OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

What Database Management System(s) Should Be Employed in Bioinformatics Applications?

PETER D. KARP

THIS PAPER ADDRESSES THE QUESTION: What database management system(s) (DBMS) should be employed in individual bioinformatics database (DB) applications? I posit that part of our research agenda should be to develop well-supported answers to this question, and to educate bioinformaticians and molecular-biologists in the answers.

There are several possible answers:

1. Do not use any DBMS (simply use flat files, which is how a significant number of molecular-biology DBs are now constructed)
2. Use open-source DBMSs such as MySQL
3. Use commercial DBMSs, which fall into two major categories:
   3a. Relational DBMSs
   3b. Object-oriented DBMSs
4. No existing DBMSs are suitable—we must build new ones

Let us discuss each alternative.

1. *No DBMS:* Although most computer scientists would dismiss this alternative as ridiculous, I would wager that a significant number of the hundreds of molecular biology DBs that now exist are not constructed using any DBMS. This should make us all stop and think. Either the biologists who are constructing these DBs know something that computer scientists don't know, or they don't know something that computer scientists do know. I believe the latter is the case—I believe that the vast majority of biologists have not received even the most elementary education in DBs, and have little idea of how DBMSs can help them, nor of how to go about using a DBMS. Note that the situation today is far improved over that of ten years ago. Today most large DB projects do employ a DBMS, whereas 10 years ago virtually none did so. Bioinformatics is still suffering from those mistakes.
2. *Open-source DBMSs:* Many of the molecular-biology DB projects that are employing DBMSs are employing open-source DBMSs to save money and because they believe that access to the source code will be useful in fixing any DBMS bugs that are discovered, or in extending the DBMS. My limited understanding of these DBMSs is that they have significant technical limitations compared to commercial DBMS product, yet little hard information about this issue is circulating in the bioinformatics community. We should obtain and circulate such information, whether from the literature or by empirical study. I also believe that the DBMSs are sufficiently complex that access to their source code is essentially useless to any but the most experienced DBMS hacker (and perhaps not even to them), and that this point must be articulated to the community.
3. *Commercial DBMSs:* Many projects use commercial relational DBMSs, specifically, Oracle. Virtually no projects use commercial object-oriented DBMSs (OO DBMSs). Why? OO DBMSs were developed to overcome limitations of relational DBMSs, and have now existed for about a decade. Little hard information about this issue is circulating in the bioinformatics community. We should obtain and circu-

SRI International, Menlo Park, California.

late such information. In my own experience, the Ocelot OO DBMS developed by my group has been an extremely useful productivity tool to us.

4. *New DBMS:* Some computer scientists might reasonably argue that existing DBMSs have significant limitations, and therefore a new DBMS must be constructed for molecular-biology applications. However, it is worth considering the likelihood of such a DBMS being adopted by bioinformaticians and molecular biologists, which I consider to be low.

In point 1, I noted that many biologists do not employ any DBMS when constructing a DB, partly out of ignorance of the advantages DBMSs bring, and partly because they do not know how to use them. This group of potential users is unlikely to appreciate the benefits of a new-generation DBMS, unless tremendous ease of use is one of those benefits.

In points 2 and 3, I noted that relational DBMSs are the most commonly used DBMSs in bioinformatics applications. In my opinion the reasons for this include (a) desire to use the prevalent technology, (b) relative ease of hiring DB administrators (DBAs) who know Oracle, (c) easy access to classes and books on relational DBMSs, (d) apparent simplicity of the relational model, and (e) ignorance of the benefits of using object-oriented DBMSs.

It is important to understand the reason for past choices by DBMS users before embarking on a new DBMS quest.

In my opinion, the DB area of bioinformatics is the area where current practice lags farthest behind the state of the art in computer science. "Database malpractice" is amazingly prevalent, by which I mean the frequency of egregious mistakes and amateurism in constructing biological DBs is much too high (examples include failure to use a DBMS for large DB projects, failure to use controlled vocabularies, and bad DB schema design). Thus, education may have more potential for improving practice in this field than research.

Address reprint requests to:
Dr. Peter D. Karp
SRI International, EK207
333 Ravenswood Avenue
Menlo Park, CA 94025

E-mail: pkarp@ai.sri.com