Towards Science worth doing

Richard K. Belew

** * * DRAFT: Not for attribution * * **
May 8, 2004
`sciWorthy.tex`

1 Introduction

Few believe any longer the myth that Science is neutral, that the uncovering of facts about the world can be separated from their consequences. But in scientific education, the myth lives on through neglect: today's scientist may be technically well-trained but is almost always ill-prepared to guide his or her work in morally appropriate directions. In the resulting vacuum, most scientists and their Science has been sucked along by the vagaries of governmental and industrial funding. To an outsider, direction by such political and economic forces may seem appropriate, but the practicing scientist knows better.

A central property of Science is that the universe it studies is expanding, literally! But even more so, it is expanding conceptually. Almost daily the practicing scientist finds that his or her original question needs to be refined into subsidiary questions that each require investigation. To be successful the scientist must become a skilled gardener, pruning away less important issues to concentrate limited resources on more important ones. A government grant or corporate research directive may specify a general line of inquiry, but within this mandate the range of intersting phenomena is still virtually limitless. The universe is expanding in the sense that the number of (very good) questions to ask will always dwarf the number of scientists available to ask them. It is here that the creative scientist is in need of some "moral compass." Given two new research questions, equal in scientific merit, which is most worthy of his or her — and society's — time?

Such determinations are difficult to make in part because modern Science is a tremendously *social* activity. The connection between a scientist's activities and their consequence generally involves a long chain of other scientists and engineers. Within this social mesh of colleagues, some suggesting new lines of research, some providing funding, some providing enabling tech-

nologies, how is the individual to maintain a personal research agenda?

While this paper is primarily a personal exploration of some of these questions, I believe it has been informed by two aspects of my technical work as a computer scientist. First, I spend a fair portion of my time studying the problem of "information retrieval" (IR), how people can find free-text documents that are "relevant" to a question they have and computer technologies that can facilitate this process. One obvious application of IR technology is for scientists looking for one another's papers in the technical literature. In fact, I have argued elsewhere that it would hard to envision modern science without the huge literature of scientific papers holding it together [**Belew-FOA**]. For these reasons, I have become particularly interested in how scientists communicate and retrieve information, and this is the second area of expertise on which this paper draws.

As a computer scientist with serious qualms about current status of the military-industrial complex in this country, I have felt a tension between the kinds of questions I thought were socially responsible to investigate and the research agenda established by the primary sources of computer science funding. But although I have lived with this tension since graduate school, I feel even less able to provide an absolutist's moral compass now than I did then.

Following are three pieces of unpublished writing i've harvested from three separate sources: First is a manuscript I worked on in the early 1990s, as I attempted to puzzle my way through several decisions I faced as a brand new faculty at UCSD. Next are notes on "The kept university" I wrote in response to the JSOE "Ethos Statement" in early 2002. Finally, I include passages from a (unfunded:(proposal for my sabbatical on the topic of "Open Source meets the Genome." later in 2002.

2 The Warm Body Metric

Direction of the scientific agenda by political and economic forces may seem appropriate, and scientists must respect these considerations along with the rest of society. But attention by scientists to the pragmatic considerations of what (governments and corporations think) it would be nice to know next is comprimises their primary mission: discovery of "knowledge for knowledge's sake." **[KitcherSTD]**

Discovering new features of the universe is the scientist's primary mandate, but the job is not done until this new fact is told to others and integrated into the collective corpus of scientific knowledge. The effort required to transform the scientist's personal understanding of a new-found result into language that is comprehensible within an established paradigm is often under-appreciated. And just as the scientist exercises choice in the selection of which questions to ask, so too there is considerable latitude in describing his or her result. To be general, suppose a scientist discovers "A causes B." Almost without exception, there will be a group of experts on A interested in this result, as well as an equally interested group of experts on B. The practicing scientist is again faced with a consequential choice. As will become clear below, the fact that the consequence of this "choice of language" decision is more easily observable — via examination of the scientist's printed record, makes it particularly important for our purposes here.

First as a graduate student and now as a practicing scientist and especially as a graduate advisor, I have personally attempted to wrestle with moral questions such as these. I happen to be a computer scientist, and the moral questions that have most occuppied me involve funding by the Department of Defence. But the argument presented here is intended to go beyond the specifics of computer science and DoD. The goal is to provide a framework within which any scientist can apply his or her own social priorities to the science that they do.

I have developed a heuristic that I call the Warm Body Metric (WBM). While far from perfect, the WBM has fairly consistently helped me sort through the complicated particulars of each new dilema to a conclusion with which I can feel comfortable.

Being an intelligent consumer in today's market is difficult, and sorting out scientific choices promises to be at least as complex. Simple heuristics, such as the Underwriters' Laboratory (UL) seal of approval, nutritional contents labels on food products, etc. are useful as a way of converting thoughtful, general policies into immediate buying decisions. The WBM is proposed

in a similar vein, as a heuristic that can be used to convert high-level ethical priorities into appropriate scientific decisions that must be made daily

It is worth re-emphasizing that the WBM can be decoupled from the particular political purposes towards which I personally have applied it. Science is an inherently pluralistic enterprise. It derives much of its value from allowing individual investigators the freedom of establishing personal research agendas. Any attempt to advocate a universally shared ethical program, or even a proscribed "science policy," faces real difficulties; in any case, I make no such absolutist claims here.

Certain questions have always been pretty easy. Reagan's Strategic Defense Initiative (SDI) from the beginning seemed to me and many others both politically destabalizing [**SDI**] and technically unsound [**Parnas**]. Most of the rest of the decisions that I have encountered, however, were far less clear. In fact, after a statement of the Warm Body Metric heuristic, the basic form of this paper is an ennumeration of real dilemas that I have bumped into during the first five years of my career as examples of the application of this heuristic.

2.1 True stories

Questions concerning what kind of science I wanted to do began to appear even in graduate school (see **??** below). For these, relying on internal, affective judgment was sufficient: Conceiving of the possibility that work I did might lead to military applications took all the fun out of doing science. So I just stayed clear of DoD funding reflexively.

But the frequency with which similar questions, and very subtle variants of them, arose certainly accelerated once I began a job as an assistant professor at a research- oriented state university. Worse, I now had to make decisions that involved others, primarily graduate students. They appropriately wanted me to explicate and rationalize the vague, personal introspections on which my early decisions had been made. Like many heuristics, the WBM grew out of a need for some analytic device that could help me think about the decisions that would shape my scientific career, and indirectly those of my students and colleagues.

Collected here are some of the most memorable of these questions. They are presented as examples of how the WBM can be applied in practice. For some, I believe the WBM works quite well; for others the account presented here may well be contrived. Because they are examples drawn from my personal experience, they necessarily embody my personal political posture. Most briefly, I want to develop new computing technologies that allow large

groups of people to build things together, and I also want to avoid DoD funding. Appendix **??** characterizes my political agenda in slightly more detail, as background. But once again, while the examples presented below grew out of my personal predilections and therefore reflect these specifics to some degree, other scientists with other agendas should be able to apply the WBM in similar fashion, simply substituting their own goals for my own.

2.2 The Warm Body Metric

The basic warm body metric argument is an attempt to resolve an apparent paradox arising from the interaction between two fundamental facts of the modern scientific enterprise. The first fact is that any contribution a scientist makes must make its way into a publically accessible open scientific literature. Scientific progress absolutely requires free and open debate. There is no doubt that a great deal of advanced technological development goes on under governmental classification or industrial propriety, but the inability of these laboratories to generate prolonged progress on fundamental questions make them merely exceptions that prove the rule.

Thus no scientist is allowed the luxury of communicating their result to only those recipients he or she would like to have benefit. Publication of results is mandatory if a scientist wishes to continue to do science. Once published, it is *theoretically possible* for anyone else to make use of these results as they choose, for good or ill. The fact that this theoretical possibility is far from a practical reality is key observation underlying the WBM.

The second fundamental fact of modern science is that it is an immediate goal for almost every scientist beginning their career is to secure extramural funding for their research.† Such funding generally is the only way to acquire necessary equipment, pay for travel to professional meetings, etc. In the university setting, it is assumed that external grants will pay a significant portion (generally around 25%) of the academic salary; unfunded academics make less money. External grants are also the primary way of financing graduate student education, in the form of "research assistanceships" by which a student is payed to help with the funded research. Ideally, the student's own thesis research is part of the faculty's funded project.

†*Get UC Faculty statement.*

2.3 Examples

2.3.1 Applying to the DoD for funding

The DoD has long been the dominant funder of computer science research. The most obvious choice for the beginning academic, therefore, is to prepare

a proposal describing the research they intend to do and send it to one of the various DoD granting agencies (e.g., DARPA, Office of Naval Research, Army Research ...)† The proposal is reviewed internally, sometimes with input from external reviewers, based on both scientific merit and its relevance to that agency's research goals.

†*Check names.*

I have developed a number of research proposals, but I've not sent any proposals to DoD agencies. Graduate students in my department who I can therefore *not* help fund (let alone my family!) have asked the very legitimate question:

> Why not take the very same research proposal you generated for some other funding agency, and simply send it to some DoD office as well? Don't change the research you want to do one wit. If somebody in DoD is willing to fund work you want to do anywaym, how can that be bad?

The basic warm body metric (WBM) argument is a response to just this question: By taking my work and *bringing it to the attention of* individuals working for the DoD, I have increased the chances my results will be used towards purposes I dislike.

If I do send my proposal to a DoD office, I am suggesting to them that the work I do may be of interest. If they fund the work, it is because (at least in part) this work is consistent with the military objectives they have established. Such a determination requires at least one "warm body," a well-trained (computer) scientist capable of understanding the work I propose and relating it to the agency's objectives.

Alternatively, imagine that I do not submit the proposal to DoD, and that the work gets funded elsewhere. As stipulated above, the results of this research must be published and therefore become available for DoD purposes anyway. On the face of it, I've done DoD a service: they got the same results and somebody else payed the bill!

But this bargain is only realized if we assume that the cost to the DoD of first finding and then applying my results to their problems is zero, and that these processes take no time. In fact, the scientific literature is enormous, and wading through all of it for useful (from the DoD's perspective) results must require a large staff of well-trained *warm bodies*. Sifting through potentially important contributions for those that best accomplish DoD goals must consume the time of another large set of technically-trained managers; more *warm bodies*. Hiring and retaining a technical staff of this magnitude, capable of watching all areas of scientific investigation, must be an extremely expensive proposition.†

†*Try to get some numbers on this. Check with Arbarbanel, Karin,...*

Most significantly, these DoD employees do not have to work alone: I have no choice but to help the DoD in their interpretive endeavor. After years of funding, imagine that some general or admiral comes to visit to try to understand what all of this scientific theorizing that his organization has payed for, and that some DoD scientist believes can help, has to do with the job he or she is trying to accomplish. How can I ethically withhold my best efforts to make these people understand what my results mean to them?

Then consider the time required under the two alternatives. If a proposal is sent to DoD, they are informed about your work and its potential before the research has even begun. DoD agencies are also particularly concerned with "progress reports" and "site visits" during the course of the grant (typically two or three years). If my work is concerns a technology they consider useful, the DoD is in as advantageous position to immediately begin the (admittedly considerable) task of taking my scientific theory and putting it into practice.

Alternatively, imagine I had sent my proposal to an other agency *that I consider more socially responsible*, perhaps an organization facilitating sustainable agriculture. If I have sent my proposal, provided regular progress reports, been site-visited by them, etc., it is this other agency that has the benefit of the head start. Given that the scientific publication process itself (including peer review, revision, publication backlogs and production) can take a year or more, it is not unrealistic to assume that the DoD will come to know of my results 2-5 years later than they would have if I had submitted the proposal to them directly. This gap may not seem like much, but in terms of the fast-breaking world of high-tech and their political realities, it can be all the difference in the world.

2.3.2 Other examples to consider

```
- DOD
-- 6.1 vs. 6.2 (basic vs. applied)

-- DARPA initiatives/ RFPs
--- Neural networks
--- Tipster
--- USe of IR techniques by NSA

- NASA/other DOD-"cognate" agenciesgrants
```

- Being included in another's DoD proposal

- Reviewing DoD proposals

- Presenting work to colleagues

- Collaborating with colleagues who are funded by DoD

- Advising the PhD of a student working for DoD

- Teaching a class with DoD employees in attendance

2.4 My personal political agenda for my science

My political priorities for the science that I want to do has both positive goals that I want to achieve and counter-productive activities that I want to avoid. Constructively, the reason that I became a computer scientist was, and still is, that I believe this technology provides a mechanism by which large groups of people can coordinate their activities to build things of unimaginable scope. Some of the benefits of such coordination have been available to large organizations (governments, corporations) for some time. The increased power and availability brought on with personal computers promises to radically transform how we interact as a species. Theoretically, I would like to understand just what kind of transformation this will be. More practically, I am concerned with developing mechanisms that facilitate such interaction, and do whatever I can to make sure that this technology does not dimish our humanity in the process.

Ideally, these positive goals would provide sufficient direction to guide the choice of what kinds of science I do. But the reality of modern science (at least as experienced by junior scientists such as myself) is that the majority of such decisions are externally initiated, by funding agencies, for example by their "requests for proposals." In computer science, far and away the largest funder is the Department of Defence. It is therefore no coincidence that almost all of the decisions discussed in **??** relate to DoD funding, in one manner or another.

My political sensibilities were formed during an era (Vietnam) in which aspects of the government, particularly its military forces, did not seem to be reflect the will of the people, at least those with which I associated. I came to distrust the Department of Defence (DoD), and doubt that all of

its activities were truly concerned with the defence of our nation.

More to the modern, post-Cold War point, the DoD appears to me to be an institution solving a problem that no longer exists. The almost perennial ritual of building extraordinarily expensive weapons of mass distruction, burying them in the desert, and then declaring them obsolete so that we can begin the cycle again is surreal. The only explanation of such continued irrationality that makes any sense to me is the one General Eisenhower cautioned against nearly 50 years ago: the "military-industrial complex" has become an extremely powerful institution most concerned with its own perpetuation. As we enter an era of increasing populations, limiting resources and global economic competition, DoD and the symbiotic industry that has grown up around it reflect an illusion the rest of us can no longer afford. Thus the central plank of my current political agenda for the science that I do is to do nothing that contributes to the perpetuation of this complex.

I also have misgivings about high-technology medical applications (as opposed to low-tech, public health applications of the same dollars), the need to explore space with so many unresolved problem here on earth, and the capitalistic vision of progress generally. But because DoD funding of computer science research has become so pervasive[1] that simply avoiding this funding has limited my options dramatically. I also have more confidence that these other issues (health care, space exploration, capitalism) can be resolved democratically; these other high-technology industries may even play a useful role as part of an economic conversion effort, away from defence industries. Consequently, these misgivings play a more minor role in my agenda.

But what about cyberwarfare defence?! A very hard one, for me. ALife techniques very relevant, and it does feel more like civil defense.

But how about big pharma? Orphan drugs, vast expense of drugs also a reflection of capitalism run amock? Maybe, but on my ordering, this is

[1] Mention stats here.

3 The kept university

Written approx January, 2003 in response to the JSOE Ethos report, written in Fall, 2002. I think; the statement seems no longer available anywhere at JSOE?!

3.1 The common good

Academic prestige is built primarily around fame-at-a-distance, the opinions of others at peer institutions. But this focus on *external* validity often comes at the expense of respect *internal* to this campus. (The Ethos Statement seems to consider the former part of a JSOE "mission," in contrast to the latter as its "ethos." I do not find the distinction helpful.) Institutions have many demands that require careful, time-consuming, unglamorous work that does nothing to attract external job offers. But these jobs must be done well or they diminish the work all of us can accomplish. The deeper, more considered opinions of people with whom we have worked worked on such chores, as well as grantsmanship let alone research, needs to be at least as consequential as more superficial evaluations done by those far away.

The focus on our campus as the *physical location* of our faculty's primary activity becomes more important as we are all involved in "virtual work" with others at other locations. The benefits of physically sharing our classrooms, laboratories and seminars to sit and talk with one another are great, and we should make the most of the proximate relationships we have here.

3.2 Remaining an open forum

Concern with the dissemination of knowledge within the university, and from the university into the rest of our society comes close to the heart of the issue for me. The level of constant, silly concern regarding the transfer of IP is interfering with our primary purpose.

No one can doubt the place of an economic agenda, indeed its dominance in modern life. But it is not the only agenda, and universities are charged with the responsibility of exploring beyond our most pragmatic concerns. We should focus on long-range, speculative inquiry because by definition it is meant to complement development that has a clear economic motivation. At least in engineering, faculty have a clear choice between commerce and the academy; if our goal was to pursue an economic agenda, we would have taken different jobs. My argument is that by adopting a single-minded economic metric, the university has been diminished in the process. To follow the

economic rather than intellectual agenda is to abdicate the university's role in society. (Consider the broad range of radically new questions that have become important in the wake of Sept. 11, 2001, and how grateful we are that some academic somewhere has actually been studying this esoteric topic for years!)

The discussion has become so confused that the Ethos Statement worries about only "... UNDUE intrusion of caveats and embargoes on [a student's] ability to publish and disseminate research." I assert that within the context of the University, students should not be worrying about these constraints whatsoever. Our goal here should be the educational preparation of the students and production of public knowledge. Any and all such constraints on the flow of IP are the appropriate concern of private industry, and should be handled as constraints on behaviors of employees and consultants within corporate institutions.

3.3 Research and teaching

My favorite phrase is "... appreciation of the *dependence* between [teaching and advancement of research]." I believe addressing how educational and research activities really do and might synergistically feed into one another would make an excellent focus for further deliberation.

For me the concern with education is directly related to issues of shared knowledge. Where does the knowledge we discover in research (potentially proprietary and subject to disclosure as invention, etc.) end, and the knowledge we are supposed to teach and share in our advanced classes begin? Should we be asking these students to sign NDAs?!

3.4 JSOE as part of UCSD

There is encouragement in the latter for departments to develop their own statement and I look forward to reading CSE's. But I see no discussion of how the JSOE and its ethos relates to those of our campus and university? [The preamble references some UCOP statement??] Reaction to my resignation letter from other parts of the UCSD community makes it seem that the concerns I have experienced, while not unique within JSOE, seem especially pronounced here. (Also consider the recent statements by our colleagues in the social sciences, Michael Bernstein concerning "Institutional Growth," and the earlier exchange between Steve Shapin and Chancellor Dynes in the UCSD weekly paper last summer. Even the articles on UCSD and CalIT2 appearing in the San Diego Reader in the last year can give us a sense of

how we can be perceived by the surrounding community.)

When I originally interviewed for jobs, one of my most important criteria was the *breadth* of the institution. I was looking for a great university to help integrate my discipline with others. If I had been interested in a strictly engineering school I would have gone to Illinois.

The questions of how CSE, JSOE and UCSD all relate are not just abstract. For example, how should we to approach *interdisciplinary* research with other units on campus and on the mesa, as virtually all funding agencies are recommending? More broadly, how can we even consider the prospect of a *School* of Computer Science without having opinions about how our discipline interacts within JSOE and with the rest of the university?

3.5 Words and Actions

I very much look forward to the construction of the Web site and bulletin board discussed as part of ""promulgation" of the JSOE ethos. I believe this mechanism has real potential as a virtual Ethos Ombudsman, especially if staff and especially students can be made to trust it and particate. But beyond this mechanism, I see very few specific proposals for *action* that I could expect to correct the concerns identified.

The central problem I see with this document is that it sees its goal as the characterization of *bounds* on appropriate behaviors, rather than as real changes to the *reward systems* influencing us all. Prohibitions may be useful for dealing with aggregious outliers, but if we are to change the behavior of the mode we must address forces that shape our regular behaviors.

3.6 Anecdotes

Example: about two years ago I was unable to convince the university technology transfer lawyers that an open source (GPL) license concerning code developed under a grant from Encyclopedia Britannica was appropriate. EB and its lawyers were convinced, and it was UCSD staff who thought this would compromise some potential economic gain!

4 Open source meets genomics

"Open source" (OpenSource) software development refers to the practice of making the human-readable source code of a program publicly available, together with a license that attempts to insure that "participants" (user/codevelopers) will retain access to their mutual work. OpenSource has, within just the last few years, moved from the anachronistic activity of a few fringe computer technologists to a widespread activity engaging large numbers of programmers in an economically viable alternative. Coincidently, within this same period the Human Genome Project and related "bioinformatic" activities have begun to realize the fruits of these labors. Enormous data sets describe, in remarkable detail, genes and proteins shared by all life on this planet. Computational models of gene networks, protein pathways and cellular function are just beginning to connect these pieces of the puzzle to one another.

This project proposes to investigate the confluence of these two major dynamics, as tools for the burgeoning bioinformatic community using the mechanism of OpenSource . Specific questions include:

- How is OpenSource development of genome-related software affecting access to this data?
- How do sociological dynamics surrounding OpenSource development affect participation by computer scientists and biologists?
- How does the OpenSource foundation affect the specification of the software tools used by the CompBio community?
- What opportunities and obstacles does OpenSource introduce to the flow of intellectual properties arising from genomic data?

The central premise of this work is that the extremely active work in computational biology provides a special opportunity to watch the role of OpenSource development. An especially broad and interdisciplinary group of investigators, many of whom are still unknown to one another, have an increasing large stake in building and sharing computational tools on which they can all depend.

OpenSource for computational biology makes excellent sense for at least three reasons:

1. OpenSource will bring to bear peer review-like social dynamics, resulting in the production of higher quality software;

2. OpenSource makes especially good sense for a project dealing with a data in public trust, like the Human Genome and related resources;

3. it generates a shared infrastructure that has collateral benefit, for example towards the education of biologists trying to understand computational concepts (and vice versa).

These are strong claims, but of significant consequence if true. This project will provide a preliminary evaluation of these issues. The "Digital society and technologies" program within IIS.CISE.NSF.GOV is an appropriate funder of this work.

4.1 Peer review and OpenSource quality

"Given enough eyes, all bugs are shallow." ***Eric S. Raymond "The Cathedral and The Bazaar "*** This early mantra of the OpenSource movement captures one of the key advantages of allowing other people, beyond the original development team, have access to source code. http://www.tuxedo.org/ esr/writings bazaar/

Since these early days, the argument has become more than talk. the importance of OpenSource as a quality-control mechanism is especially well-appreciated by the computer security community. Consider:

> "... publishing source code for independent peer review increases the likelihood that security flaws will be identified and fixed in a timely manner." ***Bud Rogers, SANS FAQList, 21 Jan 01*** http://www.sans.org/infosecFAQ/co

> In critical applications, an enormous amount of untrustworthy code may have to be taken on faith.... Open-source software offers an opportunity to surmount these risks of proprietary software." [**neumann98**]

Even NSA, our government's most security-connscious agency, is actively promoting ***SELinx*** a "security enhanced" version of the popular OpenSource operating system. http://www.nsa.gov/selinux/

> Linux was chosen as the platform for the work because of its growing success and open development environment... A Linux platform also offers an excellent opportunity for this work to receive the widest possible review and perhaps provide the foundation for additional security research by others.

4.1.1 BLAST

Within bioinformatics there can be no better example of a successfully developed facility than BLAST [**karlin90,altschul90,altschul96**]. Over approximately six year, this program developed from an influential algorithm to one of the most useful tools in biology. It is provactive to consider what might have happened if this tool had been locked up by proprietary concerns; could the "lost opportunity" costs ever have been imagined!

Fortunately, this possibility needs only be considered as a *gendanken* experiment. What we can know is that "many eyes" did indeed improve this code. Consider these excerpts from the bug fix list associated with BLAST 1.4 (the last public domain version):

```
9/26/95
Fixed a long-standing bug in pressdb regarding which sequences are
tagged as having "ambiguous" nucleotide codes.  Thanks to Colin
Watanabe at Genentech for pointing this out.

2/1/95
fixed a bug in BLASTN's
calculation of the Karlin-Altschul K value.

6/22/93
Fixed bug in consistp.c's implementation of R(i,3) found by Phil
Green.

6/21/93
Fixed bug in the calculation of "consistent N counts" for those HSPs
found on minus strands in BLASTN, BLASTX, and TBLASTN.
```

For more than a decade Warren Gish and colleagues coordinated logical fixes like these with many other refinements, ports to new environments, input and out formats, etc.

4.2 Public and private access to data

Public and private sequencing efforts have now provided complete or nearly-complete sequences for scores of organisms, from yeast, worm and fly to mouse and human. While many investigators continue to improve the quality of genome sequencing and extend it to other species, others are highlighting EST (expressed sequence tags) found, identifying variations found

in the population like SNPs (single nucleotide polymorphism), cataloging structural and functional properties of key proteins, and capturing protein expression data from micro-arrays.

The result is a dizzying array of enormous, diverse data sets, produced by investigators from a range of disciplines for wildly different purposes. But even this staggering volume of "raw" data is only the tip of an iceberg which is already beginning to build from this foundation. "Annotation" refers to the process of attaching interpretative data, from gene names to cross-species homologies to important bits of sequence data. Because making sense of this data in terms of meaningful biological knowledge and hypotheses is critical before any of this new data can become useful, many labs are proving such annotations. Other analytic tools (e.g., for gene finding or protein expression clustering), derived statistical features, computational models (e.g., of regulatory pathways or cell function) can all be expected to grow into important data types, and profitable resources, in the very near future.

The recent, well-publicized race, fight, and then resolution between public and private attempts to sequence the human genome is an interesting beginning. The circumstances surrounding the official "publication" of the human genome required special publication arrangements for this data **[Science-0012??]** demonstrate just how confused lines between science and commerce have already become. Now that the hatchet has been officially buried, the alternatives will be distinguished along more subtle product differentiations.

As biology is coming to depend more and more on sophisticated computational tools, the intellectual property issuess are only becoming more complex. Consider:

- The ***OpenInformatics Initiative*** which is petitioning U.S. government funding agencies to require grantees to publish code developed under government contract. http://www.openinformatics.org

- The issue is especially acute within academic institutions. Since the 1980 Bayh-Dole Act, which allows universities and researchers to patent the results of publicly financed research, academic institutions have come to increasingly depend on licensing of research-related patents. Current estimates are that universities file patent 2,000 annually, with the resulting fees accounting for 10% of their budgets. Phil Green, author of the widel-used PHRAP, PHRED and CrossMatch assembly tools, disagrees. His tools have made money (for the University of Washington). " Open source publishing devalues what [programmers]

do.... I don't think computer programmers should be treated any differently than other scientists. It sort of diminishes the stature of the science." ***SiliconValley.com, 24 Nov 01*** . http://www.siliconvalley.com/docs/n

- Yet the University of California was just convinced to allow at least one investigator (Steve Brenner, Berkekley) to freely distribute his research product, software.

- ***ISCB Statement on Bioinformatics Software Availability*** http://www.iscb.org/pr.shtml#softw

4.3 Pedagogical uses for OpenSource

A key, rate-limiting step in the growth of this field is the availability of well-educated programming biologists. Because of the large and interdisciplinary gap between biological background and computer science background and, students generally fall into two quite distinct and modes: biologists seeking computational knowledge and computer scientists seeking biological knowledge.

We argue that *tutorials* are often a critical missing component, separating useful but underutilized code from broadly successful OpenSource examples. Towards this purpose several tutorials have been developed.

Executable versions of code are immediately useful tools to biologists. Quickly however many slight modifications suggest themselves, and they become interested in how to *perturb* the code to make it do something new. questions as to how it works in general come to the surface next.

Literate, pedagogically well-structured OpenSource programming examples can play an especially critical role for biologists trying to learn to program. Well-designed, well-described software lets biologists see concepts they know in terms of algorithms and data structures there are trying to understand.

The key advantage of "literate" programming methodologies, going back to Knuth's seminal TEXand **MetaFont** projects, is that code that is comprehensible becomes that much more useful. The Javadoc documentation generation system associated with the Java programming language, is by far the most broadly distributed and used example of a literate-programming documentation tool. [2]

[2]It is at least curious to note to that while Microsoft has developed many advanced development tools, support for literate programming facilities has not become as integrated as it has within Java.

4.4 Digital society goals

OpenSource offers key advantage to individual programmers, groups of biologists and other software users within conventional, bricks-and-mortar institutions (corporations, universities, government) and also development groups that become "virtual" institutions. These various social groups, and the effect of OpenSource technolgies on them, are precisely the object of inquery. "Sustainability" of a developed code describes features which guarantee its utility into the future.

Just as does e-mail created communication channels through organizations disrupted some standard business practices and reinforced others, the development of OpenSource across organizations is bound and to have similar effects.

The project is interdisciplinary, in two distinct respects. Most obviously, there is the mixing of computational and biological expertise. But this project also attempts to be interdisciplinary between software policy and practice. A central deliverable will be an objective assessment of current practice in and around by and dramatic OpenSource facilities.

Special types of software are already proving useful for users in bioinformatics. For example, certain graphics packages have been developed precisely to manipulate those classes of images used within the compuational biology.

Characterization of these search activities in terms of autonomous process these can migrate from host to host are now becoming possible.

4.5 Related work

4.5.1 OpenSource Background

The central, shared activity for all computer programmers is to write algorithms for accomplishing some computational task in some programming language, for example Basic, C, Fortran, or Java. Especially when software development involves large groups of people, the fact that people can read and understand the "source code" version of the software is critical. Special programs called "compilers" convert the source code these people have written into "object code" that can be efficiently executed on some computer hardware, for example the Intel Pentium II. Once a program has been compiled into object code, only this "binary" representation needs to be available (e.., distributed by floppy disk, CDROM or over the Internet) in order necessary for a computer to perform the required task.

"Open source" (OpenSource) software development refers to the practice of making the human-readable source code of a program publicly available, together with a license that attempts to insure that "participants" (user/codevelopers) will retain access to their mutual work. This mechanism has been typical within academic computer science since its beginnings. Historically, however, almost all commercial software has been distributed exclusively in binary form. The assumption has been that a company's exclusive knowledge of and control over source code provided a fundamental competitive advantage over competitors attempting to provide similar computational tools.

As software originally developed within the computer science research community has been "spun off" to form the basis for commercial activity, there has been a recurrent tension between OpenSource and close-source software development efforts. Very recently, however, a number of companies, including industry leaders like IBM and Netscape, have embraced OpenSource for distribution of software they have developed.

4.5.2 The hackers' perspective

The number of fundamental biological questions becoming accessible to computational analysis is quite extraordinary. But the range of these questions and their dependence on particular features of data sets, organisms, methodologies and techniques, require familiarity with the biological specifics. To date this has meant that individual biologists be able to develop software themselves. The lack of biologists able to develop these computational tools, and the ignorance of biological details by computer scientists, is therefore one of the process-limiting steps of modern biology.

Yet the economic factors just mentioned can mean that scientists and software developers who wish to make CompBio the focus of their professional efforts need to ensure that their reputation is not effectively indentured to whatever pharmaceutical company they happen to be employed by at the moment. Scientific publication has often served this role, with external publication providing validation of a scientist's reputation, independent of their place of employment. For this reason also, programming has much in common with science.

Software and utilities supporting the use and analysis of these data sets is especially important to have available and consistent across the entire community. Much of this is data-driven science, and depends on agreement upon standardized formats. Many XML-based standards are now in the air, such as GeneXML, Microarray Markup Language (MAML) and MIAMI. Beyond

these standards, centralized software catalogs such as the ***EBI Bio Software Catalog*** and ***Genamics SoftwareSeek*** are helping developers find and reuse one anothers' tools. Shared "distributed annotation services" and "ontologies" of key biological concepts are beginning to emerge in efforts to coordinate investigators' efforts. Should practicing programmers use them? How can they work otherwise?

http://www.ebi.ac.uk/biocat/biocat.
http://www.genamics.com/software/

D. Gilbert, a longtime participant in CompBio software developmeent, characterizes some of the trade-offs experienced by biologists attempting to choose between free and commercially available software tools:

> Free data analysis software is common in the sciences, as the scientists in need of new analyses develop algorithms for it, then crystallize the algorithms as software. Most of the basic biosequence analyses are scientist-developed, including FastA, BLAST, Clustal, MFOLD, PHYLIP, Paup, CAP, to name a few. The source code of these is often shared freely. But often these algorithm programs lack ease of use and integration with other functions. Commercial software developers have incorporated such algorithms, along with their own, and added a much greater usability and integration, to allow you to analyze your data without spending oodles of time learning how to run the programs....
>
> Today with the rapid growth in the fields of bioinformatics and biocomputing, more good programmers are developing software, many making it freely available. You will find more sophistication in attention to user interfaces and ease of your learning to use the software. Still there is no common means for funding development of free software: government agencies do not generally fund projects from individual research/programmers, the main source of many of the free software packages. The shareware concept of users paying for software they use has never worked well. For most scientific software, the potential market is so small that we see a strong distinction between expensive commercial packages and free software.
>
> ***D. Gilbert, 1998, "Free Software in Molecular Biology for Macintosh and MS Windows computers"***

ftp://iubio.bio.indiana.edu/molbio/I

4.5.3 PERL or JAVA or CORBA or PYTHON or ...?!

A subsidiary hypothesis concerns the role of programmingg language choice. The contrast will be with Perl and CORBA. It is hypothesized that Java

will provide a more robust foundation for OpenSource development then either PERL or CORBA, but two different arguments distinguish it with respect to each. PERL is arguably still the dominant language used within CompBio. It was one of the first languages to provide quick, interpreted shall scripting facilities like the most useful for munging one data format into another. PERL's CPAN (Comprehensive Perl Archive Network) repository is a central toolchest for many.

But PERL is not inherently object-oriented , and OO features are often mentioned as important to successful OpenSource development efforts [**??**]. More modern implementations like PERL5 do provide object oriented support, and Python is another OO scripting language that many consider an advanced version of Perl.

Java and its "virtual machine" architecture provides other advantages that operating-system dependent scripting languages cannot provide. It does not support development across multiple hosts as easily as Java. Many security mechanisms that may prove useful in Java are similarly not available in basic PERL or Python.

BioJava is an especially interesting example for several reasons. First, the original (Perl) scripts hacked together for standard data exchange formats and presentation formats have been rewritten into pedagogically superior Java examples). More important, these basic utilities have been been extended to include some of the most mathematically sophisticated analyses. Durbin et al. have written the standard textbook, showing how dynamic programming over Hidden Markov Models (HMM) provides very useful tools for sequence analysis [**Durbin98**]. HMMs were originally developed for speech recognition, and have been more recently applied to text analysis [**McCallum**]. The broad utility of these basic mathematics and computational tools makes then an especially worthwhile educational goal. It is for this reason that HMMs generally and the `biojava.dp` package in particular are the focus of my early tutorial/class material development; cf. Section **??**.

CORBA is designed for a much more rigid interaction between organizations. That is, deontic (rights and obligations) commitments figure more prominently in the relationships between calling and called routines. There is a BioCorba organization, the EBI has in the past been especially active in promoting this standard, and NetGenics.com, in particular has promoted the use of CORBA-based infrastructure. Despite these efforts, however, CORBA development seems not to have kept pace.

4.5.4 Supporting software

An interesting argument *against* the use of OpenSource for bioinformatics focuses on how difficult software is to understand. G. Ropella thinks this is not a bug but a feature. He points to a division of labor he recommends, between computer professionals and others users:

> [S]ure the linux kernel is "open source"... but, one has to be an operating systems expert to understand it. And this encourages non-operating systems people to become experts in operating systems, thereby removing their attention to other things... thereby reducing the efficiency of the whole organism.
>
> It encourages Bad Behavior by relieving the experts of their interfacing responsibilities. *(personal communication)*

This perspective highlights two of the basic activities surrounding OpenSource :

- users attempting to understand software; and
- "interfacing responsibilities" of the software developer: all those things a software developer must do to help make code understandable.

In a corporation selling a software product, support questions must be handled by paid employees. These people generally do not know how to program, although they often work hand-in-hand with a technical staff who does have direct contact with program code. Their task is, in most situations, simply to route user queries to the most relevant piece of extant documentation. OpenSource code developers, on the other hand, are almost always responsible for their own support and function.

New tools, such as ***Wiki*** have become one of the major documentation tools

http://c2.com/cgi/wiki

4.6 Policy aspects of OpenSource

Lawrence Lessig [**Lessig99**] has recently highlighted the potential role OpenSource methods can have in the regulation of "code" defining cyberspace.

> What makes a system open source is a commitment among its developers to keep its core code public - to keep the hood of the car unlocked. That commitment is not just a wish. Stallman encoded it in a contract that sets the terms that control the

future use of much open source software. This is the Free Software Foundation's general public license (GPL), which requires that any code licensed with GPL (as Linux is) keep its source free. GNU/Linux was developed by an extraordinary colllection of hackers worldwide only because its code was open for others to work on.

Its code, in other words, sits in the commons. Anyone can take it and use it as he wishes. Anyone can take it and come to understand how it works. The code of GNU/Linux is like a research program whose results are always published for others to see. Everything is public; anyone, without having to seek the permission of anyone else, may join the project. [p. 105]

Say you are a Soviet propagandist, and you want to get people to read lots of information about Papa Stalin. So you declare that every book published in the Soviet Union must have a chapter devoted to Stalin. How likely is it that such books will actually affect what people read?

Books are open source software: they hide nothing; they reveal thel they are their source! A user or adopter of a book always has the choice to read only the chapters she wants. If it is a book on electronics, then the reader can choose not to read the chapter on Stalin. There is very little the state can do to modify the reader's power in this respect.

The same idea liberates open source code. The government's rules are rules only to the extent that they impose restrictions that adopters would want. The government may coordinate standards (like "drive on the right"), but it certainly cannot impose standards that constrain users in ways they do not want to be constrained. This architecture, then, is an important check on the government's regulatory power. Open code means open control - there is control, but the user is aware of it.

Closed code functions differently. With closed code, users cannot easily modify the control that the code comes packaged with. Hackers and very sophisticated programmers may be able to do so, but most users would not know which parts were required and which parts were not. Or more precisely, users would not be able to see the parts required and the parts not required because the source code does not come bundled with closed code. Closed code is the propagandist's best strategy - not a separate

> chapter that the user can ignore, but a persistent and unrecognized influ- ence that tilts the story in the direction the propagandist wants. [p. 107]

It is not surprising that Lessig (a lawyer, not a programmer) treats the distinction between "hackers and very sophisticated programmers" and "most users" as qualitative. But as managers and educators of computer professionals know well, there is in fact nearly-continuous variation between those capable of hacking/dissembling/obfuscating programming code and those who simply consume it. In addition, there are differences both in how well-structured, documented and understandable OpenSource code is, and wide variation in programmers' abilities to read others' code, open or within a corporate team. Further, most of these same conditions and dimensions of variation apply as well to data exchange formats as they do to executable code per se.

The key contribution Lessig makes is in connecting these technical features to their political consequences:

> ...open codes reduces the reward from burying regulation in the hidden spaces of code. It functions as a kind of Freedom of Information Act for network regulation. As with ordinary law, open code requires that lawmaking be public, and thus that lawmaking be transparent. In a sense that George Soros should understand, open code is a foundation to an open society. [p. 108]

While debate surrounding OpenSource is now primarily focused in the U.S., the global reach of these virtual, Net-mediate activities is guaranteed to raise similar discussion elsewhere. For example, a French proposal has been introduced concerning ***"Open Standards & Source Code Access"*** has recently been introduced by legislators. http://www.osslaw.org/motifs_en.htı

4.7 OpenSource for CompBio

The central premise of the proposed work is that the extremely active work in computational biology provides a special opportunity to watch the role of OpenSource development. An especially broad and interdisciplinary group of investigators, many of whom are still unknown to one another, have an increasing large stake in building and sharing computational tools on which they can all depend.

The CompBio community has already developed a number of shared software utilities for managing and exchanging standard data formats; these are exemplary in both in their utility to the community and their software engineering.

4.8 OpenSource economics

Science can be modeled as a sort of economy, with each scientist "producing" results that other scientists and engineers require, and simultaneously "consuming" others' results. While it is probably no more possible for any one scientist to change the course of science than it is possible for any one consumer to shape an economy, the thoughtful and consistent actions of large groups of scientists exercising their personal ethical priorities is a market force to be reckoned with.

Economists are only just beginning to make sense of the OpenSource movement. From a macro-economic point of view, Shirkey [**Shirkey00**] has characterized OpenSource within the biotech community, as a donut:

> Like companies in all information-driven industries, biotech companies thrive in environments where there exists balance between limited and free access to information. In an environment where access to information is controlled, companies can profit from investments in R&D or from their own pioneering innovations. However, the downside of control is constriction - too much control prevents the market from expanding, causing it to stagnate.
>
> On the other hand, the free flow of information helps create a level playing field, ensures low barriers to entry, and encourages innovation from all quarters, which in turn keeps the overall market vibrant. What we've learned from the economy created by the Internet over the last half-decade is that we can strike the best sort of balance between control and freedom by forming a "donut" economy.
>
> In a donut economy no one owns the intellectual property, which rests in the center. Value lies in the ring of innovation wrapped around this core. [**Shirkey00**]

Adopting a micro-economic view, from the perspective of programmers p articipating in OpenSource development, Lerner & Tirloi use the notion of a "signaling incentive" experienced by participating programmers [**LernerTiroli**]. This factor is meant to include such issues as: "career concerns

(future job offers, shares in commercial OpenSource -based companies), and ego gratification (desire for peer recognition)." A central feature of the Lerner & Tirloi analysis is "strategic complementarity": To have an audience, programmers will want to work on software projects that will attract a large number of other programmers."

But to practicing programmers, raw number of other participants is not the most salient aspect. In fact, contact with and guidance by a few experienced "wizards" counts for much more, and the "bandwagon" stage when "newbies" come into a project is generally to be dreaded.

4.9 Modeling hypothetical interactions

Simulation experiments of OpenSource dynamics are only just beginning. Khalak has explored conditions in which commercial software might still be profitable when OpenSource software of equivalent quality is available without cost [**Khalak99**].

The central theoretical organization I will use is "distributed cognition:" software design is treated as a *local activity* among interarcting agents who communicate through a shared space of representations. [[]Fig 8.4]hutchins95.

In contrast formal *economic* theories have typically been most successful as models of large systems of participants near equilibrium conditions. the fact that the participants (programmers, commercial developers, software users) and the various markets connecting them are in huge flux means that our understanding of OpenSource economics demands individual-based simulation techniques.This general class of techniques has been called Agent-Based Computational Economics [**Tesfatsion**].

The simulations of most interest to this project focus on decision-making of the programmers:

- when is it worth buying software;
- when is it worth learning how an OpenSource software package is constructed (vs. reading closed source documentation);
- when is it worth developing software yourself;
- when does it make sense to contribute developed code back into an OpenSource project?

4.10 Broader impact

This project should provide important new data and analysis concerning OpenSource development. The proposed visits should also serve a useful "cross-fertilization" role, bringing insights gleaned from one lab (tools, methods, emerging standards) to the attention of others. My software development efforts should help further a number of ongoing OpenSource projects. The proposed documentation and tutorials should also have significant impact on the accessibility of existing packages. The entire project is motivated by the promise that increased understanding of and participation in OpenSource development efforts can dramatically broaden the impact of bioinformatic tools within the shared activity of shared computational biology.

4.11 References

Altschul, SF, and W Gish (1996). Local alignment statistics. ed. R. Doolittle. Methods in Enzymology 266:460-80.

Altschul, SF, Gish, W, Miller, W, Myers, EW, and DJ Lipman (1990). Basic local alignment search tool. J. of Mol. Biol. 215:403-10.

Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, and DJ Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-402.

Gish, W, and DJ States (1993). Identification of protein coding regions by database similarity search. Nature Genetics 3:266-72.

Karlin, S, and SF Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87:2264-68.

Karlin, S, and SF Altschul (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. 90:5873-7.

Peter G.Neumann,
Robust Open-Source Software", Communications of the ACM, 41:2 (February, 1998): 128