Journal of Biomedical Informatics 42 (2009) 66–73

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Biosurveillance of emerging biothreats using scalable genotype clustering

Blanca Gallego [a,*], Vitali Sintchenko [a,b,c], Qinning Wang [b], Lester Hiley [d], Gwendolyn L. Gilbert [b,c], Enrico Coiera [a]

[a] Centre for Health Informatics, University of New South Wales, Coogee Campus, Sydney, NSW 2052, Australia
[b] Centre for Infectious Diseases and Microbiology, Institute of Clinical Pathology and Medical Research, Sydney West Area Health Service, Westmead, NSW 2145, Australia
[c] Western Clinical School, The University of Sydney, Sydney, NSW 2145, Australia
[d] Queensland Health Forensic & Scientific Services, Brisbane, Qld 4001, Australia

ARTICLE INFO

Article history:
Received 24 December 2007
Available online 29 July 2008

Keywords:
Biosurveillance
Molecular genotyping
Salmonellosis
Infectious disease clusters

ABSTRACT

Developments in molecular fingerprinting of pathogens with epidemic potential have offered new opportunities for improving detection and monitoring of biothreats. However, the lack of scalable definitions for infectious disease clustering presents a barrier for effective use and evaluation of new data types for early warning systems. A novel working definition of an outbreak based on temporal and spatial clustering of molecular genotypes is introduced in this paper. It provides an unambiguous way of clustering of causative pathogens and is adjustable to local disease prevalence and availability of public health resources. The performance of this definition in prospective surveillance is assessed in the context of community outbreaks of food-borne salmonellosis. Molecular fingerprinting augmented with the scalable clustering allows the detection of more than 50% of the potential outbreaks before they reach the midpoint of the cluster duration. Clustering in time by imposing restrictions on intervals between collection dates results in a smaller number of outbreaks but does not significantly affect the timeliness of detection. Clustering in space and time by imposing restrictions on the spatial and temporal distance between cases results in a further reduction in the number of outbreaks and decreases the overall efficiency of prospective detection. Innovative bacterial genotyping technologies can enhance early warning systems for public health by aiding the detection of moderate and small epidemics.

1. Introduction

Prospective infectious disease surveillance requires the ongoing collection and monitoring of infection-specific data and related information such as infectious disease counts or syndromic data. The goal of surveillance is to detect and then prevent and control outbreaks in real-time. For some infectious diseases, one or two confirmed cases are sufficient to raise an alarm (e.g. SARS, meningococcal disease). However, for many types of infections, detection requires clustering of the data based on similarity of isolates. A broad range of statistical techniques have been applied in order to improve the performance of prospective surveillance and have been extensively reviewed elsewhere [1–4]. In its simplest form, a statistical surveillance method consists of a process control algorithm for a single time-dependent variable. More complex methods involve the analysis of multi-variate spatio-temporal data sets. These early warning systems can identify large disease epidemics but there are usually significant delays and low sensitivity in detecting moderate and small outbreaks. This is due to the high level of noise in laboratory and syndromic surveillance data [4]. Better surveillance often allows the size of outbreaks to be limited as a consequence of public health interventions, as more outbreaks are detected and controlled at an earlier stage and fewer continue to a large size [5].

The molecular fingerprinting of pathogens with epidemic potential offers new opportunities for detecting and confirming clusters of community and hospital-acquired infections [6–8]. It involves rapid subtyping of isolates from infected patients for the purpose of strain discrimination. Although the discriminatory power varies according to the subtyping method, molecular genotyping is often useful to identify sources and routes of transmission [9]. However, identifying patients that share the same genotype is not enough to uniquely provide an operational definition for an outbreak. In practice, the decision to proceed with a public health intervention will depend on the severity, communicability and local epidemiology of the disease as well as on the availability of public health resources to conduct investigations and institute corrective measures [2,10]. It is therefore critical to have an outbreak definition (in the absence of epidemiological information) that optimizes the limited resources of public health practitioners while preventing further spread [11].

* Corresponding author. Fax: +61 2 9385 9006.
E-mail address: b.gallego@unsw.edu.au (B. Gallego).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.07.002

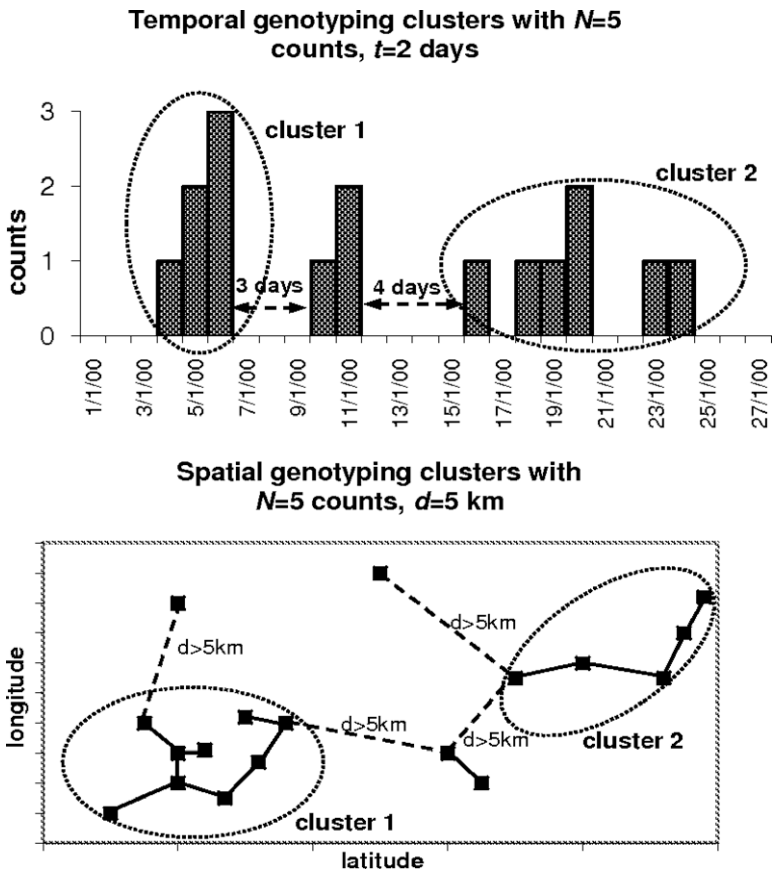

Fig. 1. Sketch depicting examples of temporal (top panel) and spatial (bottom panel) genotyping clusters. The top panel shows two temporal clusters defined as a maximal set of at least five counts with consecutive cases occurring at most 2 days from each other. The bottom panel shows two spatial clusters defined as a maximal set of at least five counts forming a spanning tree with edges at most 5 km long.

One setting that allows in-depth study of the impact of cluster definitions on prospective monitoring of bacterial genotypes is surveillance of *Salmonella enterica* serovar Typhimurium (STM) infections [12,13]. Rapid genotyping of STM has recently been widely used to characterize salmonella outbreaks. In particular, multilocus variable-number tandem repeat analysis (MLVA) of STM is a stable, easily implemented method and its results can be shared between laboratories over the Internet [14,15]. However, evidence about performance and timeliness of STM cluster detection systems remains limited [16].

To address these generic deficiencies, we introduce a working outbreak definition based upon temporal and spatial clustering of genotypes that provides unambiguous clustering of isolates and that can be tuned to accommodate the requirements and resources available for outbreak investigations. We compare this definition against statistical and epidemiologically confirmed clusters and evaluate its performance in prospective surveillance.

2. Methods and data

2.1. Genotype cluster definitions

We define a *genotyping cluster* as a *maximal set* of at least N isolates that share the same (or closely related) genotype, among a set of isolates from infected patients, each with an associated date and location (e.g. collection date and patient's address). To account for clustering in space and time, we specify:

Temporal cluster: A genotyping cluster, for which the time difference between any two consecutive cases is at most t days (see top panel in Fig. 1). The limit of $t = 0$ corresponds to clusters that last one day.

Spatial cluster: A genotyping cluster, for which locations of all cases are connectable by a spanning tree (a graph connecting a set of nodes [i.e. case locations] without any cycles) with all edges no more than d kilometers long (see bottom panel in Fig. 1). The limit of $d = 0$ indicates a cluster occurring in one location.

Spatio-temporal cluster: A combined temporal and spatial cluster characterized by parameters t and d.

These spatial and temporal cluster definitions satisfy two important properties. First, they provide a unique way of clustering cases that is independent of the order in which the isolates are considered. This property guarantees that any two cases assigned to a cluster at a given time will remain in one cluster in the presence of additional cases. This makes it possible to search, retrospectively, for clusters (for given parameters N, t and d) in historical data, compute the number of clusters and determine how early they would have been detected, prospectively. In this way, one can adjust future values of N, t and d according to prospective surveillance needs and availability of public health resources. For simplicity we have assumed that the parameters N, t and d are independent of genotype. Second, except for the limits $t = 0$ and $d = 0$, the duration and area of a cluster is not prescribed, making definitions scalable. A more naive outbreak definition as a set of at least N isolates of a given genotype occurring within a fixed duration and/or fixed area does not fulfill these properties. Furthermore, definitions with fixed duration are obviously not appropriate for prospective surveillance.

An algorithm that implements the working definitions of outbreaks described in this paper has three steps: