Evo Edu Outreach (2010) 3:451–455
DOI 10.1007/s12052-010-0258-5

INTERNET COLUMN

The NCBI Databases: an Evolutionist's Perspective

Adam M. Goldstein

Published online: 10 August 2010

Abstract The National Center for Biotechnology Information (NCBI) organizes information resources for life scientists on an evolutionary scheme. This facilitates research about present-day organisms. The recent discovery of a new arenavirus, the LUJO virus, illustrates the utility of adopting evolution as a central architectural principle for life sciences databases: using the NCBI's resources, clinicians were able to classify the new virus in real time—soon enough to aid in the diagnosis and treatment of a hemorrhagic fever caused by the LUJO virus. Topics fundamental to the study of evolution, often thought of as useless, are indeed vital because they inform how life science information ought to be organized.

Keywords NCBI · National Center for Biotechnology Information · Database architecture · Arenaviridæ · Evolution · Information resources · Genome · High-throughput gene sequencing

Introduction

The National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine, is responsible for a suite of databases, at http://www.ncbi.nlm.nih.gov/, which it maintains and continues to develop. These databases, available worldwide free of charge and requiring no authentication by password or any other registration, are among the most important information resources used by researchers: they contain records for a significant proportion of the world's literature and data in the life sciences. In accord with the famous statement "nothing makes sense except in light of evolution" (Dobzhansky 1973), the manner in which the NCBI organizes information about the living world reflects the organization, by lines of descent, of the living world itself.

In this paper, I illustrate the power of organizing information this way. I report on recent research in epidemiology and clinical medicine: using gene sequencing technology, the NCBI's information resources, and exceptional scientific sleuthing skills, an international team has developed a method to identify a previously unknown pathogen rapidly enough to influence decisions about medical treatment and measures to protect public health almost as soon as an infected individual's symptoms first appear.

An Unexplained Outbreak of Hemorrhagic Fever

Paweska et al. (2009) describe the case of a patient who contracted and died of an illness which could not be identified at the time of her death. Near Lusaka, Zambia, the patient first suffered chills and digestive symptoms and was treated for food poisoning and flu. New symptoms arose: chest pain, sore throat, and fever; a head-to-toe rash; facial swelling; and myalgia (muscle pain). The patient visited a local clinic, where these new symptoms were attributed to an allergic reaction

A. M. Goldstein (✉)
Department of Philosophy, Iona College,
715 North Avenue, New Rochelle,
NY 10801, USA
e-mail: z_californianus@shiftingbalance.org

to one of the drugs she was taking to treat her earlier symptoms. She was discharged but returned later the same day and was moved to a local hospital. She was subsequently airlifted to another in Johannesburg, South Africa. In Johannesburg, physicians found the patient unresponsive to light stimuli and to be suffering from acute respiratory distress syndrome, kidney failure, serious blood pathology, and a cerebral edema; an eschar (scab), believed to be a result of rickettsiosis, appeared on her foot. Despite aggressive treatment in Johannesburg for likely causes of her illness, the patient died within approximately two weeks of the onset of her symptoms in Lusaka, approximately four days after her admission to the Johannesburg hospital.

In the course of her transport to Johannesburg and her treatment there, three other individuals who had come into contact with the Lusaka patient or with one another became violently ill and died within approximately two weeks of their initial contact with an infected individual. A fourth new patient, the fifth if the Lusaka patient is counted, was also infected. In retrospect, after the death of the third newly infected patient, physicians identified the illness as hemorrhagic fever, a disease characterized by a broad spectrum of symptoms including those suffered by the Lusaka patient. The hemorrhagic fevers are divided into classes. What determines whether a given person's hemorrhagic fever falls into one class or another depends on what type of pathogen caused it, its locale, and its usual host—mouse, mosquito, or chicken, for example. There are four taxa of viruses responsible for the variety of hemorrhagic fevers, including the filioviruses, which include two of the most notorious viruses, Marburg and Ebola (CDC, Special Pathogens Branch 2010; National Library of Medicine, USA 2010b). Taking previous cases as a guide, a reasonable person would be justified in estimating the fifth patient's life expectancy at two weeks from the time her symptoms began.

Pathologists at the National Institute for Communicable Diseases (NICD) of South Africa tested samples from infected patients using a technique known as reverse-transcriptase polymerase chain reaction (RT-PCR). The polymerase chain reaction will create many, many copies of DNA genes from a sample containing only a few. This makes it possible to detect and study DNA from the sample, even if it contains a small amount of DNA. Reverse transcriptase PCR is similar. Arenaviruses use RNA as their hereditary material; RT-PCR creates a DNA template corresponding to viral RNA genes. This reverses the process of transcription that, in organisms using DNA as their primary genetic material, transcription creates a strand of RNA corresponding to the DNA genes; this RNA carries the information encoded in the DNA forward into the process of protein synthesis. The results of the NICD's RT-PCR tests were negative for hemorrhagic fever.

In the meantime, tissue and blood samples were sent to the United States' Special Pathogens Branch of the Centers for Disease Control (CDC), in Atlanta, Georgia. There, immunohistochemical testing was performed. In an immunohistochemical test, a sample from the patient is treated with antibodies known to react with the suspected pathogen; if the reaction occurs, a stain applied to the sample will take on a distinctive color. Of course, if nothing is known about the likely pathogen, no antibodies can be designed for the test. Not having ruled out the arenaviruses, the CDC tested for them using antibodies that would react with a range of arenaviruses known to cause hemorrhagic fever. The test was positive and physicians in Johannesburg started treating the patient with ribavirin, an antiviral drug. The NICD sent tissue and blood samples to Atlanta on 9 October 2008; the samples were received in Atlanta on 10 October, and the CDC reported the results of its immunohistochemical assay on 11 October; the ribavirin treatment began on that same day. By 2 December, the patient had recovered from hemorrhagic fever; tests for viral RNA in the patient's blood and urine were negative.

The achievement of the medical personnel and scientists in Zambia, South Africa and the United States is impressive. Their first patient died for no reason anyone could discern at the time, and her illness was transmitted to four others, three of whom died in a similarly mysterious manner. A fifth patient contracted the illness and faced near certainty of deterioration and suffering followed by death within two weeks' time. As it happened, however, the transatlantic collaboration learned enough about the pathogen within 36 hours of the patient's admission to confidently recommend a therapy, which led to the patient's complete recovery.

A New Species of Arenavirus

Except in particularly clear cases, a distinction between clinical medicine, public health measures, and pure research in biology and epidemiology cannot be drawn. The events connected with the Lusaka patient exemplify this close interrelationship. Recall that the NICD had conducted RT-PCR tests on the samples from the infected individuals but found nothing. The CDC repeated these tests and found evidence of arenavirus. The immunohistochemical tests conducted by the CDC were conclusive enough to warrant treating

the patient with ribavirin; the RT-PCR tests, conducted at the same time, confirmed the immunohistochemical results. Moreover, the RT-PCR tests led to important results in the genetics and taxonomy of the *Arenaviridæ*. The NCBI's evolutionary framework for organizing information is essential to these kinds of discoveries.

As I described above, RT-PCR creates DNA templates corresponding to viral RNA, copying them many, many times, so that even if there is only a small number of genes in a sample, they can be detected and isolated for study. In some cases, in which the genes of interest are known, the purpose of RT-PCR is to obtain more copies of them to study their properties. In this case, the virus' genes were not known, and the purpose of the RT-PCR test was to aid in their identification. In short, if enough genes can be identified, their order can be determined, and longer stretches can be constructed. These longer stretches are then tested for matches against a library of gene sequences from organisms whose genetic makeup is already known.

To match gene sequences with one another, scientists use an NCBI digital tool called the Basic Local Alignment Search Tool (BLAST). The researcher enters the gene sequence in which he or she is interested into the BLAST system, often by simply cutting and pasting it into a text entry box on the BLAST web site. This gene sequence is then compared by the BLAST program to the gene sequences in the NCBI's databases. These databases are themselves remarkable. When scientists identify a gene sequence, they submit it online to the NCBI, which provides access to it by way of a range of research tools, including BLAST. This exemplifies the idea that science is a collective enterprise; members of the scientific community have submitted enough gene sequences to construct 2,467 viral genomes.

If there is a very small probability that the sequence under study would be identical to a sequence in the database by chance, the two are considered a match. In the Lusaka virus case, the immunohistochemical staining indicated the presence of an arenavirus, but did not indicate one or another species. The BLAST system indicated that the gene sequences detected by the CDC's RT-PCR match "approximately 50% of a prototypic arenavirus genome" (Briese et al. 2009a). This is strong evidence that the unidentified virus is indeed a species of the *Arenaviridæ*. A phylum becomes more diverse—i.e., new species form—when a group of organisms in an already-existing species separates from others in the species and, with time, changes significantly enough to differ in kind from its former conspecifics. Scientists disagree about what biological changes are necessary before two diverging populations separate enough to be considered different species, although many agree that reproductive isolation is sufficient. Divergence is never total, however: related species share a considerable portion of their genomes, a legacy of their shared past.

The logic of the BLAST tool is the same that a professor might use to identify a plagiarized term paper. Suppose that a term paper contains the phrase "LCMV infection is only rarely fatal in immunocompetent adults; however, infection during pregnancy bears serious risks for mother and child and frequently results in congenital abnormalities." Suspecting that a student would not write this sentence, the professor searches the Internet for "LCMV infection is only rarely fatal in immunocompetent adults; however," the first part of the suspicious passage. It appears only once online, in the paper cited above (Briese et al. 2009a) about the Lusaka arenavirus. On one hypothesis, the passage from the article and the term paper match because the student used the same words to express ideas he or she shares with the scientists but did so independently. This is highly improbable. On a contrasting hypothesis, the passages match because they share a common history: Briese and colleagues expressed their thoughts by writing the passage, and the student copied the passage. This is clearly the stronger hypothesis. If the two instances of the passage share this common history, it is highly probable that they would be identical.

To complete the analogy, a viral gene sequence submitted to BLAST is like the student paper; the sequences in the BLAST database are like the Internet, searched by the professor for the source of the student paper. If the BLAST database contains a sequence *S*, and *S* matches a viral DNA sequence that is highly unlikely to have formed independently, it is returned as a search result. This provides warrant for the belief that the viral gene sequence and the sequence in the BLAST database are historically related. The more viral DNA that scientists can test against the BLAST database and the larger the BLAST database, the more reliable and more precise the results. As I mentioned previously, the CDC determined that the Lusaka virus and the "prototypic arenavirus genome" were a 50% match, enough to safely presume that the Lusaka virus is indeed an arenavirus. This took 72 hours.

Discovering that the previously unknown pathogen is a previously unidentified arenavirus is an important achievement and, as I have stressed, eminently useful in real time for the clinician. Nonetheless, the researchers in this case went further: They sequenced

the *entire genome* of the new virus. Samples of the NICD-CDC materials were sent to Columbia University in New York, where researchers pieced together the otherwise-unaccounted-for genetic material using a high-throughput gene sequencing apparatus—so called because it can sequence a large number of genes in a short time. Although the sequence of events is hard to reconstruct from the published papers, it is clear that the Columbia researchers began to sequence the remainder of the new virus' genome without delay as soon as they received the samples, and most likely finished in a matter of days. Now in possession of the complete genome of the virus, epidemiologists went from having no information whatever about a pathogen and the illness it causes to having a broad base of important taxonomic and genetic information about it. These data were submitted to the NCBI's databases. Physicians, public health officials and researchers, and scientists doing basic research can now check new unknowns against it. The new virus is named LUJO, in recognition of its first known victims in Lusaka and Johannesburg. Its genome can be found in the NCBI's Entrez Genome database with the classification key `txid649188[orgn]` (Briese et al. 2009b).

The Importance of Comparative Studies

Knowledge of the LUJO virus' genome puts scientists in an excellent position to understand how the virus works. What is its life cycle? How does it survive in its usual hosts? How does it spread in host populations? By what means does it construct its viral "outer shell," which forms the compartment in which the viral genetic material is stored before it is released into host cells? What physiological processes in a human being does it disrupt in order to cause hemorrhagic fever? Contrary to what might seem obvious, these questions are not ahistorical, and the information resources used to explore them are organized to reflect this.

To be clear, the kind of research required to understand how the LUJO virus works makes use of present-day, real-time experiment and observation. Nonetheless, there is a comparative method fundamental to planning experiments and deciding what observations are most likely to be interesting and useful. The reason that placing the LUJO virus among the *Arenaviridæ* is so important is that taxonomic differences, as a rule, reflect differences important to understanding how related organisms work. Take, for instance, any present-day population of LUJO virus. The ancestry of this population can be traced back to some population in the parent species—individuals that began the separation which would become a permanent species difference. Some changes that occurred during that time are due to natural selection, that is, the adaptation of the LUJO virus to an environment different than that of its parent. Others are purely historical, in the sense that they do not result from natural selection, for instance, they make no difference or very little difference to any functional aspect of the virus's biology. Comparing the present-day genome of the LUJO virus to related species or species in other related phyla is a good technique for learning about how the LUJO virus works. Suppose a certain gene in the LUJO virus has a certain function: how does that same gene differ in species sharing LUJO's ancestry? Accordingly, does the biology of the LUJO virus differ from its ancestors in any ways tied to their genetic differences?

The NCBI databases are organized in a way that facilitates this kind of comparative study. Using the NCBI Nucleotide database, a researcher can enter in the identification code for a given gene as a search key to find identical or similar genes across taxa. Similar searches can be carried out for across-taxa comparison using other databases such as Online Mendelian Inheritance in Man and Online Mendelian Inheritance in Animals, Protein, Structure, PopSet, and Conserved Domains, to name a few; these and others can be accessed from the NCBI web portal (National Library of Medicine, USA 2010a). With this comparative strategy, researchers can learn what a given gene does in a given taxon or set of taxa and, reasoning that it probably does something similar in the taxon under study, begin to see how genetic differences result in biological differences. Often, a daughter species differs from its parent species as a result of having adapted by natural selection to a new environment. Having information about what the structure and function (if any) of genes in related species offers a starting point for experiment and observation of a recently discovered species.

Concluding Remarks

Answering apparently ahistorical questions in clinical medicine and public health requires information about biological taxonomy and, more generally, about changes over time in genetics and the biological changes that result in new species or historically related populations within a species. Accordingly, the NCBI organizes scientific literature and data in a way

that reflects historical relationships at all levels of biological organization. By exploring NCBI databases, a researcher is, in effect, exploring evolutionary relationships among taxa, essential to drawing conclusions about present-day organisms. Our understanding of molecular biology and genetics is relatively young, compared with the practice of organizing plants and animals into kinds, one of humanity's oldest and widely pursued activities. Nothing makes sense in biology except in light of evolution. An important part of the great utility of the NCBI databases derives from the kinds of discoveries in molecular biology about present-day organisms that everyone recognizes enthusiastically as great advances: new medicines; deeper understanding of the connection between mind and body; insight into our physiology and metabolism; and improved crops and livestock. We would do well to show similar enthusiasm for what might seem at first to be useless studies in quaint subjects such as natural history, paleontology, biogeography, ecology, and experiments and observations of natural selection in guppies, finches, and snails and other apparently uninteresting organisms.

References

Briese T, Paweska JT, McMullan LK, et al. Genetic detection and characterization of LUJO Virus, a new hemorrhagic fever-associated Arenavirus from Southern Africa. *PLoS Pathogens*. 2009a; 5(5):e1000455. doi:10.1371/journal.ppat.1000455.

Briese T, Paweska JT, McMullan LK, et al. txid649188[orgn]—genome results. Entrez Genome records for the complete genome of the LUJO virus. 2009b. http://www.ncbi.nlm.nih.gov/portal/utils/pageresolver.fcgi?recordid=1278041 684752660. Accessed 1 July 2010.

CDC, Special Pathogens Branch. Filoviruses. 2010. http://www.cdc.gov/ncidod/dvrd/spb/mnpages/dispages/filoviruses.htm. Accessed 25 June 2010. Page reviewed 4 May 2010.

Dobzhansky TG. Nothing in biology makes sense except in the light of evolution. Am Biol Teach. 1973. http://www.pbs.org/wgbh/evolution/library/10/2/l_102_01.html. Accessed 2 July 2006.

National Library of Medicine, USA. National center for biotechnology information. NCBI Databases portal. 2010a. http://www.ncbi.nlm.nih.gov. Accessed 2 July 2010.

National Library of Medicine, USA. Hemorrhagic fevers. 2010b. http://www.nlm.nih.gov/medlineplus/hemorrhagicfevers.html Accessed 25 June 2010. Page updated 21 June 2010.

Paweska JT, Sewlall NH, Ksiazek TG, et al. Nosocomial outbreak of novel arenavirus infection, Southern Africa. Emerg Infect Dis. 2009; 15(10):1598–602.