Computer security in academia—a potential roadblock to distributed annotation of the human genome

Dov Greenbaum, Shawn M Douglas, Andrew Smith, Joanna Lim, Michael Fischer, Martin Schultz & Mark Gerstein

With the Blaster and SoBig virus outbreaks of the past summer and the daily nuisance of spam, computer security is, unfortunately, grabbing headlines (see **Fig. 1**)[1–5]. Security considerations adversely affect science, but in a somewhat different fashion from the rest of society. They directly increase the costs of using interoperating computers, diverting scarce resources from other activities. Furthermore, as we argue here, they exact a vast 'opportunity cost,' particularly in computationally intensive fields, such as genomics, by hindering important avenues of research.

Why security is a problem

Modern science is completely dependent on computers to analyze and archive data. Moreover, certain subfields, such as genomics, increasingly use interoperating computers extensively for communication and dissemination (**Fig. 1**). In these fields, servers house large centralized archives, which are continually accessed by a broad community of researchers (*e.g.*, PubMed[6] and Uniprot[7]). These central hubs link together a vast constellation of smaller more specialized resources, focusing on particular organisms or specific aspects of molecules (*e.g.*, Ensembl[8], flybase[9], MolMovDB[10] and scop[11]), and the online text of books and journals. Intricate interoperation is required between databases and tools at different sites to enable collaboration and annotation of large data sets. Often, this is through conventional web links, but increasingly more complex interfaces are employed[12]. For instance, to find out all the information associated with a particular human protein, one currently has to perform a distributed query over many disparate sites. Many believe, in fact, that the future annotation of the human genome will involve a massive federation of interconnecting and interoperating information servers[13–15].

Funding agencies must recognize that security is part of the bill for doing research and be willing to provide the necessary resources.

Unfortunately, computer security considerations make realizing this vision increasingly difficult. Frequent hacker attacks have made maintaining servers exposed to the internet very expensive. The most obvious cost is the mundane administration—for example, installing patches, maintaining nightly backups and monitoring for suspicious processes—that wastes countless hours better spent elsewhere. More subtly, security considerations make building intricate systems for interoperation between databases all the more difficult, as researchers have to continually check their interfaces for holes that would allow intruders in.

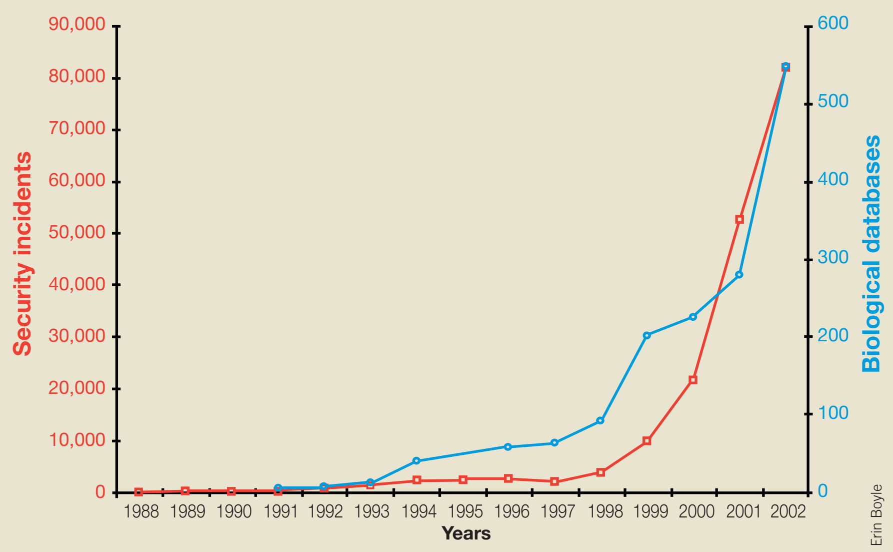
The nonscientific world, of course, also faces the costs of computer security—often with much greater financial resources. However, security considerations affect science differently from the rest of society. Academic research is uniquely structured in a way that makes it especially hard to protect. Free and broad dissemination of ideas between independent laboratories and the public is the hallmark of research. Preserving openness precludes standard security practices often employed in a corporate or military environment, such as using private networks for secure communications and imposing sophisticated systems of authority and control over the use of computers and information. Furthermore, academic computer users exhibit great variability, making effective security procedures and controls all the more difficult to implement. Users range from free-software gurus, who chaff at any restrictions, to students, who are often transient and in training. Students, moreover, in their hurry to get to the cutting edge of research often carry out rather delicate operations from a security standpoint quickly, with minimal orientation—for example, setting up a simple website in a biology laboratory with no formal computational training.

Finally, the aim of hackers attacking university computers is different from those aiming at corporate or military targets. In these later settings, the privacy of information, whether financial records or battle plans, is paramount. However, the computer hacker is not interested in the (usually

Dov Greenbaum is in the Department of Genetics; Shawn M. Douglas and Joanna Lim are in the Department of Molecular Biophysics & Biochemistry; Michael Fischer and Martin Schultz are in the Department of Computer Science; Andrew Smith and Mark Gerstein are in the Departments of Molecular Biophysics & Biochemistry and of Computer Science, P.O. Box 208114, Yale University, New Haven, Connecticut 06520, USA.
e-mail: Mark.Gerstein@yale.edu

obscure) data on university servers, but rather the hardware that houses it. (This situation with regard to privacy, of course, does not apply in a clinical setting, where ensuring the confidentiality of patient information is vital.) In effect, academic servers are the low-hanging fruit of the internet. A typical machine will have a high-speed connection, plenty of disk space and processing power, but will be protected only by the most minimal staff.

More pernicious than the considerable direct costs of computer security are the lost opportunities. Fear of assault against computing systems forces researchers to place undesired limitations on access to data and impedes database interoperation—that is, the site serving a particular bit of genome annotation is just never opened up, despite its obvious scientific value.



Erin Boyle

Figure 1 The number of computer security incidents over the past 15 years. Incidents, in fact, have increased at a faster rate than the growth of the internet. For the sake of comparison, the growth of online databases within the field of bioinformatics is shown. Security incidents are taken from http://www.cert.org/. Although representative, these statistics are only a small portion of the total number of incidents worldwide. Data for growth in the number of databases are taken from the annual *Nucleic Acids Research* issue on databases; again, although not exhaustive, the growth of the annual is representative of the general trends in the field of bioinformatics databases. (No annual was published in 1995.)

Evasive action

What to do? Unfortunately, when it was originally envisioned, the internet was not intended to be a secure platform for communication. Gone are the days when easy-to-use but insecure protocols, such as telnet, anonymous ftp and nfs, formed quick links in collaborative efforts.

Funding agencies must recognize that security is part of the bill for doing research and be willing to provide the necessary resources, and individual biomedical researchers need to understand that adequate security requires the guidance of professionals. Resources have to be allocated for full-time information-technology specialists in academia, and laboratories that are not focused on computational issues should work with them. A natural extension of this implies a movement toward centrally administered sites (perhaps on a university or departmental scale) aggregating the content from many small laboratories.

The US National Science Foundation (Washington, DC, USA) has emerging cyberinfrastructure efforts that may be worthwhile to build upon[16]. The Foundation's Middleware Initiative, which extends the Internet2 Shibboleth[17] and Globus[18,19] architectures, has been used for the underlying infrastructure in several projects, such as GriPhyN[20], the Network for Earthquake Engineering Simulation[21] and the Teragrid Supercomputing grid[22]. The international biomedical research community should develop efforts similar to these. Perhaps, the US National Institutes of Health (Bethesda, MD, USA) could start the ball rolling by setting up sponsored working groups.

A good initial goal of such efforts would be creating a community-wide system of identity management and authentication (perhaps via Shibboleth). This would greatly help in the interoperation of tools within a federated database framework. Many of the complicated interfaces of these databases and tools could be 'hidden' and accessible only after authentication. As such, community-wide authentication would provide a lightweight 'perimeter' defense against hackers interested in trying to exploit holes in interconnected systems, but would not significantly impede access to legitimate researchers. Although no panacea, measures such as these, which systematically reduce the exposure of servers to the internet at large, can diminish the severity of the security problem.

1. Butler, D. *Nature* **425**, 3 (2003).
2. Richmond, R. Workplace Security. *Wall Street Journal*, R4 (29 September 2003).
3. Richardson, R. *Eighth Annual CSI/FBI Computer Crime and Security Survey* (Computer Security Institute, San Francisco, CA, USA, 2003).
4. Committee on Institutional Cooperation. *Incident Cost & Analysis Modeling Projects (ICAMP) I* (CIC; Champaign, IL, USA, 1998). http://www.cic.uiuc.edu/groups/ITSecurityWorkingGroup/archive/Report/ICAMP.shtml
5. Committee on Institutional Cooperation. *Incident Cost & Analysis Modeling Projects (ICAMP) II* (CIC; Champaign, IL, USA, 2000). http://www.cic.uiuc.edu/groups/ITSecurityWorkingGroup/archive/Report/ICAMP.shtml
6. http://www.pubmed.gov/
7. http://www.uniprot.org/
8. Clamp, M. *et al. Nucleic Acids Res.* **31**, 38–42 (2003).
9. The FlyBase Consortium. *Nucleic Acids Res.* **31**, 172–175 (2003).
10. Echols, N, Milburn, D. & Gerstein, M. *Nucleic Acids Res.* **31** 478–482 (2003).
11. Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. *J. Mol. Biol.* **247**, 536 (1995).
12. Dowell, R. *et al. BMC Bioinformatics* **2**, 7 (2001). http://cabig.nci.nih.gov/caBIG/
13. Hubbard, T. & Birney, E. *Nature* **403**, 825 (2000).
14. Stein, L. *Nature* **417**, 119 (2002).
15. Gerstein, M. *Science* **288**, 1590 (2000).
16. http://www.communitytechnology.org/nsf_ci_report
17. http://shibboleth.internet2.edu/
18. http://www.globus.org/
19. Foster I. & Kesselman, C. *Int. J. Supercomputer Appl.* **11**, 115–128 (1997).
20. http://www.griphyn.org/
21. http://www.eng.nsf.gov/nees
22. http://www.teragrid.org/