold value with minimal classification error was found to be 81% for a 1.7% band position tolerance (Fig. 4).

For subtype classification at a 2.5% band position tolerance, the Dice and Jaccard coefficient (Fig. 3C) classifications resulted in fewer false-negative classifications and, conversely, more false-positive classifications than the Ochiai and Jeffrey coefficients. The same is true for the absolute numbers of misclassifications (data not shown).

Regarding the type classification, band-based similarity coefficients also performed equally well (Fig. 3B), but the heterogeneity of band patterns included in each type is reflected by the persistence of false-negative classifications for wider band position tolerance values. At a band position tolerance of 1.7%, the four band-based similarity coefficients are nearly indistinguishable in terms of the contribution of false-positive and false-negative classifications to the type classification error.

These calculated optimal position tolerance settings apply only to the data analyzed in this study, although it is a very good starting point for data obtained by the same PFGE protocol, since the running conditions should be similar and should generate similarly resolved band patterns.

As suggested by the results plotted in Fig. 3, the fact that the four band-based similarity coefficients performed equally well for the same band position tolerance implies that there are equivalent, but not necessarily similar, threshold values between each of the band pattern similarity measures. This equivalence is confirmed in Fig. 4, where, for optimal band tolerance (1.7% for type; 2.5% for subtype), the ROC curves and corresponding threshold values are displayed. Figure 4, as discussed in the next section, can be used to determine the appropriate threshold values for the desired proportion of false-positive and false-negative classifications in the total classification error. Figure 4 can be analyzed to produce optimal threshold values for arbitrary cost-benefit ratios. For example, if FP and FN classifications are equally undesirable, the four band-based similarity coefficient should be used with the band identity tolerance values indicated in Table 3.

DISCUSSION

The classification of *Streptococcus pneumoniae* isolates by PFGE has followed the typical method of visual recognition of similar patterns by the absolute number of band differences within existing isolates. New isolates are classified by incrementally assigning them to a type or a subtype of the previously classified isolates already described in databases. This solution pragmatically produces guidelines for group recognition, as prescribed by the widely used criteria of Tenover et al. (23) detailed in the introduction. However, when enough isolates have been processed in this fashion, the collection of results can be analyzed to identify an equivalent computational procedure. In order to achieve that goal of automation of manual classification, it is necessary to assess alternative metrics to quantify band pattern dissimilarity and also to determine its most discriminant settings: the band identity tolerance and the similarity threshold value for positive classification in the same group as another band pattern. This work was made possible by the extensive collection of *S. pneumoniae* isolates that had been manually annotated. Accordingly, the collection of 1,798 *S. pneumoniae* isolates was analyzed for determination of the settings that maximize the goodness of classification by use of the alternative band-based similarity coefficients—the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients—and also, for reference, densitometric curve-based correlation coefficients—the Pearson and Cosine coefficients.

As expected, discrete band-based similarity coefficients clearly outperformed the correlation coefficients, leading to a much higher goodness of classification, as assessed by the area under the ROC curve. Surprisingly, all of the band-based similarity coefficients tested were found to be equally discriminant for both type and subtype (Fig. 3). That is, all of the four similarity coefficient band-based formulations (Table 1) will produce the same percentage of erroneous classifications for a given band identity tolerance value (Fig. 3A and B). However, this does not necessarily imply that the erroneous classifications will include the same proportion of false-negative and false-positive classifications.

As noted above, the results presented in Fig. 3 for the dependence of goodness of classification, as assessed by the corresponding AUC value, suggest not only that the four band-based methods will perform equally well but also that they will perform equally well for the same band tolerance values (Fig. 3A and B). This observation was observed to be valid for both

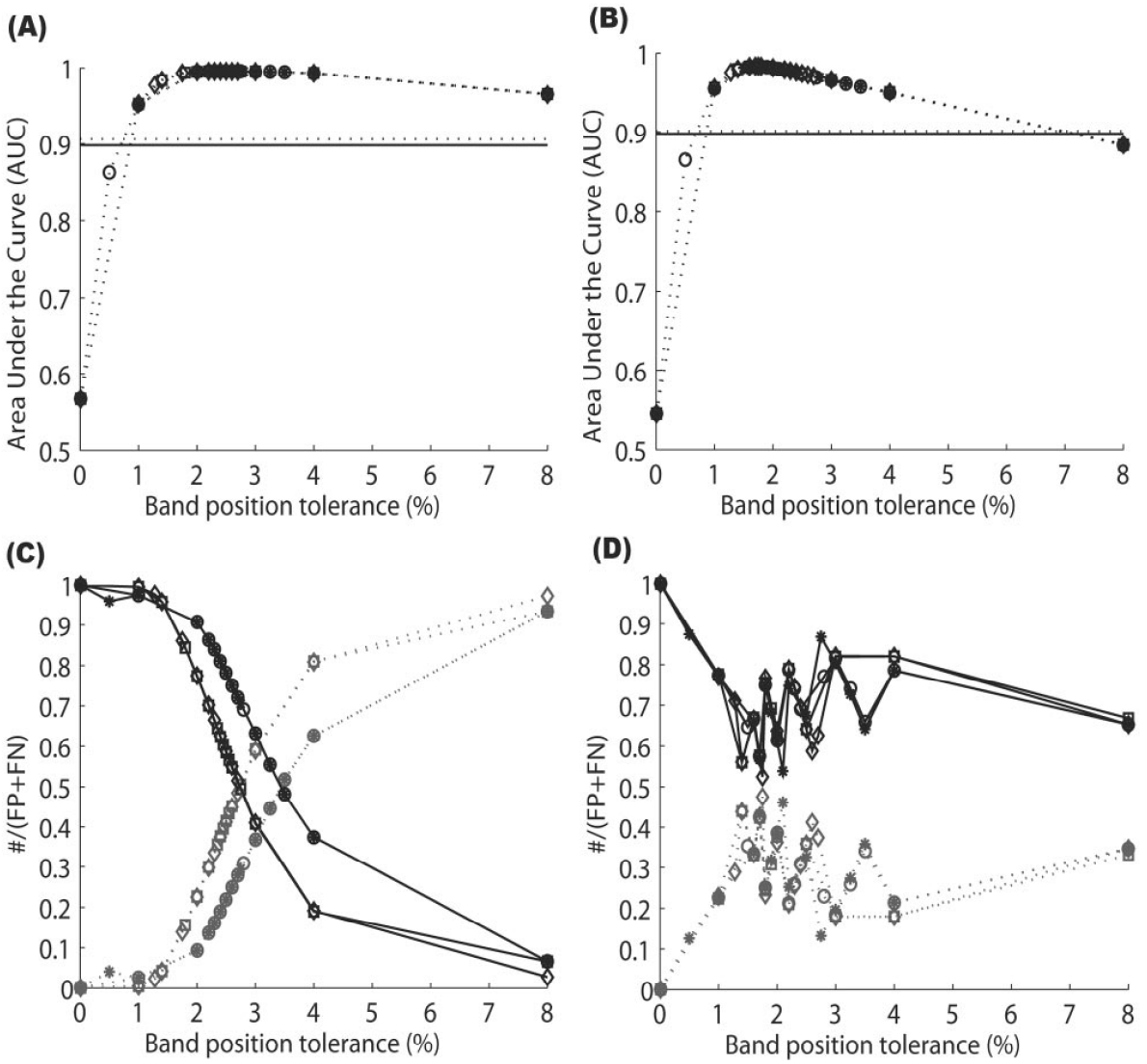


FIG. 3. Area under the curve of ROC curves of the coefficients tested for different band position tolerances for subtype (A) and type (B) classification. Contribution of false-positive and false-negative classifications for the total classification error in subtype (C) and type (D). The Dice coefficient is identified by squares, the Jaccard coefficient is identified by diamonds, the Ochiai coefficient is identified by asterisks, the Jeffrey's X coefficient is identified by circles, the Pearson coefficient is identified by a dotted line without markers, and the Cosine coefficient is identified by a solid line without markers. For panels C and D, FP classifications are represented by gray dotted lines, and FN classifications are represented by black solid lines.

type and subtype classifications. However, inspection of the corresponding proportions of FP and FN classifications (Fig. 3C and D) shows that, for subtype classifications (Fig. 3C), the Dice and the Jaccard coefficients will yield comparatively more FP classifications and fewer FN classifications than Jeffrey's X or the Ochiai coefficient. This distinction is the most pronounced when the goodness of classification (AUC) is maximal. It is also interesting that for subtype classification with exaggerated band identity tolerance values, the erroneous classifications will be heavily dominated by false-positive classifications. In contrast, neither of these observations is valid for type classification (Fig. 3D), where the proportion of false-positive and false-negative classifications is not noticeably different between the band-based methods assessed, and high band tolerance values do not cause false-positive classifications to predominate. It is also noteworthy that the proportions themselves (Fig. 3D) are somewhat erratic, which is a reflection of the fact that any of the two band patterns classified as the same type can have up to six band differences, allowing for a great heterogeneity of patterns.

The discussion above highlights the observation that if bands that discriminate between types are in close proximity to each other and are possibly bands of lower molecular size (from approximately 19 kbp to 100 kbp), misclassification will even-

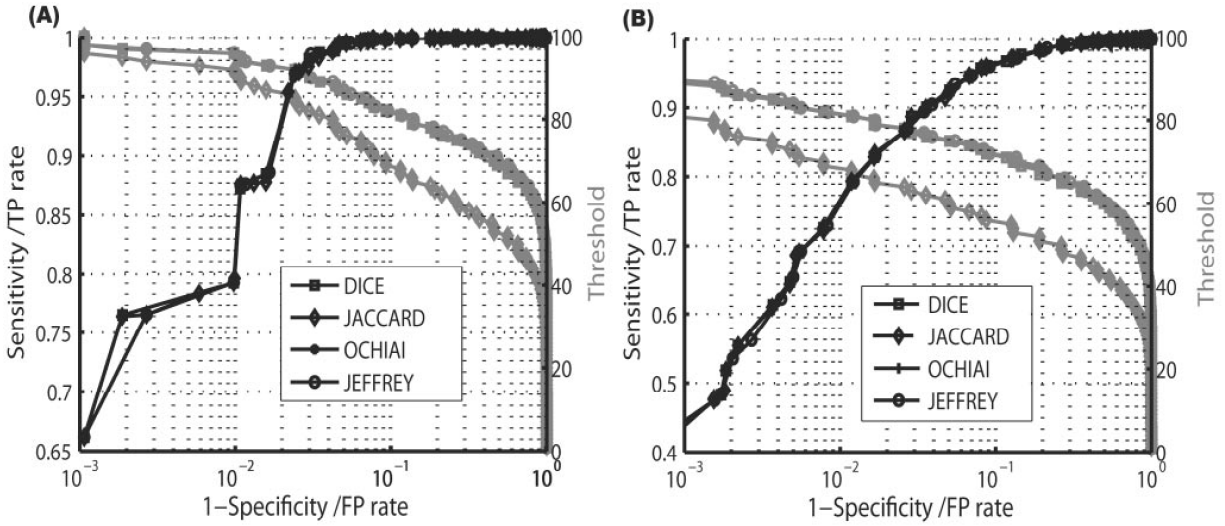


FIG. 4. ROC curves and threshold representation for subtype (A) and type (B). This figure allows the choice of a threshold value as a function of the false-positive rate/true-positive rate, for the optimal band position tolerance settings that provide a maximum discrimination between types. Note that the false-positive rate (which corresponds to 1 − specificity) is represented on a logarithmic scale.

tually occur as more subtypes are identified. This heterogeneity of patterns for isolates of the same type and the blurring of arbitrary type distinctions by new isolates justify why the number of false-negative classification contributions did not decrease for higher band identity tolerance values. This is in sharp contrast to what happens in subtype classification (Fig. 3C), where band patterns should be exactly equal (band-based similarity coefficient value of 100%) between strains of the same subtype. In practice, experimental conditions can cause small distortions in the gel that are not compensated for by software or visual classification, and that is why the determination of the optimal band identity tolerance is critical for automation of band pattern classification of subtypes. Accordingly, the similarity levels for strains of the same subtype oscillate in the 95 to 100% interval, even after selection of an optimal band position tolerance setting. These results suggest that while automation of the classification of PFGE band patterns of visually recognized subtypes and types is achieved with considerable accuracy by the proposed method (maximum AUC values of 0.9954 and 0.9837, respectively, for the Dice coefficient), visual assignment mostly delimits arbitrary groupings of subtype patterns. On the contrary, the automated classification of PFGE band patterns of subtypes confines defined groups where the band positions oscillate only very slightly around a reference value.

TABLE 3. Threshold similarity values for the point where there are a minimum of misclassifications (minimum of false positives and false negatives) of subtype and type

Coefficient	Subtype			Type		
	FP rate	TP rate	Threshold	FP rate	TP rate	Threshold
Dice	0.002	0.76	97	0.012	0.79	81
Jaccard	0.002	0.76	95	0.012	0.79	67
Jeffrey's X	0.001	0.66	98	0.012	0.79	81
Ochiai	0.001	0.66	98	0.012	0.79	81

The immediately useful result of this paper is delivered in Fig. 4. It plots, for the optimal band position tolerance value, the logarithm of the false-positive rate versus the true-positive rate and the respective threshold values. The logarithm of the false-positive rate provides easier reading of the values for the lower false-positive rates (from 0.001 to 0.1). Figure 4 allows the choice of a threshold value as a function of the false-positive rate/true-positive rate for the optimal band position tolerance settings that provide a maximum discrimination (measured by the AUC). This choice weighs the relative cost of having a false-positive or a false-negative classification. For example, if the objective of a study is to recognize membership in a specific PFGE type, the threshold should be chosen to minimize the number of false-negative assignments. If the Dice similarity coefficient was the metric chosen and the acceptable false-positive classification was only 1%, then the threshold obtained by inspecting Fig. 4 would be about 80%. If, instead, the goal was the maximization of the true discovery rate, then the appropriate threshold for the same method would be 97% (this result is also listed in Table 3). This exercise also illustrates the conclusion that although the similarity coefficients perform equally well, they are not interchangeable, as different proportions of false-negative and false-negative classifications may result. Conversely, Fig. 4 can also be used to determine what threshold values will render two similarity coefficients equally discriminant for the optimal band position tolerance value.

Over the past few years large databases of genotyped clinical strains have been assembled. These repositories contain a unique record documenting both the diversity and the dynamics of the emergence of new strains. Furthermore, it has been consistently shown that PFGE has a higher discriminatory power than newer sequence-based methods, such as multilocus

sequence typing, which justifies the prospect that the cost-effective use of PFGE will be seamlessly integrated with other genotyping methods in even larger repositories. In that regard, the study reported here leads to the following conclusions.

First, we have found that the perception that band-based similarity coefficients are superior to correlation methods is correct, provided that they are correctly parameterized. This observation puts a prize not only on the correct parameterization method but also on the use of robust image analysis software for gel lane alignment and band recognition.

Second, we have used a repository of 1,798 PFGE types isolates of *S. pneumoniae* to assess the relative merits of the different band-based similarity coefficients: the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients. Surprisingly, they were all found to be equally able to classify the isolates from the reference database, with equivalent performances occurring for distinct thresholds but the same band position tolerances. The goodness of classification was assessed by use of the AUC of the ROC curve.

Third, the equivalence in AUC with the same proportion of erroneous classifications was found to correspond to different proportions of false-positive and false-negative classifications, which will play a role in the selection of a similarity coefficient for use in a fully automated bioinformatic implementation. Consequently, the assessment and parameterization of PFGE similarity coefficients are delivered as ROC curve plots with the corresponding threshold values (Fig. 4), where the cost-benefit assigned to the different types of erroneous classifications can be weighted quantitatively and the most appropriate method and threshold values can be selected.

Fourth, the automated procedure was found to perform satisfactorily, with an optimal AUC of 0.984. This result supports the conclusion that the implementation of automated classification is highly advantageous, particularly since multiparametric statistics can be associated to select those patterns that warrant subsequent visual inspection.

The optimal parameterization of band-based similarity coefficients opens the prospect of revisiting the identification of types as a dynamic entity defined by unsupervised classification algorithms such as nearest means (*K* means) or self-organized maps. Therefore, the identification of similarity metrics that reproduce and automate the classification of typing results enables the redefinition of heterogeneous types in *S. pneumoniae* with time-dependent identities that converge to the confinements of the natural populations as more isolates are characterized. The tracking of how the definitions evolve could be solved automatically by the implementation of repositories that can be queried by use of the shortest similarity coefficient value. The methods used in this paper can be used in any database to determine which similarity metric is more adequate to describe the data and also which parameters optimize the classification procedure.

ACKNOWLEDGMENTS

We thank Alexander Tomasz, The Rockefeller University, for the gift of strain *S. pneumoniae* R6. We also acknowledge Susana Vinga for help on ROC curves and Luc Vauterin for bibliographic help on the similarity coefficients.

Partial support for this work was provided by contracts EURIS (QLK2-CT-2000-01020) and PREVIS (LSHM-CT-2003-503413 from the European Community) awarded to H. de Lencastre and J. S. Almeida. J. A. Carriço and F. R. Pinto were supported by grants SFRH/BD/3123/2000 and SFRH/BD/6488/2001, respectively, both from the Fundação para a Ciência e Tecnologia of Portugal. S. Nunes and N. G. Sousa were supported by grants 011/BIC/01 and 043/BIC/00, respectively, from contract QLK2-CT-2000-01020; S. Nunes, N. G. Sousa, and N. Frazão have been supported since March 2004 by grants 010/BIC/2004, 009/BIC/2004, and 011/BIC/2004, respectively, from contract LSHM-CT-2003-503413. C. Simas was supported by a grant from IBET, project WLP (grant 31 CEM/NET); and N. Frazão was also supported by IBET grant 28/12/02.

REFERENCES

1. **Baldi, P., S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen.** 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics **16:**412–424.
2. **Bradley, A. P.** 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30:**1145–1159.
3. **Bronzwaer, S. L., U. Buchholz, J. L. Kool, J. Monen, and P. Schrijnemakers.** 2001. EARSS activities and results: update. Euro. Surveill. **6:**2–5.
4. **Bronzwaer, S. L., O. Cars, U. Buchholz, S. Molstad, W. Goettsch, I. K. Veldhuijzen, J. L. Kool, M. J. Sprenger, and J. E. Degener.** 2002. A European study on the relationship between antimicrobial use and antimicrobial resistance. Emerg. Infect. Dis. **8:**278–282.
5. **Brueggemann, A. B., D. T. Griffiths, E. Meats, T. Peto, D. W. Crook, and B. G. Spratt.** 2003. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J. Infect. Dis. **187:**1424–1432.
6. **Chung, M., H. de Lecanstre, P. Matthews, A. Tomasz, I. Adamsson, M. Aires de Sousa, T. Camou, C. Cocuzza, A. Corso, I. Couto, A. Dominguez, M. Gniadkowski, R. Goering, A. Gomes, K. Kikuchi, A. Marchese, R. Mato, O. Melter, D. Oliveira, R. Palacio, R. Sá-Leão, I. Santos Sanches, J.-H. Song, P. T. Tassios, and P. Villari.** 2000. Molecular typing of methicillin-resistant *Staphylococcus aureus* by pulsed-field gel eletrophoresis: comparison of results obtained in a multilaboratory effort using identical protocols and MRSA strains. Microb. Drug Resist. **6:**189–198.
7. **Dice, L. R.** 1945. Measures of the amount of ecological association between species. Ecology **26:**297–302.
8. **Egan, J. P.** 1975. Signal detection theory and ROC-analysis. Academic Press, Inc., New York, N.Y.
9. **Frazão, N., A. Brito-Avô, C. Simas, J. Saldanha, R. Mato, S. Nunes, N. G. Sousa, J. A. Carriço, J. S. Almeida, I. Santos-Sanches, and H. de Lencastre.** 2004. Effect of the seven-valent conjugate pneumococcal vaccine on carriage and drug resistance of *Streptococcus pneumoniae* in healthy children attending day-care centers in Lisbon. Pediatr. Infect. Dis. J. **24:**243–252.
10. **Hoskins, J., W. E. Alborn, Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck, Jr., P. L. Skatrud, and J. I. Glass.** 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. **183:**5709–5717.
11. Reference deleted.
12. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 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.
13. **McDougal, L. K., C. D. Steward, G. E. Killgore, J. M. Chaitram, S. K. McAllister, and F. C. Tenover.** 2003. Pulsed-field gel electrophoresis typing of oxacillin-resistant *Staphylococcus aureus* isolates from the United States: establishing a national database. J. Clin. Microbiol. **41:**5113–5120.
14. **McGee, L., L. McDougal, J. Zhou, B. G. Spratt, F. C. Tenover, R. George, R. Hakenbeck, W. Hryniewicz, J. C. Lefevre, A. Tomasz, and K. P. Klugman.** 2001. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. J. Clin. Microbiol. **39:**2565–2571.
15. **Miragaia, M., I. Couto, S. F. Pereira, K. G. Kristinsson, H. Westh, J. O. Jarlov, J. Carrico, J. Almeida, I. Santos-Sanches, and H. de Lencastre.** 2002. Molecular characterization of methicillin-resistant *Staphylococcus epidermidis* clones: evidence of geographic dissemination. J. Clin. Microbiol. **40:**430–438.
16. **Mueller, U. G., and L. L. Wolfenbarger.** 1999. AFLP genotyping and fingerprinting. Trends Ecol. Evol. **14:**389–394.
17. **Murchan, S., M. E. Kaufmann, A. Deplano, R. de Ryck, M. Struelens, C. E. Zinn, V. Fussing, S. Salmenlinna, J. Vuopio-Varkila, N. El Solh, C. Cuny, W. Witte, P. T. Tassios, N. Legakis, W. van Leeuwen, A. van Belkum, A. Vindel, I. Laconcha, J. Garaizar, S. Haeggman, B. Olsson-Liljequist, U. Ransjo, G. Coombes, and B. Cookson.** 2003. Harmonization of pulsed-field gel electro-

phoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. J. Clin. Microbiol. **41:**1574–1585.

18. **Ochiai, A.** 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jpn. Soc. Fish Sci. **22:**526–530.

18a. **Pena, S. D. J., R. Chakraborty, J. T. Epplen, and A. J. Jeffreys.** 1993. DNA fingerprinting: the state of the science, p. 1–19. Birkhauser Verlag, Basel, Switzerland.

19. **Sá-Leão, R., A. Tomasz, I. Santos-Sanches, S. Nunes, C. R. Alves, A. B. Avo, J. Saldanha, K. G. Kristinsson, and H. de Lencastre.** 2000. Genetic diversity and clonal patterns among antibiotic-susceptible and - resistant *Streptococcus pneumoniae* colonizing children: day care centers as autonomous epidemiological units. J. Clin. Microbiol. **38:**4137 4144.

20. **Schwartz, D. C., and C. R. Cantor.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:**67–75.

21. **Silva, S., R. Gouveia-Oliveira, A. Maretzek, J. Carrico, T. Gudnason, K. G. Kristinsson, K. Ekdahl, A. Brito-Avo, A. Tomasz, I. S. Sanches, H. de Lencastre, and J. Almeida.** 2003. EURISWEB—Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers. BMC Med. Inform. Decision Making **3:**9.

22. **Sneath, P. H., and R. R. Sokal.** 1973. Numerical taxonomy. W. H. Freeman & Co., San Francisco, Calif.

23. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. **33:**2233–2239.

24. **van Belkum, A., M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc.** 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin. Microbiol. Rev. **14:**547–560.

25. **van Belkum, A., W. van Leeuwen, M. E. Kaufmann, B. Cookson, F. Forey, J. Etienne, R. Goering, F. Tenover, C. Steward, F. O'Brien, W. Grubb, P. Tassios, N. Legakis, A. Morvan, N. El Solh, R. de Ryck, M. Struelens, S. Salmenlinna, J. Vuopio-Varkila, M. Kooistra, A. Talens, W. Witte, and H. V. L. Verbrugh.** 1998. Assessment of resolution and intercenter reproducibility of results of genotyping *Staphylococcus aureus* by pulsed-field gel electrophoresis of SmaI macrorestriction fragments: a multicenter study. J. Clin. Microbiol. **36:**1653–1659.

26. **Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, et al.** 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:**4407–4414.

Chapter V

5. Illustration of a Common Framework for Relating Multiple Typing Methods by Application to Macrolide-Resistant *Streptococcus pyogenes*

Published in: Carriço, J.A.; C. Silva-Costa; J. Melo-Cristino; F. R. Pinto; H. de Lencastre; J.S.Almeida; M.Ramirez, Illustration of a Common Framework for Relating Multiple Typing Methods by Application to Macrolide-Resistant Streptococcus pyogenes, J Clin Microbiol. 2006 Jul;44(7):2524-32

JOURNAL OF CLINICAL MICROBIOLOGY, July 2006, p. 2524–2532 Vol. 44, No. 7
0095-1137/06/$08.00+0 doi:10.1128/JCM.02536-05

Illustration of a Common Framework for Relating Multiple Typing Methods by Application to Macrolide-Resistant *Streptococcus pyogenes*†

J. A. Carriço,[1]* C. Silva-Costa,[2] J. Melo-Cristino,[2] F. R. Pinto,[1,2] H. de Lencastre,[4,5] J. S. Almeida,[1,3] and M. Ramirez[2]

Grupo de Biomatemática, Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal[1]; Instituto de Medicina Molecular e Instituto de Microbiologia, Faculdade de Medicina de Lisboa, Lisbon, Portugal[2]; Department of Biostatistics and Applied Mathematics, University of Texas MD Anderson Cancer Center, Houston, Texas[3]; Laboratório de Genética Molecular, Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal[4]; Laboratory of Microbiology, The Rockefeller University, New York, New York[5]

Received 6 December 2005/Returned for modification 14 March 2006/Accepted 10 May 2006

The studies that correlate the results obtained by different typing methodologies rely solely on qualitative comparisons of the groups defined by each methodology. We propose a framework of measures for the quantitative assessment of correspondences between different typing methods as a first step to the global mapping of type equivalences. A collection of 325 macrolide-resistant *Streptococcus pyogenes* isolates associated with pharyngitis cases in Portugal was used to benchmark the proposed measures. All isolates were characterized by macrolide resistance phenotyping, T serotyping, *emm* sequence typing, and pulsed-field gel electrophoresis (PFGE), using SmaI or Cfr9I and SfiI. A subset of 41 isolates, representing each PFGE cluster, was also characterized by multilocus sequence typing (MLST). The application of Adjusted Rand and Wallace indices allowed the evaluation of the strength and the directionality of the correspondences between the various typing methods and showed that if PFGE or MLST data are available one can confidently predict the *emm* type (Wallace coefficients of 0.952 for both methods). In contrast, *emm* typing was a poor predictor of PFGE cluster or MLST sequence type (Wallace coefficients of 0.803 and 0.655, respectively). This was confirmed by the analysis of the larger data set available from http://spyogenes.mlst.net and underscores the necessity of performing PFGE or MLST to unambiguously define clones in *S. pyogenes*.

Typing methods are major tools for the epidemiological characterization of bacterial pathogens, allowing the determination of the clonal relationships between isolates based on their genotypic or phenotypic characteristics. Recent technological advances have resulted in a shift from classical phenotypic typing methods, such as serotyping, biotyping, and antibiotic resistance typing, to molecular methods such as restriction fragment length polymorphism (8), pulsed-field gel electrophoresis (PFGE) (25), and PCR serotyping (4). With the availability of affordable sequencing methods, another shift occurred towards sequence-based typing methods such as multilocus sequence typing (MLST) (18) and *emm* sequence typing (2). Sequence-based methods have a wide appeal since they provide unambiguous data and are intrinsically portable, allowing the creation of databases that, if publicly available through the internet, enable the comparison of local data with those of previous studies in different geographical locations. Ideally, an analysis of each typing method, in terms of discriminatory power, reproducibility, typeability, feasibility, and other characteristics as suggested by Struelens (31), should be performed to better determine which method is appropriate in a given setting.

Several molecular epidemiology studies of clinically relevant microorganisms provide a characterization of isolates based on different typing methods (6, 8, 20, 23). Frequently these studies focus on a comparison between the assigned types of different typing methods, from a qualitative point of view, i.e., indicating correspondences between the types of the different methods. Although this may be useful for the comparison of the genetic backgrounds of the particular set of isolates under study, it does not allow for a broader view of how the results of the different typing methods are related.

As more bacterial genomes are completed, novel typing methods will appear based on the new information available. Comparisons of these new methods to those currently available should be complemented by a quantitative measure of how much information is gained from a new method in terms of discriminatory power, type assignment, or even phylogenetic information about the isolates. It is conceivable that less-sophisticated molecular methods can recover levels of information about the relationships between the isolates that are similar to those obtained with newer sequence-based methods. Since typing schemes analyze different phenotypic or genotypic properties of bacteria, if some congruence between the methods is found, it suggests that a phylogenetic signal is being recovered by both methods, allowing greater confidence about the evolutionary hypothesis or clonal dispersion of the strains under study. These quantitative comparisons should allow the informed choice of which typing method is more appropriate in a given clinical or microbiological research setting, also taking into account other factors, such as the ability to identify isolates of interest, execution time, cost-effectiveness, or ease of

* Corresponding author. Mailing address: Rua da Quinta Grande 6 2780-156 Oeiras, Portugal. Phone: 351 21 446 98 55. Fax: 351 21 442 87 66. E-mail: jcarrico@itqb.unl.pt.

† Supplemental material for this article may be found at http://jcm.asm.org/.

interpretation of the results. A great diversity of typing methods is used to characterize bacterial isolates, rendering the comparison of the various studies difficult. If one could infer the missing information from the available data provided in each study, one could overcome this problem. In order to do this, a method offering a quantitative assessment of the confidence of predicting an unknown character from another typing method or set of methods is needed.

Streptococcus pyogenes or group A streptococci (GAS) are known to cause infections ranging from mild manifestations, such as pharyngitis, to severe invasive infections, such as streptococcal toxic shock syndrome and necrotizing fasciitis (7). These human pathogens provide a good case study for mapping relationships between typing schemes, since multiple typing methods have been used in their characterization, including T and M serotyping, antibiotic resistance typing, PFGE, restriction fragment length polymorphism (*vir* typing), *emm* sequence typing, and MLST (9, 17, 21). Although all these methods have proven useful for the characterization of GAS isolates, phenotypic methods have declined in popularity and the mainstream methods are now *emm* sequence typing, PFGE, and MLST. In the study that defined the MLST scheme for GAS, the authors compared MLST and *emm* sequence typing and concluded that the majority of *emm* types define clones or clonal complexes (9). This conclusion, together with the existence of extensive data on serological M types that are directly comparable to *emm* sequence type data and a technically simpler and more economic determination of *emm* types as opposed to the characterization by MLST, led to the frequent use of *emm* typing as the main typing technique for GAS clone definition. Notwithstanding the advantages of sequence-based methods, GAS virulence has been related to the presence of phages and to horizontal transfer of large fragments of DNA (1, 32). These observations suggest that techniques, like PFGE, that probe genomic organization could be more discriminative than sequence-based methods, since phage insertions can alter band positions in an agarose gel and, consequently, create more diversity within PFGE types.

In this paper we propose the use of measures of clustering concordance—Adjusted Rand (15, 24) and Wallace (34) coefficients—to compare type assignments, allowing a quantitative approach for exploring the concordance between typing methods. The proposed methods were applied to a set of 325 macrolide-resistant GAS for which extensive typing information was available and, when possible, we generalized the conclusions based on this data set by using typing data available from the MLST database. The proposed framework also allows the evaluation of possible gains in discriminatory power obtained by using different methods or any combination of typing schemes and the identification of which of the typing methodologies used will be more informative in clone definition. Ultimately, this framework may allow a mapping of type equivalences between typing methods.

MATERIALS AND METHODS

Strain collection. A collection of 325 macrolide-resistant *S. pyogenes* isolates recovered from throat swabs associated with a diagnosis of tonsillopharyngitis, from the period between 1998 and 2003 in Portugal, was analyzed. Results of antimicrobial susceptibility testing, T typing, macrolide-resistant phenotyping and genotyping, and *emm* typing were reported previously (29). Eleven T types were identified (1, 2, 4, 5/27/44, 6, 9, 12, 13, 25, 28, and B3264). Since twelve isolates were nontypeable by this method, the typeability of this method in our collection was 97%. All isolates were analyzed by PFGE using SfiI and either SmaI or Cfr9I endonucleases. Twelve *emm* sequence types were identified (1, 2, 4, 6, 9, 11, 12, 22, 28, 75, 77, and 89). Forty-one strains were chosen for MLST analysis by selecting at least one isolate from each SmaI/Cfr9I cluster. Ten sequence types (ST) were found (20, 28, 36, 38, 39, 45, 46, 52, 75, and 150) (28).

Gel analysis. A database of PFGE patterns was created in Bionumerics version 4.5 from Applied Maths (Sint-Martens-Latem, Belgium).

The gel digital photos acquired and stored in a Kodak EDAS 290 system were imported into a Bionumerics database, as inverted 8-bit grayscale TIF images.

For each image, spectral analysis included in the software was used, to determine the disk size that should be used in "rolling disk" background subtraction (Background scale) and the cutoff threshold for least-squares filtering (Wiener cutoff scale). Furthermore, a median filter was used in the image to further smooth the densitometric curves.

After this image preprocessing, intergel and intragel normalizations of PFGE runs were done using a Lambda PFGE molecular marker (New England Biolabs, Ipswich, Ma.). All the gels had three markers in the first, middle, and last lanes. Ten lambda bands were used from 48.5 kb to 485 kb.

On all gel images, band assignment was manually curated after automatic band detection. Bands ranging from 22.8 kb to 608 kb were considered in this study.

The settings used for comparing the strains' PFGE patterns were 1.0% optimization and 1.5% band tolerance.

Diversity indices. Hunter and Gaston (16) proposed the use of Simpson's index of diversity (30) to measure the discriminatory ability of typing systems. This index indicates the probability of two strains sampled randomly from a population belonging to two different types. Grundmann et al. (13) proposed a method for determining confidence intervals (CIs) of Simpson's index, thereby improving the objective assessment of the discriminatory power of typing techniques. The formulas of Simpson's Index (D) and the CI are presented in the following equations:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} n_j(n_j - 1)$$

$$\sigma^2 = \frac{4}{N}\left[\sum \pi_j^3 - \left(\sum \pi_j^2 \right)^2 \right]$$

$$\mathrm{CI} = [D - 2\sqrt{\sigma^2}, \mathrm{D} + 2\sqrt{\sigma^2}]$$

where N is the total number of strains in the sample population, S the total number of types described, n_j is the number of strains belonging to the jth type, and π_j is the frequency n_j/N.

Other diversity indices exist, such as the Shannon-Wiener index (27) and others from the Hill family of indices, of which Simpson's index is a special case (14). Notwithstanding, the ease of interpretation of Simpson's index of diversity as a probability and the possibility of calculating a confidence interval justifies the choice of this index in our study.

Clustering comparison coefficients: Rand, Adjusted Rand, and Wallace. In molecular epidemiology studies, the term cluster is frequently used to describe a group of isolates sharing similar characteristics according to a given typing method. Frequently, the clusters are obtained by hierarchical methods, such as the unweighted-pair group method with arithmetic means (UPGMA), providing further detail on the relationships of the isolates within clusters. In these cases, the definition of types relies on partitioning the resulting dendrogram at a given similarity value. In this paper, the terms partition, cluster, and type will be used interchangeably to identify a group of isolates sharing similar characteristics according to a given typing method.

To compare two sets of results of different microbial typing methods, an objective measure of agreement is needed. Several measures were developed for comparing two sets of partitions (10, 15, 22, 24, 34), taking different approaches to how partitions should be compared. For their ease of interpretation, in this study we use Adjusted Rand's index and Wallace coefficient. Rand (24) and Adjusted Rand (15) are symmetric coefficients, i.e., they do not take into consideration which partition is considered the standard, while others, like the ones proposed by Wallace (34), do. It is also important to note in this context that none of the partitions tested are considered the "correct" partition in terms of microbial typing.

Given two partition schemes of the same data set, P and P′, all these coefficients are calculated based on the fact that a pair of points (in microbial typing, a pair of points is a pair of isolates under study) from the data set will fall into

TABLE 1. Clustering comparison coefficients: Rand, Adjusted Rand, and Wallace

Coefficient	Formula (s)[a]
Rand	$\frac{a + d}{a + b + c + d}$
	$\frac{a + d - n_c}{a + b + c + d - n_c}$
Adjusted Rand	$n_c = \frac{n(n^2 + 1) - (n + 1)\sum n_i^2 - (n + 1)\sum n_j^2 + 2\sum\sum \frac{n_i^2 n_j^2}{n}}{2(n - 1)}$
Wallace	$W_I(P, P') = \frac{a}{a + b}$
	$W_{II}(P, P') = \frac{a}{a + c}$

[a] See the text for definitions of a, b, c, and d; n represents the number of strains under study, n_i and n_j represent the numbers of strains in types i and j of partitions P and P′, respectively.

one of the following conditions: a, the number of point pairs that are in the same cluster in P and P′; b, the number of point pairs that are in the same cluster in P but not in P′; c, the number of point pairs that are in the same cluster in P′ but not in P; or d, the number of point pairs that are in different clusters in P and P′.

The coefficients can then be defined as shown in Table 1.

Rand's index represents the proportion of agreement for both matches (a) and mismatches (d). An acknowledged limitation of this coefficient is that, when comparing two random partitions, the expected value of the Rand's index does not take a null value (indicating nonagreement). To address this issue, Hubert and Arabie (15) assumed a hypergeometric distribution as the random model, adding a correction factor designed to take into account the presence of chance agreement. The Hubert and Arabie's Adjusted Rand index, here referred to simply as Adjusted Rand, allows a better quantitative evaluation of the global congruence between the two partitions.

Wallace proposed two coefficients, based on Fowlkes and Mallows ' coefficient (10). They are easy to interpret since they represent the probability that a pair of points which are in the same cluster under P are also in the same cluster under P′ and vice versa.

Wallace's coefficients provide an estimate of, given a typing method, how much new information is obtained from another typing method. A high value of Wallace's coefficient indicates that partitions defined by a given method could have been predicted from the results of another method, suggesting that the use of both methodologies is redundant.

To facilitate the use of these indices in studies conducted by others, we have made available a Bionumerics script that calculates these indices from any two sets of data generated by different typing methods. The script can be downloaded from http://biomath.itqb.unl.pt/ClusterComp.

Visual representation of cluster congruence. To facilitate the interpretation and representation of the comparisons between partitions, we developed a visual method where all the clusters and cases under comparison are represented in a figure similar to a sequence dot plot (12, 19). The strains are ordered by type and cluster size for each of the typing methods under comparison. A dot is then plotted at the intersection of the position of each strain. Vertical and horizontal lines delimit the clusters in the figure.

Examples of this visual representation are shown in Fig. 2D and in Fig. SA1 in the supplemental material.

RESULTS

Pulsed-field gel electrophoresis. For this study, 325 isolates of *S. pyogenes* were analyzed by PFGE, using two restriction enzymes: SfiI and either one of two isoschizomers—SmaI or Cfr9I. The use of Cfr9I was necessary since the majority of the 150 isolates presenting macrolide

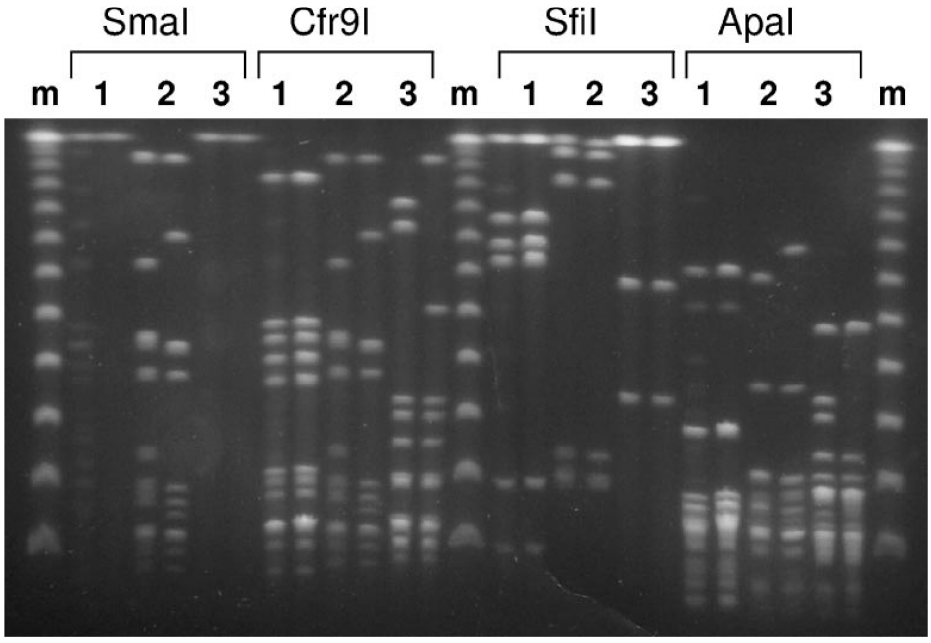


FIG. 1. PFGE profiles of three pairs of macrolide-resistant isolates generated after digestion with either SmaI, Cfr9I, SfiI, or ApaI. The isolates in pair 1 were refractory to cleavage with SmaI similarly to what was previously reported in the literature for isolates presenting the M phenotype (21) and showed identical profiles upon digestion with Cfr9I, SfiI, and ApaI. The isolates in pair 2 were digested by both SmaI and Cfr9I and, as expected, the profile of each isolate was identical with either endonuclease. The isolates in pair 2 showed different SmaI/Cfr9I and ApaI profiles but were indistinguishable by SfiI. The isolates in pair 3 were refractory to cleavage with SmaI and presented different profiles upon digestion with Cfr9I (four-band difference). These isolates presented identical SfiI profiles but exhibited different profiles with ApaI. m, lambda ladder PFG marker (New England Biolabs, Beverly, Ma.).

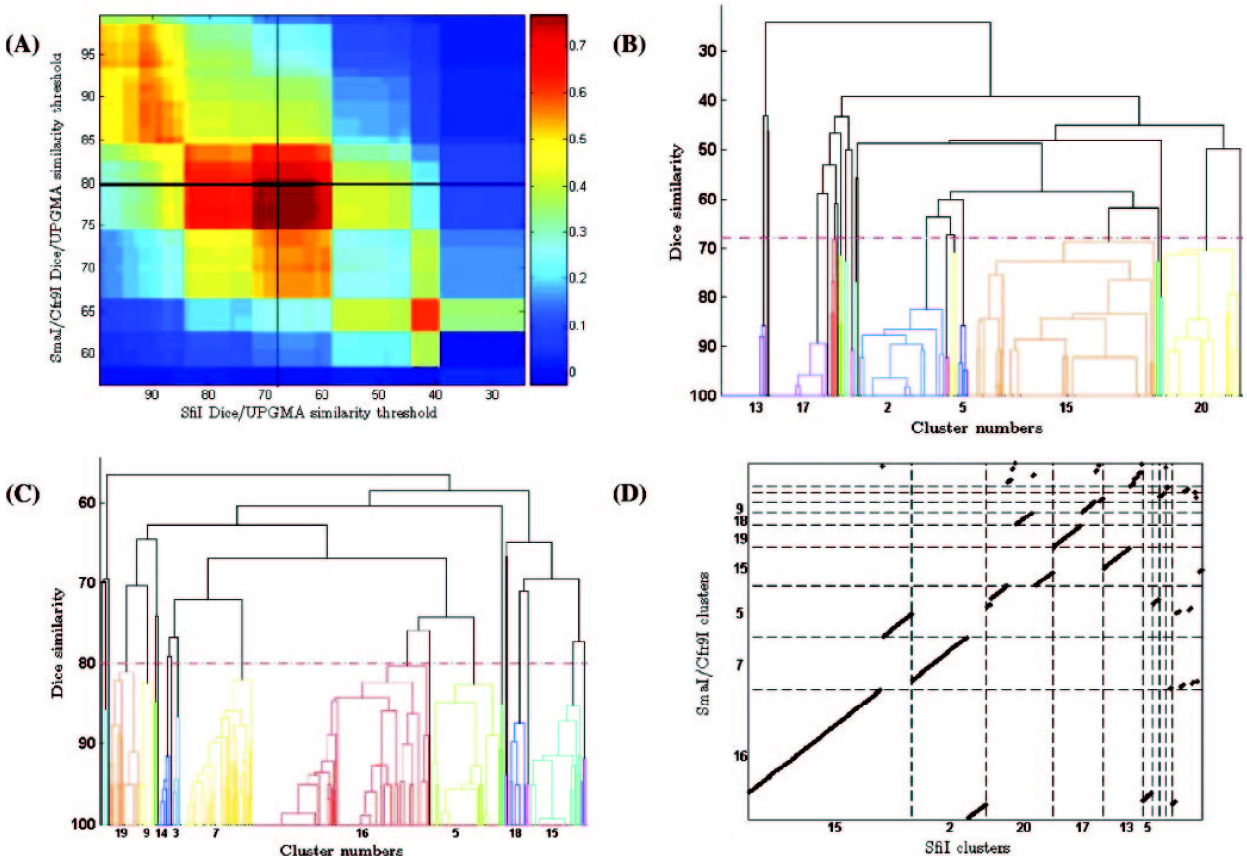


FIG. 2. Panel A. Adjusted Rand values for all possible cutoff values in each of the SmaI/Cfr9I and SfiI dendrograms (in panels B and C). Panel B. Dice/UPGMA dendrogram for SfiI band patterns of the 325 isolates. Panel C. Dice/UPGMA dendrogram for SmaI/Cfr9I band patterns of the 325 isolates. Panel D. Visual representation of cluster congruence between SmaI/Cfr9I and SfiI clusters at the cutoff levels indicated in panels B and C.

resistance phenotype M were refractory to cleavage with SmaI, in agreement with previous studies (21). As expected, DNA of isolates that were digested with SmaI presented identical band patterns when digested with Cfr9I (Fig. 1). In this way the use of Cfr9I increased the typeability of PFGE to 100%, compared to 54% (175/325) estimated for the use of SmaI alone, in our collection.

In previous studies, SfiI is the endonuclease most frequently used to characterize M isolates by PFGE (21), although a few studies use ApaI (3). There was a concern that SfiI patterns did not allow sufficient discrimination, since these patterns had only 3 to 10 bands, compared to SmaI/Cfr9I patterns that presented 8 to 17 bands (Fig. 1). This hypothesis was tested using Simpson's index of diversity and Adjusted Rand to determine the threshold that better defines types compared to those defined by SmaI/Cfr9I.

The Adjusted Rand index was calculated for each possible combination of partitions given by varying the threshold cutoffs for each UPGMA/Dice dendrogram using SfiI and SmaI/Cfr9I endonucleases similarly to what was previously described (5). The threshold value that produced the maximum coefficient value was determined, and the results are displayed in Fig. 2.

The maximum value of Adjusted Rand of 0.771 was found at a threshold level of 77% similarity in the Dice/UPGMA dendrogram of SmaI/Cfr9I and at a 68% similarity in the SfiI dendrogram.

An 80% similarity value was previously shown to be useful and concordant with proposed visual comparison criteria when defining types by Dice/UPGMA dendrograms of SmaI profiles of *Streptococcus pneumoniae* (5, 11, 26, 33). At this commonly used similarity threshold cutoff, a maximum Adjusted Rand value of 0.765 was found, corresponding to a threshold value of 68% in the SfiI dendrogram. Due to its acceptance as the cutoff value to define clusters in SmaI Dice/UPGMA dendrograms and the small difference observed in the value of Adjusted Rand at these two threshold levels, we opted to use the 80% cutoff for SmaI/Cfr9I and the 68% cutoff for SfiI in the remainder of the analysis.

At these cutoff values, 21 clusters were defined by either SmaI/Cfr9I or SfiI. Simpson's index of diversity calculations for the partitions found at these threshold levels for either endonuclease were of the same value of 0.81 (95% CI, 0.78–0.84), resulting in equal discriminatory power for SmaI/Cfr9I and SfiI (Table 2).

TABLE 2. Number of types and Simpson's index of diversity of each typing method

Typing method	No. of types found	Simpson's index of diversity (95% CI)	No. of isolates typed (% typeability)
T typing	12	0.72 (0.68–0.77)	316 (97)[a]
emm typing	12	0.77 (0.74–0.81)	325 (100)
T + *emm* typing	33	0.81 (0.77–0.84)	325 (100)
SmaI/Cfr9I 80%	21	0.81 (0.78–0.84)	325 (100)
MRP[b]	3[c]	0.51 (0.50–0.53)	325 (100)
SfiI 68%	21	0.81 (0.78–0.84)	325 (100)
MLST	10	—[d]	41 (100)

[a] Twelve strains were nontypeable. For the purpose of calculating the Simpson's index of diversity, they were all considered to belong to the same group.
[b] MRP, macrolide resistance phenotyping.
[c] Three macrolide resistance phenotypes were considered: M, $cMLS_B$, and $iMLS_B$.
[d] —, Simpson's index of diversity was not calculated for the 41 isolates since it would not be comparable to the remaining results calculated for the entire 325-isolate data set.

Comparing typing methods. Simpson's index of diversity provides a measure of the discriminatory power of the different typing methods as applied to our study data. Table 2 summarizes this coefficient for the methods used: T typing, *emm* typing, a combination of both these methods, PFGE typing defined at an 80% cutoff value on the SmaI/Cfr9I Dice/UPGMA dendrogram, macrolide resistance phenotyping, PFGE typing defined at a 68% cutoff value on the SfiI Dice/UPGMA dendrogram, and MLST data.

For comparing the congruence between type assignments of the different typing methods, adjusted Rand and Wallace coefficients were calculated for the subset of 41 isolates for which the results of all typing methods were available. These isolates are a diverse collection representing most of the types defined by the various methods in the entire collection of 325 isolates. The results are shown in Fig. 3A and in Tables SA1 and SA2 in the supplemental material. The data indicate a strong correlation between the information provided by SmaI/Cfr9I PFGE, MLST, and *emm* typing. It is interesting to note that there was a robust bidirectional correspondence between SmaI/Cfr9I PFGE types and ST. In contrast, a strong correspondence was found only in the direction of *emm* type for both ST and SmaI/Cfr9I PFGE types but not in the reverse direction (Fig. 3A).

To increase the robustness of the values of the Adjusted Rand and Wallace coefficients for the correspondences between the various typing methods utilized to characterize GAS, we excluded the MLST data and used the entire collection of 325 isolates for which information regarding all other typing methods was available. The results are shown in Tables 3 and 4. The values of Rand and Wallace coefficients were consistently higher when considering the 325 isolates. This was only expected if the smaller data set of 41 isolates already reflected the true correspondences between the various typing methods.

Previous publications indicated that *emm* types defined clones as assessed by MLST (9). Our limited data suggested a more complex relationship between *emm* typing and MLST (Fig. 4), resembling the comparison between *emm* and SmaI/Cfr9I PFGE types for our entire data set (see Fig. SA1 in the supplemental material).

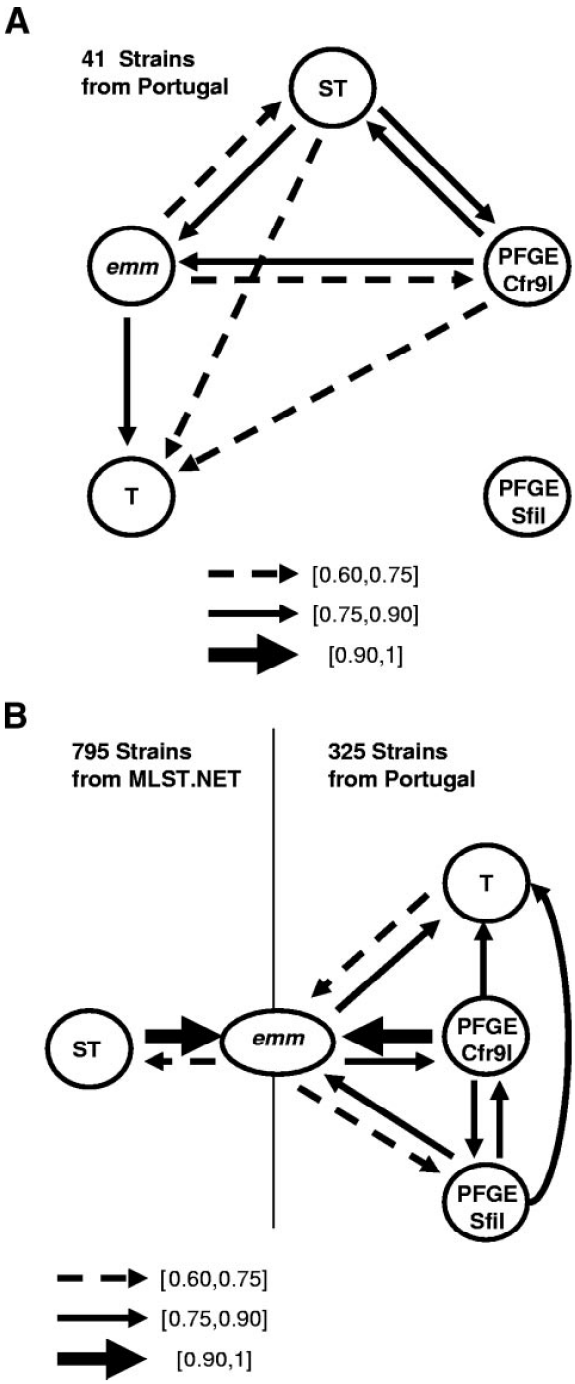


FIG. 3. (A) Representation of correspondences between the typing methods used in the 41-strain subset, calculated by Wallace coefficients. The arrows represent Wallace coefficients of >0.60.(B) Representation of correspondences between the typing methods used in the 325-strain data set and on the 795 strains from www.mlst.net, calculated by Wallace coefficients. The arrows represent Wallace coefficients of >0.60.

TABLE 3. Adjusted Rand coefficients for the methods used to characterize the 325 macrolide-resistant GAS

Typing method	Typing method					
	T typing	*emm* typing	T + *emm* typing	MRP[a]	SmaI/Cfr9I 80%	SfiI 68%
T typing	1.000					
emm typing	0.693	1.000				
T + *emm* typing	0.768	0.905	1.000			
MRP[a]	0.268	0.462	0.397	1.000		
SmaI/Cfr9I 80%	0.566	0.837	0.764	0.399	1.000	
SfiI 68%	0.514	0.724	0.663	0.382	0.762	1.000

[a] MRP, macrolide resistance phenotyping.

TABLE 4. Wallace coefficients for the methods used to characterize the 325 macrolide-resistant GAS

Typing method	Typing method					
	T typing	*emm* typing	T + *emm* typing	MRP[a]	SmaI/Cfr9I 80%	SfiI 68%
T typing	1.000	0.697	0.697	0.723	0.559	0.528
emm typing	0.861	1.000	0.861	0.989	0.803	0.724
T + *emm* typing	1.000	1.000	1.000	0.988	0.802	0.725
MRP[a]	0.415	0.459	0.395	1.000	0.392	0.386
SmaI/Cfr9I 80%	0.818	0.952	0.817	1.000	1.000	0.812
SfiI 68%	0.764	0.848	0.731	0.972	0.802	1.000

[a] MRP, macrolide resistance phenotyping.

To better evaluate the discriminatory power of MLST, we calculated Simpson's index of diversity for the 795 strains that had unambiguous information about both ST and *emm* type in the *S. pyogenes* MLST database (http://spyogenes.mlst.net; with a total of 847 isolates on 29 August 2005). Likewise, the same index was calculated for *emm* type to provide a more global view than that afforded by our data set. The results are shown in Table 5. The same could not be done for T types, since only 90 strains had T type information; 23 of those were nontypeable, and the majority of the 67 remaining strains had ambiguous information (two types). The distribution in this data set of *emm* sequence types per ST, and vice versa, as well as the overall concordance between *emm* typing and MLST among the strains in the *S. pyogenes* MSLT database can be found in the supplemental material (Fig. SA2 and SA3).

The Adjusted Rand for the comparison of the clustering by MLST and *emm* typing is 0.77, indicating a good overall match between partitions. The Wallace coefficient provides more information: considering ST as the standard for comparison, the value of Wallace's index is 0.952, i.e., the probability of two strains having the same ST also sharing the same *emm* type is 95%. However, the probability of two isolates that share the same *emm* type sharing the same ST is only 66% (Wallace's index is 0.655) (Table 5).

The correspondences between the various typing methods defined by using these expanded data sets are graphically represented in Fig. 3B.

DISCUSSION

The primary objective of this report was to provide a framework for the quantitative assessment of correspondence between type assignments obtained by different microbial typing methods. This quantification is achieved by the use of Simpson's index of diversity, Hubert and Arabie's Adjusted Rand, and the Wallace coefficient and is complemented by the visualization of the congruence between partitions generated by different typing methods.

An important application of the proposed framework is in evaluating if clusters generated by a given typing method could have been predicted by another methodology, allowing the evaluation of the usefulness of using several typing methods to characterize the same collection of isolates. This is also important in benchmarking the novel information offered by new typing schemes and in establishing if one can infer unknown typing information for a given isolate from other known characters.

While Simpson's index was previously used in this context (16) and it allows for a measure of the discriminatory power of a typing method, it does not evaluate the degree of equivalence between type assignments of two distinct typing methods. This goal is achieved here by the Adjusted Rand index, which provides an overall measure of the congruence between two typing methods. On the other hand, Wallace's coefficient is more informative and offers a clear interpretation since it represents the probability that a pair of strains which are assigned to the same type by one method are also classified in the same type by the other method. To facilitate the use of these indices in studies conducted by others, we have made available a Bionumerics script that calculates these indices from any two sets of partitions generated by different typing methods. To test the validity of the framework proposed, we applied it to a collection of macrolide-resistant GAS characterized by T serotyping, *emm* sequence typing, PFGE using two different endonucleases, and MLST.

Although intuitively we could expect that the PFGE band patterns generated after SfiI digestion would be less discriminatory than those of SmaI/Cfr9I, since the former presented fewer bands, this was not the case. The clusters defined by either SmaI/Cfr9I or SfiI, at the cutoff levels showing the highest agreement, showed the same Simpson's index value, namely, 0.81 (95% CI, 0.78 to 0.84).

In spite of the similar discriminatory powers, SmaI/Cfr9I and SfiI assigned a significant number of isolates to different types (Adjusted Rand, 0.765). The value of the Wallace index was almost the same in either direction and indicated that as many as one out of every five pairs of isolates classified in the same cluster by an endonuclease are in separate clusters using the other endonuclease. This is represented in Fig. 2D, where it can be seen that isolates that belong to a single cluster when using one endonuclease were scattered into at least two other clusters when using the other.

In view of this data, which endonuclease is more suitable for typing macrolide-resistant *S. pyogenes*? A weaker correspondence between SfiI PFGE and the other typing methods is shown in Fig. 3A, where no line connects SfiI PFGE and T types whereas multiple correspondences are associated with SmaI/Cfr9I PFGE types. When using the full data set of 325 isolates this was not so pronounced, but the correspondences established between SmaI/Cfr9I PFGE and the other typing methods were consistently stronger than those observed for

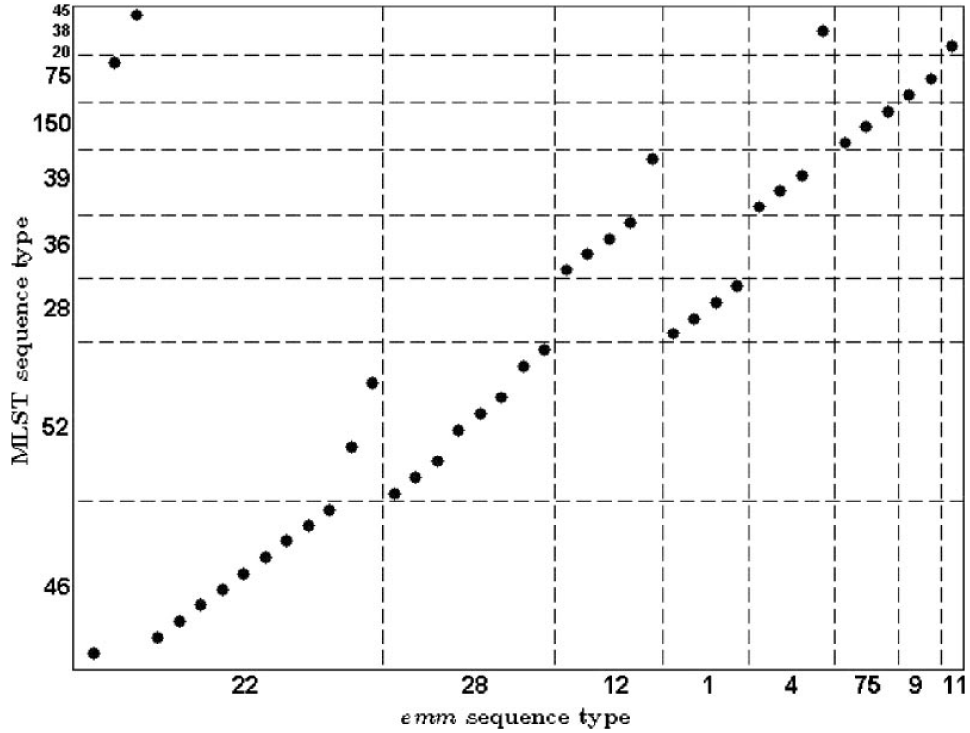


FIG. 4. Relationship between MLST and *emm* sequence typing for the 41 strains. In this example, the figure illustrates the fact that the majority of strains with *emm* type 22, the most frequent *emm* type in this subset, presented ST 46, and two isolates presented ST 52. However, the majority of the strains with ST 52 have *emm* sequence type 28.

SfiI PFGE (Fig. 3B). We would therefore argue that the groups defined by SmaI/Cfr9I, by being more concordant with other typing methods, are more likely to accurately represent the true relationships between the isolates, justifying the use of these endonucleases in typing GAS isolates by PFGE. The use of the SmaI isoschizomer Cfr9I was essential in increasing the typeability of macrolide-resistant GAS, since the majority of the isolates presenting the M phenotype were refractory to cleavage by SmaI but were susceptible to digestion with Cfr9I (Fig. 1). A combination of the two endonucleases is therefore recommended for the characterization of macrolide-resistant GAS.

If two isolates belong to the same cluster according to SmaI/Cfr9I, one can predict confidently that they will also share the same *emm* sequence type (Wallace, 0.952). The reverse correspondence, on the other hand, cannot be made with the same certainty since only four out of five pairs of strains sharing the same *emm* sequence type were also grouped into the same SmaI/Cfr9I cluster (Wallace, 0.803). Even considering that SmaI/Cfr9I PFGE was comparable to *emm* sequence typing in terms of discriminatory power (Table 2), it was more informative given its predictive power over *emm* sequence typing (Fig. 3B).

TABLE 5. Diversity indices and Wallace coefficients for *emm* typing and MLST data available from the online MLST database

Typing method	No. of types found	Simpson's index of diversity (95% CI)	Wallace coefficient	
			MLST	*emm* typing
MLST	274	0.985 (0.982–0.988)	1.000	0.952
emm typing	155	0.978 (0.975–0.981)	0.655	1.000

T typing alone, although presenting a Simpson's index of diversity similar to that with *emm* sequence typing, separated the isolates into groups that could not be mapped confidently in any of the other typing methods, indicating that this methodology is poorly congruent with any of the others (Fig. 3). The combination of T typing and *emm* sequence typing showed a discriminatory power comparable to that of SmaI/Cfr9I (Table 2), but it did not improve the mapping relative to *emm* typing alone as demonstrated by Adjusted Rand and Wallace coefficients (Table 3 and Table 4). A drawback of using the combination of these two methods was the reduced typeability of T typing compared with all other typing methods and the necessity of maintaining a set of specific sera for T typing.

When comparing the ST of the 795 isolates referenced in the MLST database, what had started to emerge with our limited data set of 41 isolates became clear: the majority of *emm* types subdivide themselves into two or more ST (see Fig. SA3 in the supplemental material). This is supported by a Wallace coefficient of 0.655 when comparing *emm* types using MLST as a standard. In the reverse comparison, a Wallace coefficient of 0.952 implies that when two strains have the same ST we will only misclassify them according to their *emm* types once out of every 20 pairs. These results show that clone definition by MLST is more consistent than by *emm* sequence typing.

The correspondences between the various typing methods illustrated in Fig. 3B argue that performing either PFGE using SmaI and Cfr9I endonucleases or MLST is sufficient to predict the *emm* type of the isolates with less than 5% error but that one cannot accurately predict ST or SmaI/Cfr9I PFGE types from *emm* data.

A comprehensive comparison between SmaI/Cfr9I PFGE and MLST is outside the scope of this paper. The limited data available from the smaller data set of 41 isolates for which we had MLST information suggested that there is a strong mapping between SmaI/Cfr9I PFGE types and MLST (Fig. 3A; also see Tables SA1 and SA2 in the supplemental material). This was also supported by the similar relationship of each of these methods with *emm* typing (Fig. 3B). In spite of the role of bacteriophages and of the horizontal exchange of large fragments of genomic DNA in the evolution of virulent GAS strains (1, 32), these observations argued in favor of equally good results when using SmaI/Cfr9I PFGE or MLST to characterize GAS. However, the choice of isolates for which MLST was determined reflects the SmaI/Cfr9I PFGE type assignment, so further studies are necessary to clarify which of the two typing methods would provide a more discriminatory and informative clone definition.

Our data set represented a diverse group of GAS as documented by the use of the various typing methods; for instance, 21 SmaI/Cfr9I PFGE types were defined. However, it could be argued that these isolates do not accurately represent the global diversity of *S. pyogenes* since they are restricted to macrolide-resistant GAS recovered in Portugal during a limited time period. This would prevent the generalization of the results presented. Although it is certainly true that there was limited diversity in our collection, the expected clonal structure of such a geographically and temporally limited population would reinforce the correspondences between the different typing methods, increasing the values of adjusted Rand and Wallace coefficients. This was not observed for all methods and Wallace coefficients showed strong asymmetries depending on the directionality, suggesting that the results did not reflect a particular clonal composition of the studied population but are general properties of the typing methods used. This was further supported by the analysis of the more extensive data available from the MLST online database that strengthened the conclusions emerging from the study of our data set regarding the relationship between *emm* typing and MLST.

When using PFGE to characterize macrolide-resistant GAS, the results were in favor of the use of SmaI, complemented by its isoschizomer Cfr9I to circumvent the resistance to cleavage of M isolates as documented previously (21), against the alternative endonuclease SfiI. The analysis also highlighted the importance of using PFGE and MLST, in addition to *emm* typing, in the characterization of GAS due to the poor predictive value of the latter over the groups defined by the former.

As the data compiled in online databases (such as www.mlst.net) increases, the framework of methods presented will provide further insights into the relationships between isolates, eventually enabling a generic mapping between the different typing methods. The congruence of results between typing methods suggests that a phylogenetic signal is indeed being recovered by the typing data generated by different methods. Accordingly, the progressive identification of mapping functions, such as the probability matrices for agreement of type assignments represented in the supplemental material (Fig. SA4.1 and SA4.2), indicates that a consensus assessment of the relationships between the different types will soon be at hand. Such a tool would allow not only for comparisons of typing results obtained by different methods but would also facilitate the joint analyses of multiple typing methods.

ACKNOWLEDGMENTS

J. A. Carriço and F. R. Pinto were supported by grants SFRH/BD/3123/2000 and SFRH/BD/6488/2001, respectively, both from the Fundação para a Ciência e Tecnologia, Portugal.

Partial support for this work was provided by PREVIS (LSHM-CT-2003-503413 from the European Community awarded to J. S. Almeida, H. de Lencastre, and J. Melo-Cristino) and by a grant from the Fundação Calouste Gulbenkian awarded to J. Melo-Cristino and M. Ramirez.

This publication made use of the Multi Locus Sequence Typing website (http://www.mlst.net) at Imperial College London developed by David Aanensen and Man-Suen Chan and funded by the Wellcome Trust.

REFERENCES

1. **Aziz, R. K., R. A. Edwards, W. W. Taylor, D. E. Low, A. McGeer, and M. Kotb.** 2005. Mosaic prophages with horizontally acquired genes account for the emergence and diversification of the globally disseminated M1T1 clone of *Streptococcus pyogenes*. J. Bacteriol. **187:**3311–3318.
2. **Beall, B., R. R. Facklam, J. A. Elliott, A. R. Franklin, T. Hoenes, D. Jackson, L. Laclaire, T. Thompson, and R. Viswanathan.** 1998. Streptococcal emm types associated with T-agglutination types and the use of conserved *emm* gene restriction fragment patterns for subtyping group A streptococci. J. Med. Microbiol. **47:**893–898.
3. **Billal, D. S., M. Hotomi, K. Yamauchi, K. Fujihara, S. Tamura, K. Kuki, R. Sugita, M. Endou, J. Mukaigawa, and N. Yamanaka.** 2004. Macrolide-resistant genes of *Streptococcus pyogenes* isolated from the upper respiratory tract by polymerase chain reaction. J. Infect. Chemother. **10:**115–120.
4. **Brito, D. A., M. Ramirez, and H. de Lencastre.** 2003. Serotyping *Streptococcus pneumoniae* by multiplex PCR. J. Clin. Microbiol. **41:**2378–2384.
5. **Carriço, J. A., F. R. Pinto, C. Simas, S. Nunes, N. G. Sousa, N. Frazão, H. de Lencastre, and J. S. Almeida.** 2005. Assessment of band-based similarity coefficients for automatic type and subtype classification of microbial isolates analyzed by pulsed-field gel electrophoresis. J. Clin. Microbiol. **43:**5483–5490.
6. **Coenye, T., T. Spilker, A. Martin, and J. J. LiPuma.** 2002. Comparative assessment of genotyping methods for epidemiologic study of *Burkholderia cepacia* genomovar III. J. Clin. Microbiol. **40:**3300–3307.
7. **Cunningham, M. W.** 2000. Pathogenesis of group A streptococcal infections. Clin. Microbiol. Rev. **13:**470–511.
8. **de Lencastre, H., I. Couto, I. Santos, J. Melo-Cristino, A. Torres-Pereira, and A. Tomasz.** 1994. Methicillin-resistant *Staphylococcus aureus* disease in a Portuguese hospital: characterization of clonal types by a combination of DNA typing methods. Eur. J. Clin. Microbiol. Infect. Dis. **13:**64–73.
9. **Enright, M. C., B. G. Spratt, A. Kalia, J. H. Cross, and D. E. Bessen.** 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. Infect. Immun. **69:**2416–2427.
10. **Fowlkes, E. B., and C. L. Mallows.** 1983. A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78:**553–569.
11. **Gertz, R. E., Jr., M. C. McEllistrem, D. J. Boxrud, Z. Li, V. Sakota, T. A. Thompson, R. R. Facklam, J. M. Besser, L. H. Harrison, C. G. Whitney, and B. Beall.** 2003. Clonal distribution of invasive pneumococcal isolates from children and selected adults in the United States prior to 7-valent conjugate vaccine introduction. J. Clin. Microbiol. **41:**4194–4216.
12. **Gibbs, A. J., and G. A. McIntyre.** 1970. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur. J. Biochem. **16:**1–11.
13. **Grundmann, H., S. Hori, and G. Tanner.** 2001. Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. J. Clin. Microbiol. **39:**4190–4192.
14. **Hill, M.** 1973. Diversity and evenness: a unifying notation and its consequences. Ecology **54:**427–432.
15. **Hubert, L., and P. Arabie.** 1985. Comparing partitions. J. Classification **2:**193–218.
16. **Hunter, P. R., and M. A. Gaston.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. **26:**2465–2466.
17. **Kataja, J., P. Huovinen, A. Efstratiou, E. Perez-Trallero, and H. Seppäla.**

2002. Clonal relationships among isolates of erythromycin-resistant *Streptococcus pyogenes* of different geographical origin. Eur. J. Clin. Microbiol. Infect. Dis. **21:**589–595.

18. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 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.
19. **Maizel, J. V., Jr., and R. P. Lenk.** 1981. Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci. USA **78:**7665–7669.
20. **Malachowa, N., A. Sabat, M. Gniadkowski, J. Krzyszton-Russjan, J. Empel, J. Miedzobrodzki, K. Kosowska-Shick, P. C. Appelbaum, and W. Hryniewicz.** 2005. Comparison of multiple locus variable number tandem repeat analysis with pulsed-field gel electrophoresis, *spa* typing, and multilocus sequence typing for clonal characterization of *Staphylococcus aureus* isolates. J. Clin. Microbiol. **43:**3095–3100.
21. **Malhotra-Kumar, S., C. Lammens, S. Chapelle, M. Wijdooghe, J. Piessens, K. Van Herck, and H. Goossens.** 2005. Macrolide- and telithromycin-resistant *Streptococcus pyogenes*, Belgium, 1999–2003. Emerg. Infect. Dis. **11:**939–942.
22. **Milligan, G. W., and M. C. Cooper.** 1986. A study of comparability of external criteria for hierarchical cluster analysis. Multivariate Behav. Res. **21:**441–458.
23. **Nemoy, L. L., M. Kotetishvili, J. Tigno, A. Keefer-Norris, A. D. Harris, E. N. Perencevich, J. A. Johnson, D. Torpey, A. Sulakvelidze, J. G. Morris, Jr., and O. C. Stine.** 2005. Multilocus sequence typing versus pulsed-field gel electrophoresis for characterization of extended-spectrum beta-lactamase-producing *Escherichia coli* isolates. J. Clin. Microbiol. **43:**1776–1781.
24. **Rand, W. M.** 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66:**846–850.
25. **Schwartz, D. C., and C. R. Cantor.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:**67–75.
26. **Serrano, I., J. Melo-Cristino, J. A. Carriço, and M. Ramirez.** 2005. Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal. J. Clin. Microbiol. **43:**1706–1715.
27. **Shannon, C. E.** 1948. A mathematical theory of communication. Bell Syst. Tech. J. **27:**379–423, 623–656.
28. **Silva-Costa, C., M. Ramirez, and J. Melo-Cristino.** 2006. Identification of macrolide-resistant clones of *Streptococcus pyogenes* in Portugal. Clin. Microbiol. Infect. **12:**513–518.
29. **Silva-Costa, C., M. Ramirez, and J. Melo-Cristino.** 2005. Rapid inversion of the prevalences of macrolide resistance phenotypes paralleled by a diversification of T and *emm* types among *Streptococcus pyogenes* in Portugal. Antimicrob. Agents Chemother. **49:**2109–2111.
30. **Simpson, E. H.** 1949. Measurement of species diversity. Nature **163:**688.
31. **Struelens, M. J.** 1996. Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. Clin. Microbiol. Infect. **2:**2–11.
32. **Sumby, P., S. F. Porcella, A. G. Madrigal, K. D. Barbian, K. Virtaneva, S. M. Ricklefs, D. E. Sturdevant, M. R. Graham, J. Vuopio-Varkila, N. P. Hoe, and J. M. Musser.** 2005. Evolutionary origin and emergence of a highly successful clone of serotype M1 group A *Streptococcus* involved multiple horizontal gene transfer events. J. Infect. Dis. **192:**771–782.
33. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. **33:**2233–2239.
34. **Wallace, D. L.** 1983. A method for comparing two hierarchical clusterings: comment. J. Am. Stat. Assoc. **78:**569–576.

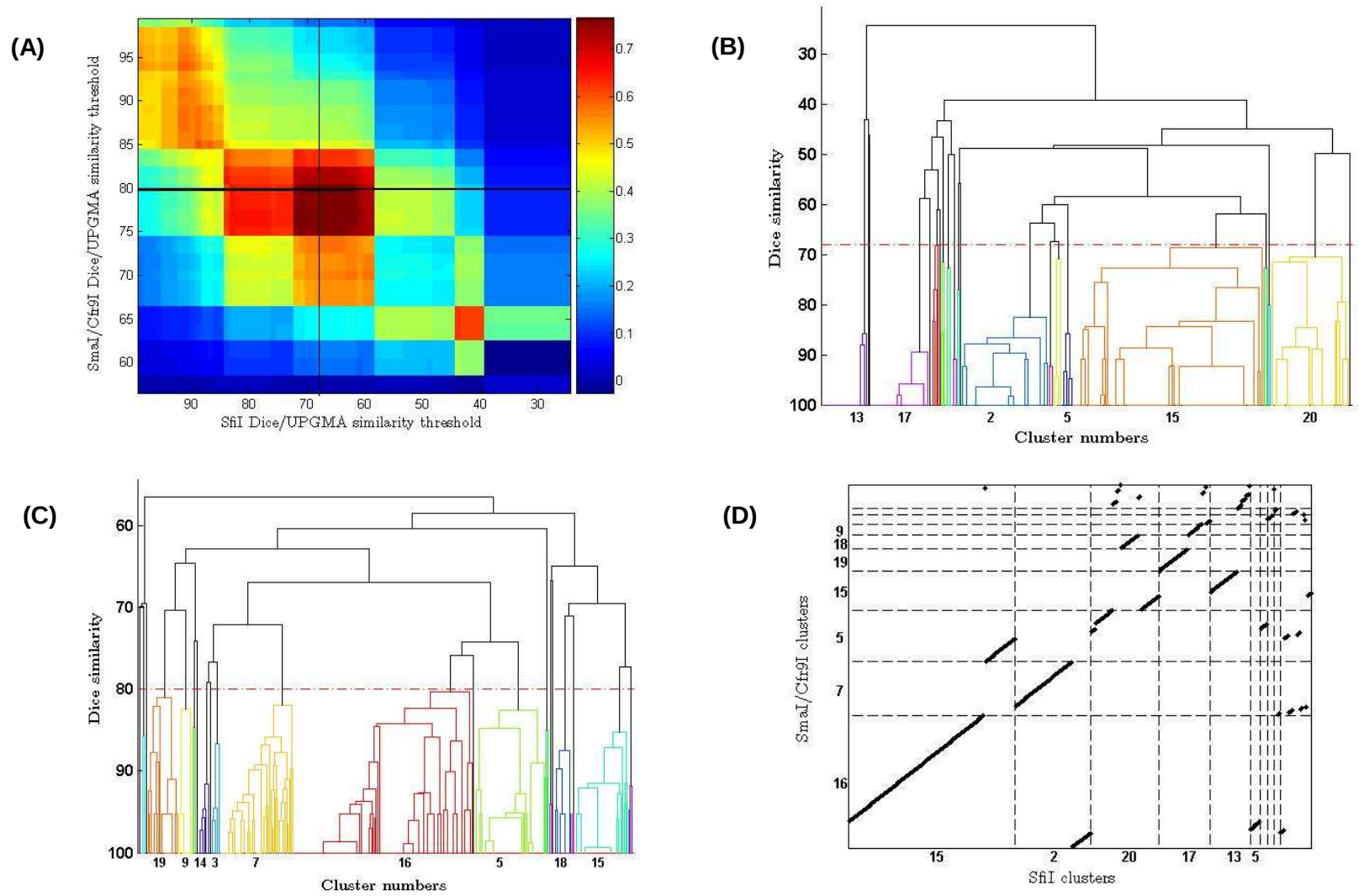


Figure 2- Panel A - Adjusted Rand values for all possible cut-off values in each of SmaI/Cfr9I and SfiI dendrograms (in panels B and C). Panel B – Dice/UPGMA dendrogram for SfiI band patterns of the 325 isolates. Panel C – Dice/UPGMA dendrogram for SmaI/Cfr9I band patterns of the 325 isolates. Panel D – Visual representation of cluster congruence between SmaI/Cfr9I and SfiI clusters at the cut-off levels indicated in panels B and C.

Supplemental Material

Table A1 - Adjusted Rand coefficients calculated for the 41 macrolide resistant GAS with MLST information

	T typing	*emm* typing	T+*emm* typing	MRP[a]	SmaI/Cfr9I 80%	SfiI 68%	MLST
T typing	***1.000***						
***emm* typing**	0.607	***1.000***					
T+*emm* typing	0.676	0.895	***1.000***				
MRP[a]	0.162	0.338	0.283	***1.000***			
SmaI/Cfr9I 80%	0.407	0.721	0.637	0.290	***1.000***		
SfiI 68%	0.287	0.354	0.318	0.188	0.396	***1.000***	
MLST	0.417	0.725	0.661	0.287	0.873	0.387	***1.000***

[a] MRP – Macrolide Resistance Phenotype

Table A2 - Wallace coefficient calculated for the 41 macrolide resistant GAS with MLST information

	T typing	*emm* typing	T+*emm* typing	MRP[a]	SmaI/Cfr9I 80%	SfiI 68%	MLST
T typing	**1.000**	0.577	0.577	0.660	0.407	0.304	0.412
***emm* typing**	0.836	**1.000**	0.836	**1.000**	0.709	0.388	0.709
T+*emm* typing	**1.000**	**1.000**	**1.000**	**1.000**	0.696	0.384	0.714
MRP[a]	0.318	0.333	0.279	**1.000**	0.286	0.216	0.284
SmaI/Cfr9I 80%	0.687	0.826	0.678	**1.000**	**1.000**	0.443	0.887
SfiI 68%	0.590	0.520	0.430	0.870	0.510	**1.000**	0.500
MLST	0.702	0.833	0.702	**1.000**	0.895	0.439	**1.000**

[a] MRP – Macrolide Resistance Phenotype

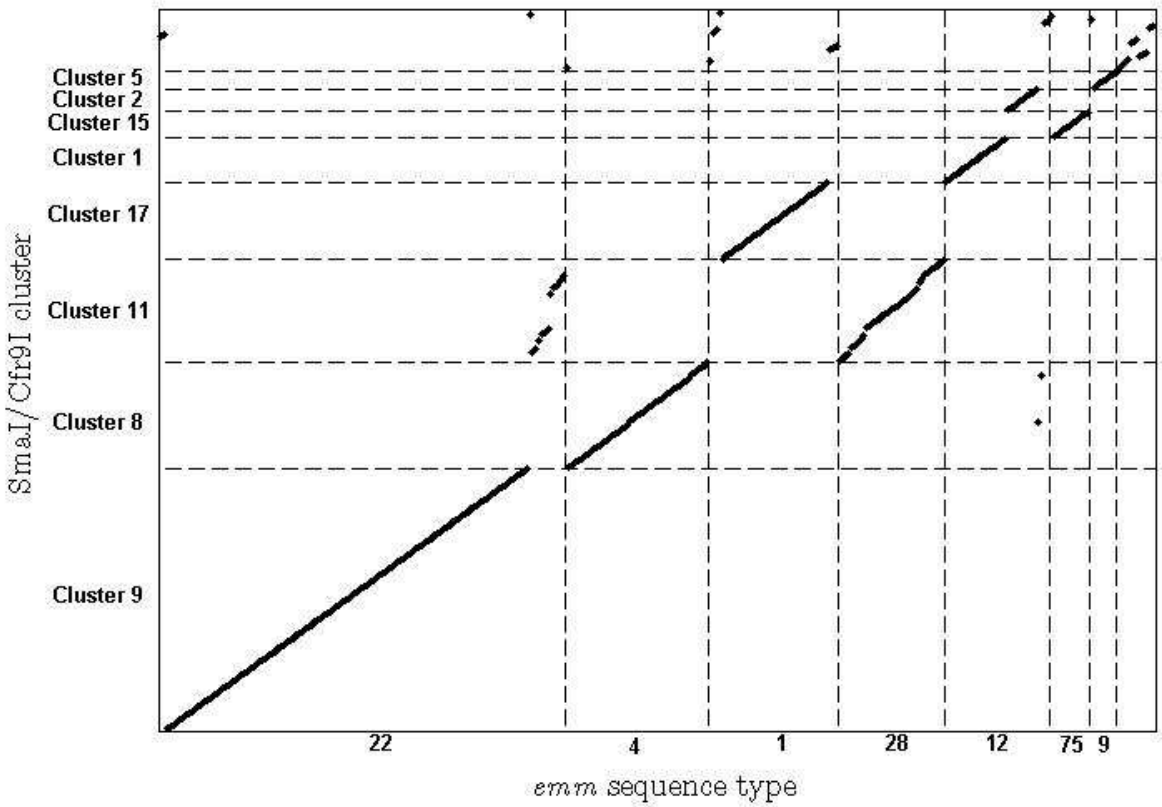


Figure A1 – Relation between PFGE SmaI/Cfr9I clusters (80% similarity cut-off DICE/UPGMA) and *emm* sequence types for the 325 strains.

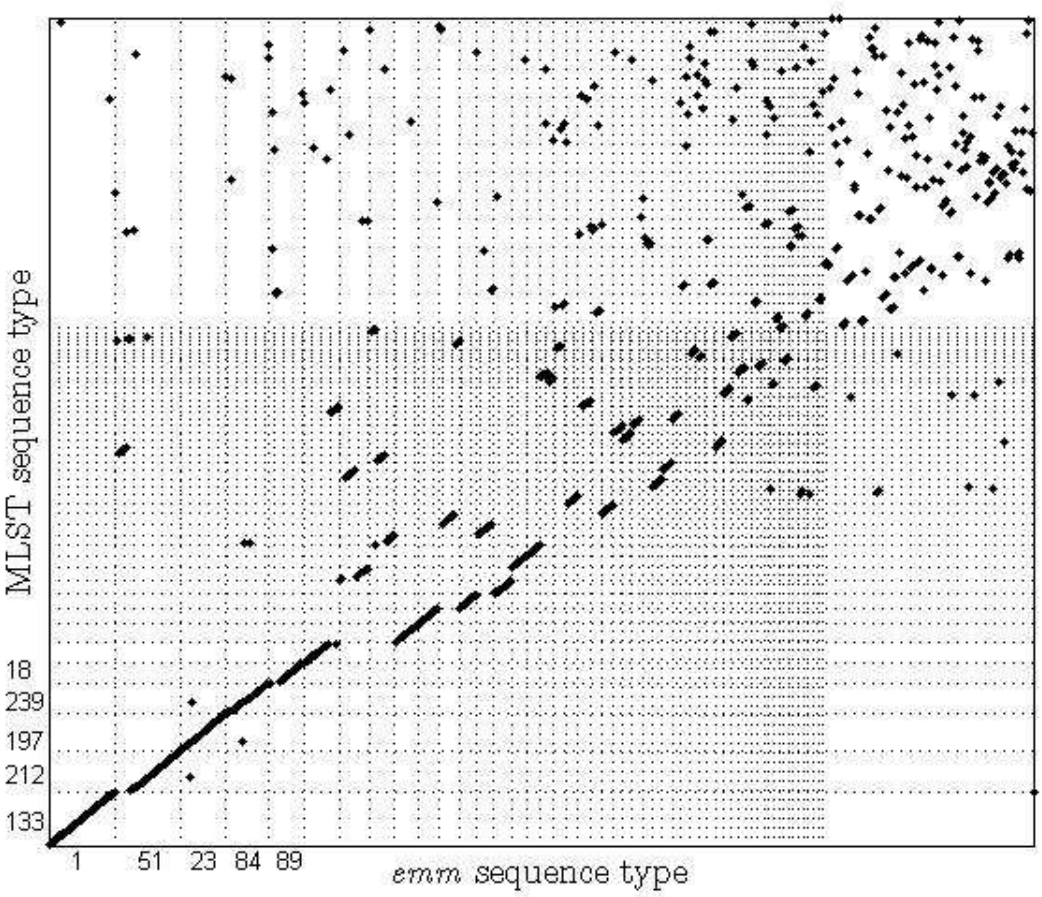
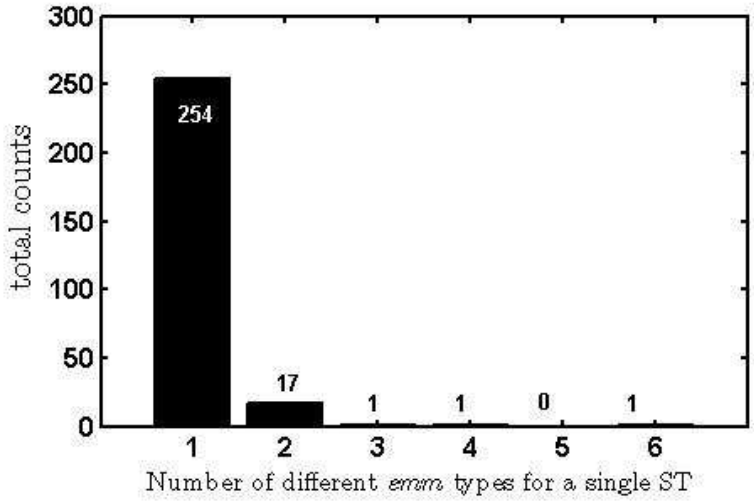


Figure A2 – Relationship between MLST sequence type and *emm* sequence type for the 795 strains that had unambiguous information about both ST and emm type in the S. pyogenes MLST database (http://spyogenes.mlst.net- presently with a total of 847 total strains). The horizontal and vertical lines limit the clusters with more than 4 elements. For each typing method, the five clusters with more elements have the corresponding type identified on the appropriate axis.

(1)



(2)

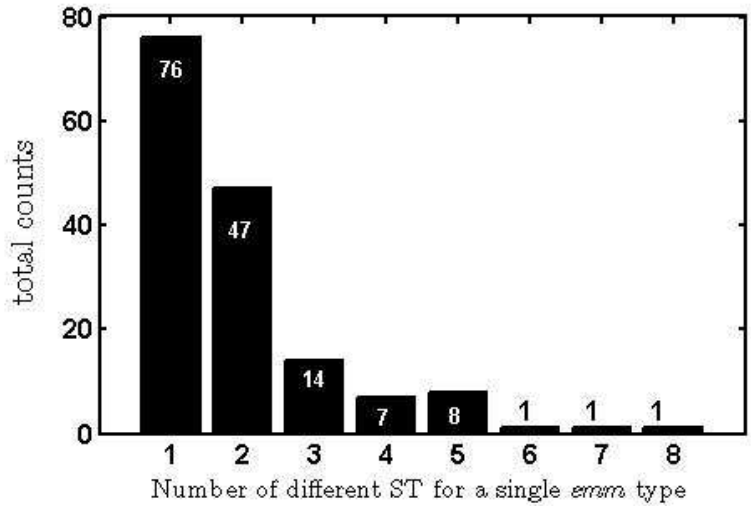
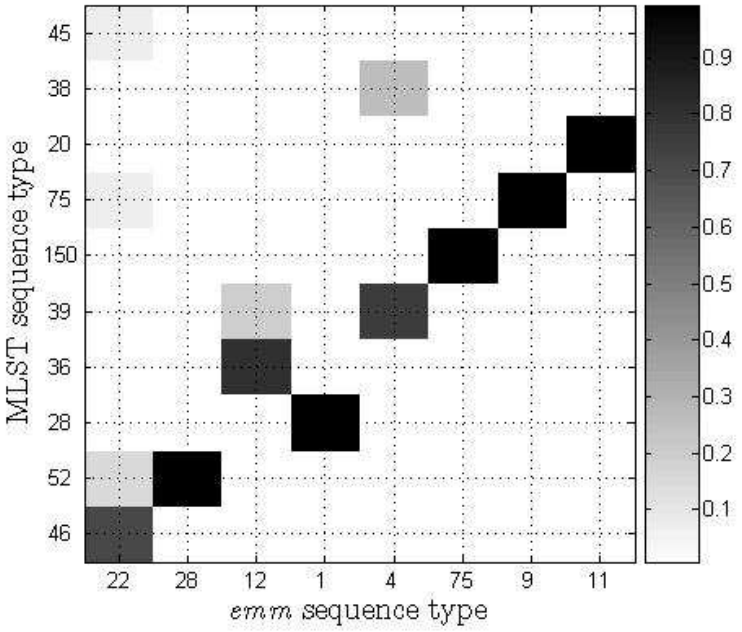


Figure A3 – Relations between emm types and ST for 795 isolates referenced in the online spyogenes.mslt.net database: (1) distribution of *emm* sequence type per MLST ST; (2) distribution of MLST ST per *emm* sequence type

(1)



(2)

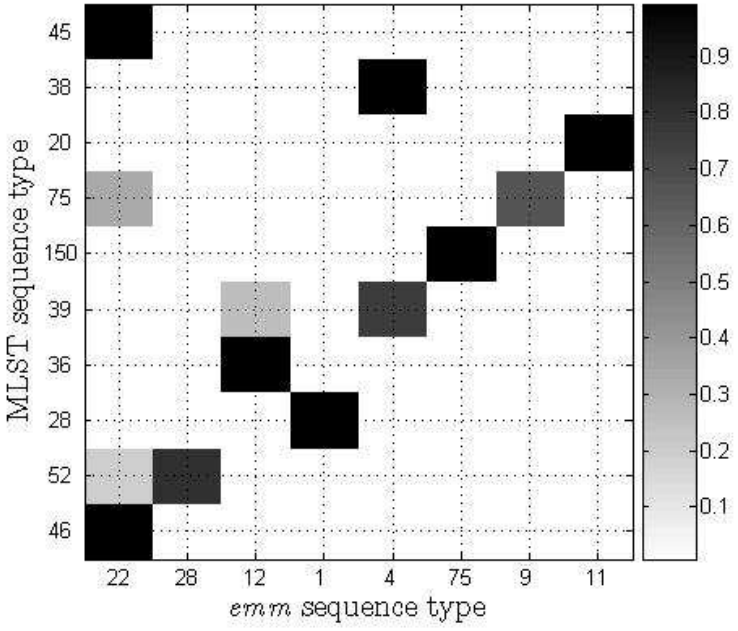


Figure A4 – Probability matrices for the concordance between *emm* sequence type and MLST ST assignments for 41 strains. The color scale represents the probability of the event in our data set of 41 strains (1) probability of a strain with a given emm sequence type having a ST type (2) probability of a strain with a given ST type having an emm sequence type.

Chapter VI

6.1. Final discussion

The technological advances in molecular biology and sequencing technology lead to the development of the new sequence based typing methods. However the ability to compared the obtained sequences to a central database and automatically assign a type based on the sequence was fundamental for these new microbial typing methodologies success. Behind such an apparently trivial procedure of submitting a sequence and, almost instantly, a type being assigned, a considerable computational power is needed to correctly store and compare the sequences, together with all other information that is available for the isolate. The comparisons algorithms that now take a matter of seconds or minutes to run, only 5 years ago have taken days or weeks. Therefore the advances in computation, algorithm development and internet technologies are also a major driving force in our capacity to analyze epidemiological data in a global manner, since we now can compare thousands of strains present in a database to our isolate of interest. Since this is a relatively new ability, methods of organizing this data and analyzing are needed to cope with these novel challenges.

In chapters II and III, we presented an online information system, where the flexibility of design allowed a multi-centric international study. The ultimate goal of this information system is the integration of data analysis and automated reporting to become a fully fledged Open Source Epidemiological Information System where a multi-centric approach is needed and data privacy of each centre can be enforced while allowing the pooling of data for integrative analyses. These Epidemiological Information Systems can be adapted for a multitude of studies and, since open source code is provided, algorithms can be easily developed, implemented and tuned for a multitude of studies from outbreak detection to population genetics studies. The data model for the EURISWEB database can also be used as the first step to construct an ontology of terms and concepts related to microbial typing and epidemiological data. This could guide the construction of new databases and facilitate the exchange of data between existing ones, which will prove he next challenge for this information technology. The ability to query multiple databases simultaneously pooling together information at different levels (phenotypic, genomic, proteomic or metabolic levels) from a single microorganism or even strain, could

provide us with new hypothesis and more refined details about the population structure of microbial pathogens and whose factors are important for the virulence and pathogenecity of some strains, providing us with new ways to circumvent the morbidity and mortality caused by infectious agents.

In Chapter IV, we propose the use of receiver operating characteristic curves, to assess the goodness of classification of several commonly used band-based similarity metrics for type and subtype classification of microbial isolates analyzed by Pulsed-Field Gel Electrophoresis (PFGE). The methodology used allowed the determination of a similarity threshold where a minimum misclassification error was observed when compared to the commonly used criteria for assigning types proposed by Tenover *et al (1)*. It also provides a way to fine tune gel analysis parameters such as band position tolerance. New similarity metrics (either band-based or correlation measures) can also be tested to a visually classified collection to determine if they provide better classification when compared to the ones commonly used and presented in the article. Another possible use of the methodology, instead of using known criteria for generic type/subtype classification, is choosing some a group of strains with certain characteristics (i.e. increased pathogenecity or virulence) and determine which set of parameters (band position tolerance and cut-off value for group assignment) will provide best classification to that group, fine tuning those parameters for the desired ratio between false positives and false negatives: if we want to be sure to classify an unknown strain to the group of selected strains we can allow for more false positives results, that can be latter excluded to belong to the group of interest, than false negatives; Conversely if we are searching a database for strains with particular characteristics to further study them using some expensive method, we can allow for more false negative results and fewer false positive ones, since analyzing a false positive result would involve an unnecessary expense. Finally this methodology can be extended to any typing method where a quantitative similarity level between isolates can be measured.
The image analysis procedures required by methods like PFGE, involve dedicated commercial software such as Bionumerics, and their implementation as an online tool is still difficult, given the bandwidth requirements for sending the gel images and the computational power that would be required to do image analysis for several concurrent users. But one of the unavoidable features of these commercial software packages is exporting the data in a multitude of formats that can be submitted to the online database and used for further studies. For instance the similarity matrixes for all isolates obtained in gel analysis software can be uploaded to the database and

algorithms such as the ones described in Chapter IV can be implemented and their results made available to all the users. Also the raw image can be displayed on the database, as described in Chapter III, allowing a visual confirmation of results.

In Chapter V, we describe a framework for relating multiple typing methods results and illustrated its use using a collection of macrolide-resistant *Streptococcus pyogenes* characterized by PFGE, T typing, *emm* typing and MLST. The aim of the framework was to provide a quantitative answer to the question if a typing method results could have been predicted by the results of another method. This is particularly useful to determine if new typing methods are providing new information or are redundant when compared with other typing methods or combination of typing methods.

The application of this framework to large collections of isolates characterized by different typing methods can also map type equivalences for different methods, allowing the inference of results for a given method when results of other methods are known. For that the indexes used could be expanded for allowing the simultaneous evaluation of several methods. This methodology could also be further adapted to provide a measure of strength of the phylogenetic signal that is recovered by different typing methods: If a group of isolates always have similar type assignments in different typing methods that can be interpreted as evidence of a clonal relationship between isolates. Since the proposed measures can be applied to any microorganism and typing method, another way to measure clonality in different bacterial species could be how this mapping of typing results equivalences evolves over time. Newly found type equivalences or diversification of types could indicate exchanges of genetic material or phenotypic adaptations that could be further investigated.

Another very useful extension of these measures could be the comparison of different data analysis methods for the same typing method results. Recently, we proposed the use of an information-theoretic similarity metric that takes into account the relative frequency of alleles of each gene in the sample of the population present in the MLST database to determine the distance between Sequence Types (See Appendix), and the groups(putative Clonal Complexes) formed by the proposed algorithm can be evaluated by the measures proposed in Chapter V.

6.2. New solutions and New Problems

The paradigm shift that high-throughput methods and computational advances are imposing on Biology in general is also being reflected in microbial typing: The huge amount of data that is fast becoming available necessitates the use of integrative approaches for the analysis of data and data-driven approaches are needed in conjunction to the classical hypothesis driven theories. This is the new Systems Biology, which is already assumed as the Biology of the XXI century.

Naturally, new ethical problems arise with these new approaches: Should all the information in databases be made freely available? How can the privacy of the subjects from whom the original samples where taken be preserved? How can we trust the data submitted by third-parties? These questions can only be solved by consortia responsible for validating the data available on these databases, and enforcing the privacy issues.

Also, the capacity to store huge quantities of data can provide the studies with a new level of quality control. Together with the data generated the complete experimental protocol can be stored: information from which reagents where used to the lab equipment can be stored and algorithms can mine the data in search for correlations. This could determine if results are being biased from reagent batches and equipment.

The new challenge will be how to store, exchange, analyze and interpret the ever-growing wave of data without loosing perspective of the details.

6.3. References

1. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol **33:**2233-9.

APPENDIX

fBURST: Enhancing eBURST with an information-theoretic similarity metric

João Carriço[1], Francisco Pinto[1], Mário Ramirez[2] , Jonas Almeida[1,3]

[1]Biomathematics Group,Instituto de Tecnologia Química e Biológica, Oeiras,Portugal
[2]Institute of Molecular Medicine, Lisbon Faculty of Medicine, Lisbon
[3]Department Biostatistics, Bioinformatics, and Epidemiology, Medical University South Carolina, Charleston, South Carolina

Keywords: Microbial Typing, Multi-Locus Sequence Typing(MLST), eBURST, Information-Theoretic Similarity metric

Abstract

Multi-locus Sequence Typing (MLST) is a microbial typing method based upon sequencing internal fragments of 7 housekeeping genes of a given strain and then assigning to each unique allele a number. The final Sequence Type (ST) of a strain is a unique 7 number code that characterizes the isolate. The algorithm used to calculate relationships between strain types is eBURST (enhanced Based Upon Related Sequence Types). Here we propose the use of an information-theoretic similarity metric that takes into account the relative frequency of alleles of each gene in the sample of the population present in the MLST database to determine the distance between STs. This metric can extend the eBURST algorithm, providing further insight to phylogenetic relations between STs, since the allele frequency for each gene varies widely in the majority of microorganisms to which MLST has been applied.

Published in: João Carriço, Francisco Pinto, Mário Ramirez and Jonas Almeida, “fBURST: Enhancing eBURST with an information-theoretic similarity metric”, Procedings of BKDB2005 -Bioinformatics: Knowledge Discovery in Biology , Lisbon, 17 June 2005

Introduction

Multi Locus Sequence Typing (MLST)(4) is a microbial typing method based upon sequencing ~450-500 base pairs internal fragments of 7 housekeeping genes of a given strain and then assigning to each unique allele a number, after comparing the sequence in an online database(1). The seven number code obtained, designated Sequence Type (ST), is also compared with the online database to determine if the ST was previously encountered.
Determining accurately the double strand sequence of internal fragments with automated DNA sequencers, allows MLST to have the inter-laboratory portability and accuracy, desired in typing methods used for tracking bacterial populations, while retaining discriminating power.
Because of its characteristics, MLST has become widely used in molecular epidemiology surveillance and microbial population studies. In those fields, the typing method must have the ability to determine the relationship between strains. For MLST, the eBURST(2) algorithm is commonly used, being preferred to dendrogram representations, which provide poor representations of clonal emergence and diversification. The first step of eBURST algorithm is the group creation. Every ST within an eBURST group has a user-defined minimum number of identical alleles (n) (typically n=6, creating the most exclusive group definition) in common with at least one other ST in the group. Group assignment of STs is mutually exclusive: a ST belongs only to a group. Using this group partition method, several groups are created and some have only an ST. These are called singletons, since they share only n-1 or less alleles with other ST in the data set. The second step in the algorithm is the primary group founder determination. The primary founder is predicted on the basis of parsimony as the ST that has the largest number of Single Locus Variants (SLVs: a single allele difference). In case of two ST sharing the same number of SLVs, the one with more Double Locus Variants (DLVs) is considered the founder. The next step in the algorithm is assigning a statistical significance for each of the group founders. This is performed using a bootstrap resampling procedure, where for each group, resampling with replacement is performed a user-defined number of times (typically 1000 times), and then the primary founders are re-assigned for each group as previously described . A bootstrap value of 100% would be assigned to a ST considered group founder for all the resamplings.
Using the eBURST algorithm, will also produce subgroups and subgroup founders. These are STs connected with the group founder and already have more than two SLVs of their own. Finally a topology optimization procedure always maximizes the SLVs to the primary group founder. The typical graphical output of eBURST algorithm is presented in Fig 1.

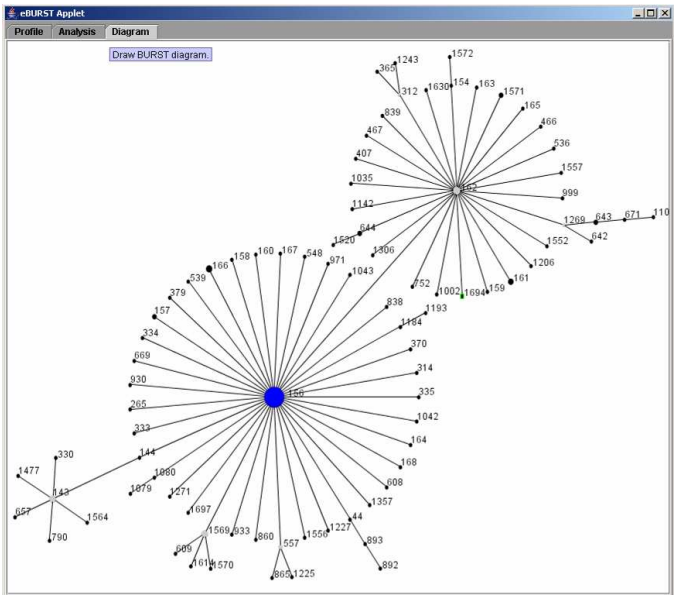


Figure 1 – eBURST graphical output for group 1 determined from the entire ST database of *Streptococcus pneumoniae*. In blue is the primary group founder (ST 156) and in grey are the subgroup founders

Observing the allele frequencies for the microorganisms in the MLST databases, it strikes out the fact that they differ from locus to locus and, at each locus, there is a predominant allele. In Figure 2, we represent the allele frequency for gene *aro*e in *Streptococcus pneumoniae* MLST database. In this study, we propose a similarity metric that reflects the different allele frequency, and allows greater flexibility in measuring ST relationships than simply counting the number of differences between two alleles. This similarity metric, proposed by Lin(3), derives of a definition of similarity using Information Theory concepts.

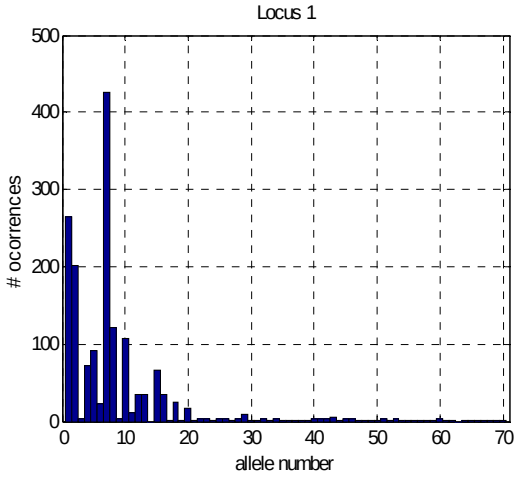


Figure 2 – Allele frequency for *aro*e gene (locus 1) in *S. pneumoniae* MLST database

In this definition , the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are. We defined the information needed to fully describe an ST as the logarithm of the product of the frequencies of the observed alleles on that ST, resulting on the following formula for similarity between STs:

$$sim(ST1, ST2) = \frac{2\log\prod_{i=1}^{m} f_i}{\log\prod_{j=1}^{t} f_j^{ST1} + \log\prod_{j=1}^{t} f_j^{ST2}}$$

Equation 1 – Similarity between STs sharing m alleles of a total t alleles.

Results

One of the characteristics of the information-theoretic similarity metric used, is the capability to distinguish between SLVs. Since the metric is based upon the frequency of the alleles, 3 STs , SLVs of each other, have different similarity values between them if the non-shared allele frequency varies. A comparison between SLVs with the non-shared allele having low frequency in both, would yield a lower similarity value than with one (or both) non-shared alleles with high frequency. Also the frequency of the shared alleles plays a part on the similarity metric. Lower frequency in the shared alleles yields higher similarity values than higher frequency. This nicely translates the fact that STs with rarer (lower frequency) alleles shared must be more related than STs with common (high frequency) alleles.

Calculating the similarity using this metric for the *Streptococcus pneumoniae* MLST database, we also found out that the lowest similarity value for SLVs in the database was 0.58 and, at those similarity similarity values, DLVs and Three Locus Variants (TLVs) were also found. The biological meaning of this finding is still under study, but it indicates that sometimes is more probable changing 2 or 3 higher frequency alleles in the population than a single lower frequency allele.

Using the 0.58 similarity value as a cut-off value for group formation (all STs with similarity greater than 0.58 where considered belonging to the same group) and comparing with eBURST group 1 for *Streptococcus pneumoniae* we also found that 17 STs found in eBURST Group 1 where not connected to ST 156 (i.e. had similarity values lower than 0.58) using our method and that 5 eBURST singletons were considered connected to ST 156.

Further studies are in progress to access this similarity in different microorganisms that have different recombination rates in the population for the genes used in their MLST schemas.

Conclusions

The information-theoretical definition of similarity used in this study is an 'evolutive' metric since as new alleles are added to the database, their frequencies change and the distances change reflecting the new allele distribution. This can effectively enhance the eBURST algorithm, opening new possibilities when exploring strains relatedness and studying microbial population biology.

References

1. **Chan, M. S., M. C. Maiden, and B. G. Spratt.** 2001. Database-driven Multi Locus Sequence Typing (MLST) of bacterial pathogens. Bioinformatics **17:**1077-83.
2. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol **186:**1518-30.
3. **Lin, D.** 1998. Presented at the Proceedings of International Conference on Machine Learning, Wiscosin.
4. **Spratt, B. G.** 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Curr Opin Microbiol **2:**312-6.

Curriculum Vitae

João André Nogueira Custódio Carriço was born in 24 September 1976, in Amadora, Portugal. From 1994 to 2000 he frequented the Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, where he obtained the degree in Applied Chemistry, Biotechnology branch. During his degree he did a research training in Epidemiological Data Analysis and Management under the supervision of Prof Dr. Jonas Almeida and Prof. Dra. Herminia de Lencastre, that he followed to his PhD project.
During his PhD training , he was responsible for the computer–assisted gel analysis module of Pulsed-Field Gel Electrophoresis EURIS workshop, held at The Rockefeller University, New York USA, from 15 February 2001 to 15 March 2001.
In 4-7 June 2002 he presented a talk on “Design of an Epidemiological System: A proposal based on the CRISTAL kernel” on the IEEE Computer Based Medical Systems 2002 in Maribor, Slovenia, as a result of a collaboration with the Centre for Complex Cooperative Systems, Faculty of Computing, Engineering and Mathematical Sciences, University of West of England, Bristol, UK and the ETT-IPR Group at CERN, Genève, Switzerland.
In 23 November 2002 he presented “The project EURIS online database at the ENEMTI (ESF Network for Exchange of Microbial Typing Information) workshop “Latest Advances in Molecular Fingerprinting Methods for Bacteria and the Construction of Computer Databases”, in Barcelona, Spain.
In 17 June 2005 his work “fBURST: Enhancing eBURST with an information-theoretic similarity metric” was selected for a talk in BKDB2005 - Bioinformatics: Knowledge Discovery in Biology, Lisboa, Portugal
During is PhD training he participated in also had several short (1-2 months) visits to the Department of Biometry, Medical University of South Carolina, Charleston, SC, USA, and had an active participation on two European union projects: 5th Framework project EURIS - European Resistance Intervention Study - QLK2-CT-2000-01020 (http://euris.itqb.unl.pt) and 6th Framework project PREVIS - Pneumococcal Resistance Epidemicity and Virulence An International Study - LSHM-CT-2003-503413(http://previs.itqb.unl.pt).

Publications in international scientific periodicals with referees

Almeida, J.S., C. Chen, R. Gorlitsky, R. Stanislaus, M. Aires-de-Sousa, P.Eleutério, J. Carriço, A. Maretzek, A. Bohn, A. Chang, F. Zhang, R. Mitra,G.B. Mills, X. Wang, and H.F. Deus. 2006. Data integration gets “Sloppy”.Nature Biotechnology. 24:6-7.

Carriço, J.A.; C. Silva-Costa; J. Melo-Cristino; F. R. Pinto; H. de Lencastre; J.S.Almeida; M.Ramirez, Illustration of a Common Framework for Relating Multiple Typing Methods by Application to Macrolide-Resistant Streptococcus pyogenes, J Clin Microbiol. 2006 Jul;44(7):2524-32

Mato R, Sanches IS, Simas C, Nunes S, Carrico JA, Sousa NG, Frazao N, Saldanha J, Brito-Avo A, Almeida JS, Lencastre HD.Natural History of Drug-Resistant Clones of Streptococcus pneumoniae Colonizing Healthy Children in Portugal.Microb Drug Resist. 2005 Winter;11(4):309-22

Carriço, J.A. , F.R. Pinto, C. Simas, S. Nunes, N. G. Sousa, N. Frazão, H. de Lencastre, and J.S. Almeida. 2005. Assessment of band-based similarity coefficients for automatic Type/Subtype classification ofmicrobial isolates analyzed by Pulsed-Field Gel Electrophoresis. J Clin Microbiol. 2005 Nov;43(11):5483-90

Sousa N. G., R. Sá-Leão, M.I. Crisóstomo, C. Simas, S. Nunes, N. Frazão, J. Carriço, R. Mato, I. Santos-Sanches, and H. de Lencastre. 2005. Properties of novel international drug-resistant pneumococcal clones identified in day-care centers (DCCs) of Lisbon, Portugal. J Clin Microbiol. 2005 Sep;43(9):4696-703

I. Serrano, J. Melo-Cristino, J. Carriço and M. Ramirez, Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal, J Clinical Microbiology,2005 Apr;43(4):1706-15

Nelson Frazão; António Brito-Avô; Carla Simas; Joana Saldanha; Rosario Mato; Sónia Nunes; Natacha Gonçalves Sousa; João Carriço; Jonas Almeida; Ilda Santos-Sanches; Hermínia de Lencastre, Effect of the 7-valent conjugate pneumococcal vaccine on carriage and drug resistance of Streptococcus pneumoniae in healthy children attending day-care centers in Lisbon Pediatr Infect Dis J. 2005 Mar;24(3):243-52.

Nunes S, R Sá-Leão, J Carriço, CR Alves, R Mato, A Brito Avô, J Saldanha, JS Almeida, I Santos Sanches, and H de Lencastre (2005) Trends in drug resistance, serotypes and molecular types of Streptococcus pneumoniae colonizing pre-school age children attending day care centers in Lisbon, Portugal – a summary of four years of annual surveillance.J. Clin. Microbiol. 2005;43 1285-1293

Sara Silva, Rodrigo Gouveia-Oliveira, Antonio Maretzek, Joao Carrico,Thorolfur Gudnason, Karl G Kristinsson, Karl Ekdahl, Antonio Brito-Avo, Alexander Tomasz, Ilda S Sanches, Herminia de Lencastre and Jonas S Almeida.Web-based epidemiological

surveillance of antibiotic-resistant pneumococci in Day Care Centers BMC Medical Informatics and Decision Making 2003, 3:9

Maria Miragaia, Isabel Couto, Sandro F. F. Pereira, Karl G. Kristinsson, Henrik Westh, Jens O. Jarløv, João Carriço, Jonas Almeida, Ilda Santos Sanches and Hermínia de Lencastre.Molecular Characterization of Methicillin Resistant Staphylococcus epidermidis (MRSE) Clones: Evidence of Geographic Dissemination J Clin Microbiol 2002 Feb;40(2):430-8

Santos Sanches, I., R. Mato, H. de Lencastre, A. Tomasz, CEM/NET collaborators: S. Nunes, C. R. Alves, M. Miragaia, J. Carriço,I. Couto, I. Bonfim, M. A. de Sousa, D. Oliveira, A. Gomes, M. Vaz, S.Fernandes, S. C. Verde, S. Ávila, F. Antunes, R. Sá-Leão, J. Almeida,and International collaborators: O. Melter, M. Chung, M. C.Brandileone, E. Castañeda, I. Heitmann, M. Hortal, W. Hryniewicz, F.Jia, K. Kikuchi, K. G. Kristinsson, J. Liñares, A. Rossi, E. Z. Savov,J. Schindler, F. Solorzano-Santos, K. Totsuka, M. Venditti, P. Villari,H. Westh, J. S. Wu, and R. C. Zanella. 2000. Patterns of multidrug resistance among methicillin-resistant hospital isolates of coagulase-positive and coagulase-negative staphylococci collected in the international multicenter study RESIST in 1997 and 1998. Microb. Drug Resist. 6:199-211

Papers in conference proceedings

Tony Solomonides, Mohammed Odeh, Richard McClatchey,Jean-Marie Le Goff, Joao Carrico,Jonas Almeida.Conceptual Modelling of an Epidemiological Information System, IADIS International Conference e-Society 2003, ISBN:972-98947-0-1

João Carriço, Francisco Pinto, Mário Ramirez , Jonas Almeida, fBURST: Enhancing eBURST with an information-theoretic similarity metric , BKDB2005 - Bioinformatics: Knowledge Discovery in Biology 2005, ISBN:972-9348-12-10