Using 'Big Data' approach to map relationships between human and animal diseases

July 21 2014

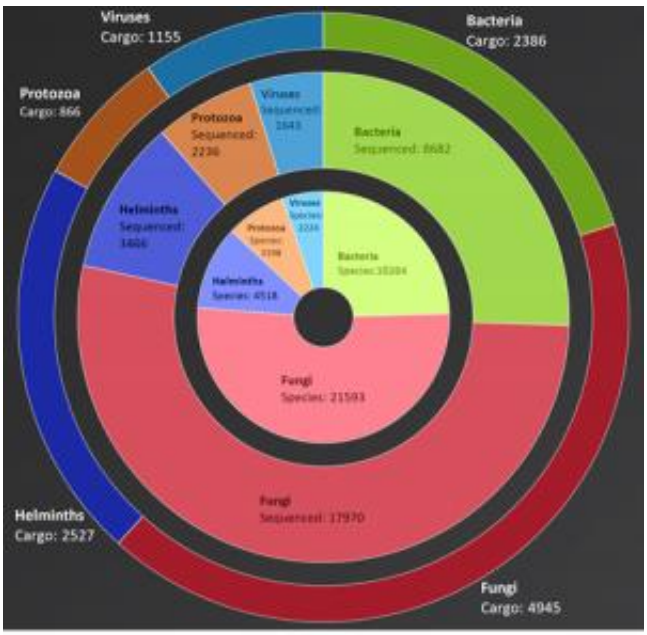

Figure 1: This visualisation depicts the pathogen species that EID2 currently has information on. The innermost circle represents the number of species listed within the NCBI taxonomy database in the major groupings that contain pathogens (denoted as 'Species', www.ncbi.nlm.nih.gov/taxonomy). The intermediate circle is those for which sequences (and metadata potentially describing their host origin) are available in the NCBI Nucleotide database (denoted as 'Sequenced', www.ncbi.nlm.nih.gov/nucleotide). The outermost

circle represents pathogen species for which data has been captured about the hosts in which they occur within the EID2 database itself (denoted as 'Cargo'). Credit: Dr Maya Wardeh, LUCINDA team

Researchers at the University of Liverpool's Institute of Infection and Global Health are building the world's most comprehensive database describing human and animal pathogens, which can be used to prevent and tackle disease outbreaks around the globe.

The Enhanced Infectious Diseases (EID2) database has been developed by the Liverpool University Climate and Infectious Diseases of Animals (LUCINDA) team and is funded by a BBSRC Strategic Tools and Resources Development Fund grant.

Effectively mapping the relationships between human and animal diseases and their hosts, disease-causing pathogens and the ways in which pathogens are transmitted can offer huge benefits when it comes to knowing what the disease risks are in a population or geographical area, and how best to manage and eliminate them.

The EID2 team realised that there was a potential treasure trove of data already available in the scientific literature and in pre-existing databases, which was just waiting to be mined for useful insights – a 'Big Data' approach. 'Big Data' is about utilising large datasets which may already have been collected, but which may be unstructured, and not fit into a conventional data-frame, by using often high performance and/or complex computing technologies. The emphasis on Big Data has increased recently because people have realised that the data that they have collected routinely, if used cleverly, can contain much more useful and potentially extra information than previously thought.

By using openly accessible information in a new way, data from EID2 has been used in work to trace the history of human and animal diseases, to predict the effects of climate change on pathogens, to produce maps of which diseases are most likely in some areas and to categorise the complex relationships between human and animal carriers and hosts of numerous pathogens.

Epidemiologist Dr Marie McIntyre, one of the EID2 team, said: "The database is matchless in scale, and has the capacity to hold data on all known human and animal pathogens, when detailed information becomes available."

"We use largely automated procedures to collate data on human and animal pathogens: where, when, and in which hosts there is evidence of their occurrence.

"After scientists have sequenced part or all of a pathogen's DNA or RNA, they usually upload the sequence to public databases, and include information (called metadata) on where, when and from which host the pathogen was obtained.

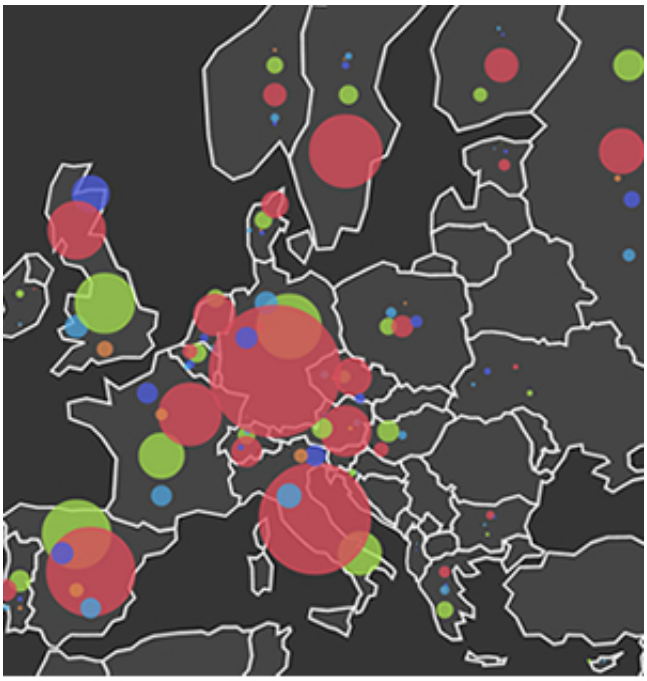

Figure 2: This image describes the number and types of pathogens found in EU countries. The size of each circle is proportional to the number of pathogen species the EID2 team has evidence of, and the different coloured circles represent: bacteria (greenish yellow), fungi (red), helminths (purple blue), protozoa (orange), and viruses (blue). Credit: Dr Maya Wardeh, LUCINDA team

"EID2 is unique in extracting the information on pathogens from this metadata and as there are already tens of millions of sequence uploads to look at (see Fig. 1), and millions more are added every year, EID2 has the capacity to become a comprehensive, definitive source of pathogen and disease information.

"In addition, the sequence data is supplemented with information from

the NCBI's database of scientific publications, PubMed. The procedures used for identifying the hosts in which pathogens occur, and where they occur, are objective, and the information the EID2 contains can be regularly updated and improved as more detailed information becomes available."

All together there are more than 60 million pieces of data that have been brought together for EID2, with new information added all the time. The database is open-access, allowing registered researchers to use it, and the data can be manipulated in lots of ways to help scientists to tackle numerous questions.

Dr McIntyre said: "EID2 is useful because it gives access to sets of information on infectious pathogens which have, until now, been difficult to acquire. For example, it describes all of the known pathogens of a host species, and all of the hosts of a pathogen species. It can generate all of the recorded pathogens in a specific country or region (for example Fig. 2), or all of the pathogens of a certain host in a specific country. It gives instant access to the raw data from which this information is built. It also allows the distribution of pathogens (and hosts) to be mapped."

This disease mapping is one of the most important areas where EID2 can be a valuable tool.

Research has shown that only four percent of clinically-important diseases in humans have been geographically mapped, despite half having a strong rationale for mapping.

Because EID2 can pull together novel data sources, it can quickly and accurately map diseases, and because it isn't limited as to which pathogens and hosts it can describe, it has the potential for large-scale global mapping of animal and crop diseases in the same way as is

currently being undertaken for human and some animal diseases.

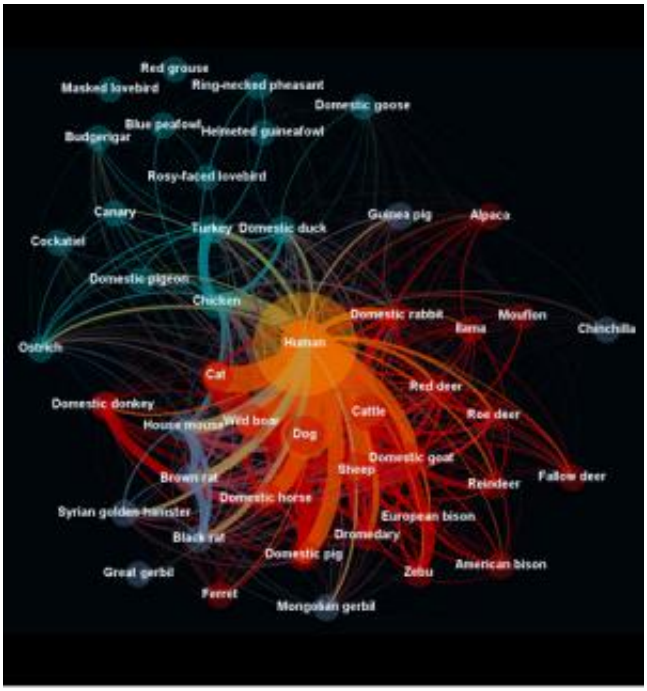

Figure 3: This depicts the relationship between information currently in the EID2 on domestic animal and human host species. The size of the nodes is relevant to the number of pathogen species found for each host; the arrows linking these nodes show the number of pathogen species shared by each pair of hosts (the thicker the arrow, the greater the number of pathogen species); and the colour of the nodes is related to the type of host (humans, other mammals, birds, rodents). Credit: Dr Maya Wardeh, LUCINDA team

This can produce country-by-country, or even county-by-county, profiles of the factors affecting disease, for example factoring in prevailing climate conditions, meaning regions can best prepare to avert or manage outbreaks of emerging infections.

Dr McIntyre added: "A further valuable data provision for the EID2 is in quantifying the interactions between pathogens and their hosts using approaches such as network analysis (for example Fig. 3). This is important because, for example, we know that humans originally acquired about 60% of our pathogens from animals, but we don't know which animals. Nor do we know where those animals got their pathogens from. Once we have a clear picture of the pathogen species found in different domestic and wild animal hosts, we will be able to study the possible routes by which pathogens make it into human populations.

"Is a new pathogen of, say, mice going to reach humans because they interact with us in our houses, or will it be via the cats that eat them? This kind of information is ordinarily very difficult to acquire because it requires knowledge of which hosts are affected by diseases and vice-versa, but also how often these infection events occur, or if they have ever occurred."

The EID2 data is already being used to contribute to work on emerging and zoonotic infections at a Health Protection Research Unit (HPRU) at the University of Liverpool, which was created at the end of 2013, a national centre of excellence in multidisciplinary research to protect the nation's health.

In fact the potential of the EID2 data for analysis is incredibly wide ranging, and a host of exciting ideas are being considered by the Liverpool team. Among these are plans to use data for risk analysis, predicting where and in which species certain diseases are most likely to occur, and producing estimates of where diseases can occur based on environmental data such as climate, demographics and vegetation.

Provided by University of Liverpool