

**\*\*Very preliminary draft\*\*** please do not cite or quote without permission

## “Open and Collaborative” Biomedical Research: Promise and Perils

Against the backdrop of an increasingly proprietary<sup>1</sup> and secretive<sup>2</sup> biomedical research arena we have recently begun to see some resistance. Public funding bodies, and some of the scientists supported by those bodies, are increasingly asserting the importance of openness. Significantly, some of these strong assertions regarding what might be called open and collaborative science move beyond what was historically the more uneven openness and collaboration of traditional biological science. Indeed, these assertions also move beyond more generalized calls for access to data and biomedical research tools.

Perhaps not surprisingly, the rise of arguments for open and collaborative biomedical research has overlapped with the well-documented emergence of “open source” methods of innovation in other arenas of research and development, primarily software. Indeed, in some recent cases, the modeling on open source software has been quite explicit – the federally funded haplotype mapping project, which aims to create a database that catalogues human genetic variation, has adopted a licensing policy that is self-consciously modeled on the “copyleft” system of open source software licensing.

This paper examines the promise and limits of the open source model for biomedical research. Because biomedical research does not always, or even generally, involve software and source code, the paper uses the term open and collaborative

---

<sup>1</sup> Cite to Walsh et al., Rai & Eisenberg

<sup>2</sup> Campbell, Gruschow (discussing empirical findings of increased secrecy associated with patents). Interestingly, this increase in secrecy may be occurring even when it is not explicitly associated with patents. See, e.g., Campbell (finding that increased competition may be one reason for increases in denials of access to research results); Gruschow (secrecy increased in period between 1980 to 1990, even for scientists who were not seeking patents).

research to encompass open source-type modes of investigation where information and research materials are generated, shared, and improved upon collaboratively without the usual sorts of exclusionary property rights. Although a few scholars have mentioned the possible extension of open source principles to biomedical research,<sup>3</sup> their analysis has not been informed by detailed examination of existing projects. This paper's analysis incorporates information gathered in interviews conducted with key scientists at the major existing open source biomedical research projects. Additionally, because these projects will typically be conducted in the university context, the analysis also incorporates interviews with university officials who have primary responsibility for making intellectual property decisions about university research.

An open and collaborative approach to biomedical research has considerable promise. As the paper notes, the model has been invoked in – and is likely to be quite useful in – three interrelated contexts: first, bioinformatics software; second, databases of biological information; and third, certain areas of “systems” biology, where many different sorts of expertise will be necessary to solve highly complex biological problems. Outside of these contexts, however, the benefits of an open and collaborative approach may be outweighed by certain disadvantages, including reduced incentives for commercialization. Indeed, even within the context of biological databases and systems biology, versions of an open and collaborative approach that rely on the copyleft version of open-source licensing may undermine commercialization.

In addition, unlike most open source software development, open and collaborative biomedical research will likely require some level of public funding: while the open source software model relies on services than intellectual property to generate

---

<sup>3</sup> Cite to Burk, Merges (?)

revenues, the extent to which open and collaborative biomedical research can rely on a services model is unclear. Finally, to the extent that pioneering scientific research has historically been driven by competition for kudos, approaches to open source that incorporate some role for competition and kudos may be preferable.

Other concerns about the open and collaborative model are more pragmatic: unlike open source software development, which largely takes place outside the university, open and collaborative biomedical research is emerging, and will continue to emerge, primarily in a university/academic setting. Accordingly, the paper examines the positions towards biomedical research licensing and towards open source software development taken by a select group of universities that have been actively engaged in one or more of the following three categories of activity: software patenting; biomedical patenting; or federally sponsored biomedical research.

Part I of this paper outlines various institutional structures for innovation, discusses their descriptive accuracy and normative appeal (or lack thereof), and then situates the open source model in the context of these institutional structures. Part II discusses the empirical methodology employed by this project. Part III discusses the empirical results of the project. Part IV then uses these results as well as the theoretical literature to elucidate the extent to which the model of open and collaborative research can be applied to biomedical research. It identifies advantages of the open and collaborative model as well as both fundamental and pragmatic difficulties that such efforts are likely to face.

Part I: The Open Source Model in an Institutional Context

### A. Institutional Models of Innovation (Other than Open Source)

The open source model can be best understood, and evaluated, by situating it within the different institutional models of innovation extant in the literature. In general, the three most prominent institutional models for organizing production are the firm, the market, and non-legal private ordering.<sup>4</sup> Firms tend to form when the costs of transacting in the market are high. Although firms lower transaction costs, they can also reduce the operation of high-powered incentives. Non-legal private ordering tends to operate in contexts where the number of relevant players is small, and these players are well known to each other. The social connection between the players lowers the cost of transacting without resort to the hierarchical model of the firm. To be sure, the simplicity of these institutional models belies the extent to which actual production in a given context may include elements of firms, markets, and non-legal ordering. But these models do represent one fruitful mechanism for organizing thinking about production.

Most scholars have not integrated these general institutional models with the rich economic and institutional literature specific to innovation. Such integration is fruitful, however, as it yields a variety of specific institutions for conducting innovation. These include: coordination with the firm; competition between firms generally; licensing of intellectual property from small firms to large firms; coordination through property rights and markets; competition through property rights and markets; and a mix of competition and collaboration in the private ordering that has traditionally existed in the context of publicly funded research. As discussed further below, each of these institutions varies

---

<sup>4</sup> In a forthcoming paper, my colleague Barak Richman offers an account of the comparative advantages of these three institutional models. See Barak Richman, *Firms, Courts, and the Market* (working paper 2004)

not only in their mix of coordination and competition but also in their openness with respect to exchanges of information.

In the context of innovation, the coordination within the firm model is probably best represented by the views of Joseph Schumpeter. Although Schumpeter is not concerned with transaction costs, and hence does not invoke Coasean terminology, he views innovation as driven by large firms with dominant positions in product markets. These dominant positions allow large firms to hedge the risks associated with innovation. The risk that large firms will become complacent is hedged by the continual threat that a competitor will arise to displace the large firm.

In contrast to Schumpeter's vision of serial monopoly, Kenneth Arrow focuses on the spur to innovation provided by contemporaneous competition between firms. In his model, coordination within the firm needs to be supplemented by competition between firms. Competition is necessary to guard against firm complacency and reluctance to innovate when such innovation would undermine a firm's existing business model. Although their models of innovation are very different, neither Schumpeter nor Arrow focuses on intellectual property and markets as a mechanism for information exchange or, indeed, on information exchange outside the boundaries of the firm.

A third model combines firms, intellectual property, and markets: specifically, it envisions small firms with creative new ideas as the primary engines of innovation. In this model, small firms are uniquely well positioned to generate, and market, new ideas. This model tends to involve an explicit role for intellectual property rights. More specifically, patent portfolios on new ideas not only help small firms directly by bringing

in licensing revenue, but they also allow small firms to attract venture capital.<sup>5</sup> Some advocates of exclusive licensing of university patents to small firms similarly support such licensing on the grounds that exclusivity is necessary for small firms to attract licensing revenues from large firms as well as venture capital.

A fourth model, involving coordination through contracts and markets, is most closely associated with Edmund Kitch. Kitch views intellectual property rights as mechanisms for efficient development of innovation. Most significantly, the holder of a broad property right, or “prospect,” on inchoate invention will be in a position to coordinate, through tailored licensing, the activities of all of those who work in the area of the property right. Such coordination will reduce duplicative work, particularly duplicative work that might otherwise occur because of “racing” towards the prospect of a downstream patent.

Another aspect of Kitch’s argument focuses on the role of property rights in fostering commercial development of a prospect, presumably where the likelihood of a downstream patent is insufficiently attractive (as opposed to excessively attractive, as in the race context). This argument places less emphasis on the role on monopoly power as a mechanism for coordination through the market and more emphasis on its role as a mechanism for ensuring that rents from development are not dissipated from competition. This second aspect of Kitch’s argument has precursors in the work of more traditional patent theorists like F.M. Scherer and Fritz Machlup, who note than patents can provide not only incentives to invent but also incentives to develop and commercialize. This model of innovation played a significant role in Congress’s decision in 1980 to pass the Bayh-Dole Act. This Act, which allows universities considerable discretion to seek

---

<sup>5</sup> Scherer, John Golden

patents on federally funded research, contemplates that exclusive licensing of such patents will create incentives for the firms that secure those exclusive licenses to commercialize the inventions.

Like Kitch, other contemporary scholars of innovation envision innovation through the lens of the market – that is, through licensing of intellectual property rights. However, while Kitch favors innovation through coordinated licensing of a broad patent, these scholars favor somewhat narrower patents that afford some protection against competition but nonetheless allow multiple parties to compete in a given area of innovation.<sup>6</sup>

A final prominent model of innovation involves federally funded research in universities. Under this model, academic scientific research is governed by communal norms rather than intellectual property rights. This model is most closely associated with the work of Robert Merton and his protégés. Mertonian norms emphasize the scientist's responsibility to contribute to a public domain of freely available scientific information, independent selection of research topics, and competition for intellectual credit through publication rather than through intellectual property. Mertonian norms represent a type of informal private ordering, albeit admittedly private ordering under the aegis of public funding. Like private ordering generally, Mertonian norms are supposed to operate within small communities – specifically, small communities of scientists working in the same or similar substantive areas. The strong social links between scientists in these communities lower the costs of transacting without resort to the hierarchical approach of

---

<sup>6</sup> Merges & Nelson

the firm.<sup>7</sup> Notably, the Mertonian model does not explicitly provide for the translation of research into commercial products – it tends to assume that such translation will occur through the natural diffusion of knowledge that has been made publicly available.

#### B. The Open Source Model

The open source model of innovation has links to the Mertonian framework for conducting scientific research. More specifically, the open source movement originated in a communal “hacker” culture that prevailed in certain academic and corporate laboratories in the 1960s and 1970s. At that time, packaged software was rare and individuals freely exchanged software and underlying source code for purposes of modification and improvement. Such exchange was facilitated when the U.S. Defense Advanced Research Project (“DARPA”) established the ARPANET computer network; this network served hundreds of universities, defense contractors, and research laboratories. Richard Stallman, then a researcher at MIT’s Artificial Intelligence laboratory, developed the basic idea of using copyright to keep source code freely available after MIT licensed some of the code created by the laboratory to a commercial firm, and the firm then prevented the MIT researchers who had developed the code from having access to it. In 1985 he founded the Free Software foundation to promote his idea. The concept was made more palatable to industry in 1998 when Bruce Perens and Eric Raymond extended Stallman’s idea to include software licenses that disclose source code but do not necessarily require improvers to disclose source code.

Although the term open source software refers to the availability of source code and not to methods of development, the free availability of source code creates a situation

---

<sup>7</sup> Cf. Star and Griesmer (positing “social worlds” theory in which coordination occurs through many nodes of a decentralized network).

where individuals working outside a single firm environment can contribute to software development. The possibility of relatively decentralized yet collaborative development has been further facilitated by the expansion of the ARPANET into the Internet. As various scholars have explained,<sup>8</sup> the transaction-cost reducing properties of the Internet allow coordination of innovative activity outside the hierarchical confines of the firm. Indeed, in some cases, transaction costs have been reduced to the point where software production can draw in large number of individuals, not just the small group traditionally contemplated by proponent of non-legal ordering.

Currently, the term open source, as used by the Open Source Initiative, encompasses over thirty different types of software licenses. Although the exact terms of these licenses vary considerably, and some involve fees for use, they share the requirement that source code be made available to the recipient of the licensed software.<sup>9</sup> Moreover, as a first approximation, open source licenses can be divided into two different categories: “copyleft” or “GPL” licenses that require licensees who make improvements to the software to make those improvements publicly available on the same terms provided open source terms that they received the original software;<sup>10</sup> a second category that essentially imposes no requirements on recipients. This paper will use the term open source in reference to software and the more general term open and collaborative production to denote projects where information and research materials are shared without the ordinary restrictions associated with property rights.

---

<sup>8</sup> Cite to McGowan, Benkler.

<sup>9</sup> Bruce Perens, The Open Source Definition, available at <http://perens.com/articles/osd.html>.

<sup>10</sup> Although Richard Stallman and some others argue that copylefted software should be called “free” software, this paper uses the term “open source” to encompass copylefted software.

Though the open and collaborative model of research and development is relatively decentralized, at least in comparison to the firm, it is more centrally coordinated than traditional biological science. In some respects, the open and collaborative model is similar to the production model put forward by Bruno Latour. Latour's "actor-network" theory focuses on how scientists build "black boxes" – stable facts and artifacts – and how they extend these black boxes into the world. By creating stable facts and dependable machines, successful network builders make themselves indispensable, compel assent, and position themselves in the center of production. As in Latour's vision of science, each open source software project has a central developer who is responsible for evaluating and integrating developments on an ongoing basis. In contrast, traditional Mertonian science accumulates in a less coordinated manner, as decentralized labs compete, make their research findings available in order to claim competitive priority, and then proceed to build upon each others' findings.

### C. Descriptive Accuracy and Normative Evaluation

In terms of descriptive accuracy and normative appeal, particularly with respect to the industries most relevant to this paper – the software industry and the biopharmaceutical industry -- each of these innovation models has advantages and disadvantages. The aspect of Kitch's model that features coordinated improvement through the licensing of broad patents fares most poorly. To be sure, broad patents are in fact issuing in some industries, including software and to some extent biopharmaceuticals.<sup>11</sup> The availability of broad patent rights has not, however, led to coordinated development through licensing. Rather, in the biopharmaceutical industry,

---

<sup>11</sup> In biotechnology, the availability of broad patents has been constrained to some extent by the manner in which the Court of Appeals for the Federal Circuit has applied the written description requirement of patentability. See *Eli Lilly, Enzo* etc.

broad patents on upstream research have led most prominently to lawsuits against unlicensed downstream developers of that research.<sup>12</sup> In the software industry, there is some evidence that broad patent rights have led to patent thickets.<sup>13</sup> Part of the reason for thickets is the legitimate availability within the patent system of blocking patents – that is, patents that cover the same innovative space. Blocking patents are not acknowledged by Kitch’s model. Another reason for patent thickets stems from the institutional reality that the Patent and Trade Office is not particularly well equipped to ensure that issued patents do not overlap in illegitimate ways. To be sure, the available evidence indicates that because software patents tend not to be used offensively, patent thickets do not generally block development.<sup>14</sup> Nonetheless, the transaction costs associated with accumulating and maintaining patent portfolios are not negligible.

In contrast with Kitch’s argument that intellectual property rights can be used to coordinate development, his view that intellectual property rights can hedge the risk associated with development has considerable appeal. In contemporary drug development, clinical testing for safety and efficacy is the Bermuda Triangle of R&D: only 8% of drugs that begin such testing on human become successful products.<sup>15</sup> Kitch provides a good explanation for why significant patent rights are necessary for promising chemical entities and biologics that must face the hazards of clinical testing.<sup>16</sup> Schumpeter provides insight on why this clinical testing process is generally managed by large pharmaceutical firms.

---

<sup>12</sup> Ariad v. Eli Lilly (Nf-Kb cell signaling patent); University of Rochester v. Pfizer (cox-2 patent).

<sup>13</sup> Bessen; also note that Ronald Mann’s study confirms existence of many patents

<sup>14</sup> Ronald Mann

<sup>15</sup> Cite to recent FDA whitepaper.

<sup>16</sup> Note that they may not need to be as broad as Kitch posits because various features of the drug regulation process provide protection from competition.

The small firm model, which focuses on the manner in which patent rights can provide a mechanism for small firms to survive by attracting venture capital and licensing revenues, also appears to have some descriptive accuracy. Various commentators have suggested the model applies in the context of small biotechnology firms – that is, firms that produce research inputs for large pharmaceutical firms but do not themselves produce therapeutic end products<sup>17</sup> – as well as small software firms.<sup>18</sup> Its descriptive accuracy notwithstanding, the small firm model makes several contestable normative assumptions. First, the model tends to assume that small firms are uniquely poised to generate and market valuable new ideas. Alternatively, or perhaps additionally, small firms are valuable because they will eventually compete with large firms. Take the second assumption first: the empirical data supporting the idea that small firms eventually “grow up” to become viable competitors to large firms is not robust, particularly in the biopharmaceutical industry. The first assumption, that small firms generate and market ideas unlikely to be generated by large firms, is more compelling. The organizational theory literature gives us reason to believe that the structure of large firms may be inimical to thinking in radical new ways. Nonetheless, the empirical case is hardly solid. To the extent that small firms do not provide “value added,” their patent claims may reasonably be viewed as creating costs that tax or even impede innovation. This is particularly true for the biopharmaceutical industry, where small firms often have patent claims on the upstream inputs necessary to produce end products.

As for the model of academic science put forward by Merton, it could be argued that Merton’s description has never represented accurately the nature of research

---

<sup>17</sup> Cite to Scherer  
<sup>18</sup> Mann

relationships in highly competitive fields like human genetics. In such fields, scientists have always kept information secret for purposes of winning the race for publication priority.<sup>19</sup> In addition, even without deliberate secrecy, there have always been impediments to the full dissemination of biological information and materials. For example, the publication process itself creates some delay. Moreover, even when an article is published, it can often be obtained only by those who subscribed to the journal and could pay the subscription price.

In any event, whatever the accuracy of the Mertonian model for the era of molecular biology prior to the late 1970s, that model has limited applicability to research in molecular biology today. To some extent, the breakdown of the model may have resulted from increased competitive pressure. For example, data collected by Eric Campbell and his colleagues indicates that, in the genetics area, problems associated with access to scientific data and materials – even scientific data and materials supporting research that has already been published – became more prevalent in the late 1990s. Campbell and his colleagues observe that part of the reason for this increased secrecy appears to be greater competitive pressure. In addition, even for those who don't deny access for reasons of competition, the cost associated with transferring biological materials can be high.<sup>20</sup>

Impediments to the Mertonian vision of scientific communication and cumulative progress have also been significantly exacerbated by the introduction of proprietary rights. The movement toward intellectual property rights was the consequence of a “perfect storm” of mutually reinforcing forces. First, by the late 1970s, with the

---

<sup>19</sup> The canonical example was perhaps the race between James Watson and Linus Salk. See also Walsh and Hong (discussing substantial secrecy in mid 1960s science, particularly biological science).

<sup>20</sup> Cite to Nature article on cost of transfer

emergence of such products as Genentech's recombinant insulin, it was clear that ostensibly basic research techniques such as the Cohen-Boyer method for producing recombinant DNA could have tremendous market value. These market forces led the primary producers of such upstream research, publicly funded universities, to begin seeking patents.<sup>21</sup> As university administrators were becoming more interested in patenting, Congress was becoming increasingly persuaded that intellectual property rights on publicly funded inventions could stimulate development and commercialization in the manner contemplated by theorists such as Scherer, Machlup, and Kitch. In particular, in passing the Bayh-Dole Act of 1980, Congress wanted to encourage universities to patent inventions and then license those patents exclusively to commercial firms. Although Congress may have been contemplating the use of such rights to foster commercial "scaling up" of inchoate end products, such as promising drug candidates, universities quickly began to seek rights on more upstream biomedical research. Finally, case law began to encourage the patenting of more upstream biomedical research. In particular, the 1980 *Diamond v. Chakrabarty* case, and the 1982 formation of the Court of Appeals for the Federal Circuit, made it clear that legal doctrines such as patentable subject matter, utility, and nonobviousness would not pose significant obstacles to the patenting of upstream research.

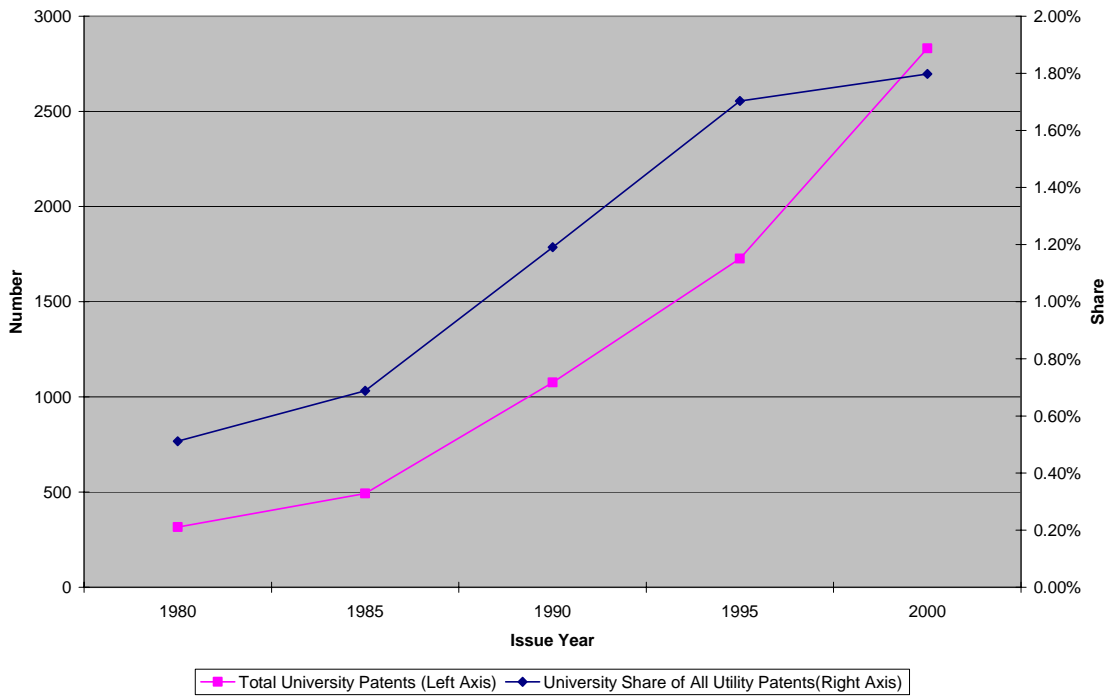
The result of these converging factors was not simply overall patent growth within universities but particularly steep levels of growth in university biotechnology patenting. Figure 1 shows the overall increase in patenting by U.S. universities that grant doctorates (the so-called "Carnegie" universities). As a percentage of all utility

---

<sup>21</sup> Mowery et al. 2001 Cf. Harold Demsetz (where private value of rights exceeds private cost of enforcing rights, transition to property rights regime will occur)

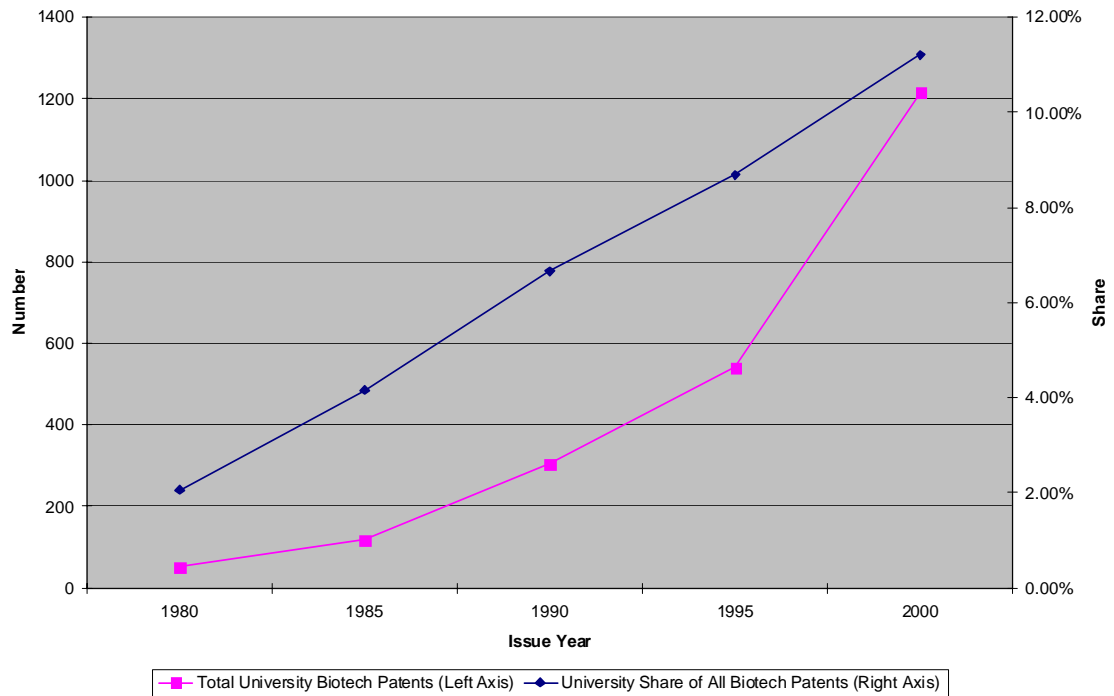
patents, overall patenting rose from approximately 0.2% of all patents in 1980 to approximately 1.9% in 2000.

Figure 1: University Patenting: 1980-2000



Moreover, according to one estimate that uses International Patent Classification categories to identify “biotechnological” patents,<sup>22</sup> the share of such patents obtained by doctorate granting universities rose from 2% of the total to approximately 11% in 2000.

Figure 2: University Patenting in Biotech: 1980-2000



(Data courtesy of Bhaven Sampat, University of Michigan)

The study by Eric Campbell and his colleagues indicates that, in addition to competitive pressure, restrictions imposed by commercial sponsors, as well as complications encountered in university negotiations over agreements for the transfer of biological materials, have contributed to decisions to deny access. Indeed, some have argued that even in cases where a decision to patent is not the direct reason for secrecy, they may contribute to such secrecy. In particular, the increased assertion of proprietary

<sup>22</sup> Biotechnology was defined to include the following 4-digit international patent classifications (IPCs): A61K, C07H, C07K, C12N, C12P, C12Q, and G01N. Note problems with these classifications.

rights by some may contribute to a more general breakdown in the culture of collegiality and reciprocity.<sup>23</sup>

Finally, even if were entirely accurate as a descriptive matter, Merton's model of scientific progress also has some normative difficulties. As an initial matter, it does not provide for mechanisms of information integration. Particularly when the volume of data available is large, the need for integration of information by methods other than the sheer diligence of the individual scientist can be significant. More importantly, the Mertonian model does not explicitly provide for commercialization of research. Rather, as noted earlier, it appears to assume that commercialization will occur automatically through diffusion of the university-generated knowledge. This view is likely to be accurate in situations where commercialization costs are relatively low or where the research can be used immediately in a manner that leads to further patents downstream. In contrast, where commercialization costs are high and downstream patents to protect investments associated with commercialization are not available, the Mertonian model is incomplete.

What about the open source model of innovation? Although we are far from a definitive verdict, this innovation model does appear to have enjoyed some success in the development of software. Outside of human capital, software development does not require significant capital investment. Moreover, commercial firms that produce open source software defray the cost of human capital by relying on the volunteer labor of programmers outside the firm. These firms can then survive on the relatively modest revenue stream generated by selling services related to the software. The apparent success of the open source software movement has coincided with, and in some cases led to, calls for open source-type development in other areas, including biomedical research.

---

<sup>23</sup> Gruschow 2003.

The remainder of this article is devoted to examining those open and collaborative projects that have thus far emerged. In the next Part I discuss the methods I used to identify important biomedical scientists and projects that operate under an open source model. I also discuss methods for identifying technology transfer offices that are likely to be central to open and collaborative biomedical efforts.

## Part II: Methods

I used snowball sampling to identify biomedical research projects that operate under an open and collaborative model. First and most obviously, the projects include a significant amount of bioinformatics software. A second prominent category of projects involves databases of biomedical information, including the now completed Human Genome Project; the SNP Consortium; various projects for sequencing other genomes; and the ongoing Haplotype Map (“HapMap”) project. A third category of project is database annotation. Finally, a few “systems biology” projects appear to fall under the general rubric of open and collaborative model. The most prominent of these is the Alliance for Cell Signaling, a consortium directed by Nobelist Alfred Gilman of the University of Texas/Southwestern Medical School.

I asked the scientists associated with each of the projects two categories of questions. One set of questions focused on the science of the project: in particular, what are the primary scientific goals and how does the open and collaborative nature of the project advance or hinder those goals. The other set of questions assessed the social science. Specifically, I inquired into the mechanisms of coordination, reward structure, and publication associated with the projects. I also inquired into intellectual property

rights; have such rights been sought and how have technology transfer personnel reacted to the project's decisions regarding such rights.

In order to determine which university technology offices to contact, I identified universities that fell into one or more of the following categories: 1) top 10 software patentees in 2000<sup>24</sup>; 2) top 10 biotechnology patentees in 2000<sup>25</sup>; 2) 3) top 10 recipients of federal biomedical research funding in 2000. In total, 19 universities fell within one or more of these categories. These are: the University of California system and the University of Washington (categories 1, 2, and 3); Stanford University (category 1); the California Institute of Technology (category 1); Cornell University (categories 1 and 2); University of Pittsburgh (categories 1 and 3); Massachusetts Institute of Technology (category 1); Columbia University (category 1); the University of Wisconsin (category 1); Georgia Institute of Technology (category 1); University of Texas (category 2); Rockefeller University (category 2); Harvard University (category 3); Yale University (category 3); University of Pennsylvania (category 3); Johns Hopkins University (category 3); Baylor University (category 3); Washington University, St. Louis (category 3); and the University of Michigan (category 3). I asked the technology transfer representatives at these institutions were questioned about the institution's policies towards technology transfer in the context of biomedical research and software, including open source software. For those institutions within the group of 19 that had participated or were participating in "open and collaborative" biomedical research projects, I asked these technology transfer officers about their attitudes towards those projects.

### III. Results

---

<sup>24</sup> Software was defined using the following IPC classifications: \_\_\_\_\_. Note that this is a conservative definition; also the same one used by Graham and Mowery in their 2003 NAS paper.

<sup>25</sup> Biotechnology was defined using the same IPC classifications used to derive the data in Figure 2.

## A. Software Projects

According to computational biologist Steven Brenner, an assistant professor at the University of California, Berkeley and a founding developer of BioPerl, open source software is “extremely prevalent” in the bioinformatics software community. Many software projects, particularly small software projects, operate under an open source model. By those who participate in such projects, open source is seen as a good mechanism for information dissemination and for “moving science forward.”<sup>26</sup>

Before coming to Berkeley, Brenner had participated in open source software development, specifically the development of SCOP software, at the U.K. Medical Research Council. This experience convinced Brenner that his lab at Berkeley needed to be able to produce open source. When Brenner started negotiating with the University of California, one of his requests was that his lab be allowed to produce open source software. Open source was “totally foreign to them.” The negotiations ultimately headed to President’s Office in Oakland, to the Vice-President responsible for IP. It was a time consuming process. Brenner was initially told that all patents belonged to the university. Eventually, however, the office agreed to allow Brenner’s lab to produce open source. He understands that open source is an option for other research labs, but these labs have to show new software to the UC, Berkeley Office of Technology Transfer. Brenner believes that this disclosure requirement is in tension with the philosophy behind open source because code is being changed every day. Brenner does not believe, however, that all bioinformatics software should necessarily be open source.

In this last sentiment, Brenner appears to reflect a significant segment, perhaps the predominant segment, of the bioinformatics software community. This segment believes

---

<sup>26</sup> Interview with computational biologist Steven Brenner, UC Berkeley.

that there is a role for both open source and closed source software within the community. In 2001, when some publicly funded software developers asked NIH to mandate that all software developed with its funding be open source, the software community engaged in a vigorous debate over the merits and demerits of any mandate.<sup>27</sup>

\*\*Add discussion of Defense Advanced Research Project Agency's sponsoring BIOSPICE\*\*

### B.Database Projects

The first, and still most important, “open source” biomedical research database project was the publicly funded project to sequence the human genome. The history of the Human Genome Project has been discussed by a number of authors and scholars, including one of the scientists, Sir John Sulston, who was a primary force behind the 1996 decision to release raw sequence data within 24 hours in the public domain.<sup>28</sup> The large-scale sequencing centers in the U.S. and U.K. that signed on to these so-called “Bermuda rules” also agreed to refrain from seeking any proprietary rights in the data. In January 2003, the National Human Genome Research Institute extended this policy regarding immediate data deposition without accompanying intellectual property rights to all large-scale data “infrastructure” projects. Indeed, at this meeting, NHGRI prioritized immediate and full access to data over the traditional scientific norm that the investigator who generates the data has the right to do the first analysis of this data.<sup>29</sup>

The scientists who first formed the human genome sequencing coalition emerged from the community that was working on the genome of the nematode, or *c. elegans*. Unlike many in the human medical genetics community, this community had always

---

<sup>27</sup> Russ Altman et al., Whitepaper on Open Source Software in Bioinformatics (on file with author)

<sup>28</sup> John Sulston, *The Common Thread* (2002).

<sup>29</sup> 421 *Nature* 875 (2003)

been characterized by a norm of free information exchange. Early in the process, they insisted on free information exchange. This insistence on information exchange eventually led to the adoption of the Bermuda principles.<sup>30</sup>

Unlike traditional human genetics, which revolved around individual laboratories that tended to be highly competitive – and hence uneven in their willingness to share information, particularly pre-publication – the Human Genome Project was, from the outset, an intensively collaborative endeavor. Particularly after the public project was faced with a challenge from Craig Venter, the leader of a private effort to sequence the genome, the major sequencing centers – the so-called “G-5” – reported their progress in weekly conference calls with the funding entities, principally the National Human Genome Research Institute.

\*\*\*Human Genome Project explicitly decided not to “copyleft” their data; new International HapMap project has a modified version of the copyleft, however\*\*\*

\*\*\* SNP Consortium: publication to defeat patents\*\*\*

### C. The Distributed Annotation System

The distributed annotation system (“DAS”), itself an open source software program, is set up to facilitate collaborative annotation of various genomes, including human, mouse, *C. elegans*, fruit fly, and rice. Any interested party can set up an annotation server. Annotation information is stored in a series of databases that are connected to the Internet. End users of the information – in other words, researchers – choose the annotations they want to view by typing in the URLs of the databases. According to Lincoln Stein, one of the designers of the DAS, it was “designed to facilitate comparisons of annotations among several groups. The idea is that an

---

<sup>30</sup> Discussion with John Sulston, 3/26/04.

annotation that is similar among multiple groups will be more reliable than an annotation that is noted by one group.”<sup>31</sup> The quality of the annotation is also judged by looking at published papers that describe the annotation technique.

#### D. “Systems Biology” Projects

One of the least studied, and potentially most interesting, application of the open source model, involves various projects that aim to study not a single gene or protein but, rather, biological systems. In the last five years, the National Institute of General Medical Science (“NIGMS”) has funded five large grants that are intended to “make resources available for independently funded scientists to form research teams to solve a complex biological problem that is of central importance to biomedical science . . . and that would be beyond the means of any one research group.” These grants depart from the traditional R01 grant model, as they aim to stimulate research beyond the individual steps in biological processes and towards problems of “global control and integration” that draw upon not only mainstream biomedical science but also physics, mathematics, and computer science.

Notably, the NIGMS Request for Applications (RFA) for the Glue Grant program stresses the important not only of the science but also of the social science. As the RFA notes, “[a] high level of resources may be requested to allow participating investigators . . . to approach a research problem of overarching importance in a comprehensive and highly integrative fashion.” The RFA also emphasizes that applicants submit a plan for sharing research resources generated through the award as well as a plan for addressing intellectual property rights. In this respect, the RFA is similar to the efforts NIH made in the context of its initial foray into generation of information regarding single nucleotide

---

<sup>31</sup> Interview with Lincoln Stein, 3/26.

polymorphisms (“SNPs”) -- that is, single base points within the genome at which the DNA sequence of individuals differs. Before a private group of pharmaceutical companies stepped forward to put SNPs in the public domain, NIH had decided to allocate public funds for SNP identification. In its RFA for SNP-related grants, NIH asked grant applications to specify their plans for sharing data, materials, and software. More recently, NIH has asked all applicants who apply for more the \$500,000 in grant funding to include within their grant application a plan for dissemination of data that will be generated from the grant. The NIGMS RFA goes beyond these prior precedents, however, in that it specifically notes that “because dissemination [of information] is a critical aspect and fundamental purpose of this RFA, evidence of the commitment of the large-scale leadership to the sharing of research resources and to effective management of intellectual property issues will be part of the *scientific merit review*.”<sup>32</sup>

The Alliance for Cell Signaling (“AFCS”) was the first of these glue grants to be funded. It is also the glue grant that most clearly reflects an open and collaborative approach to research. The Alliance is led by Nobelist Alfred Gilman of the University of Texas, Southwestern Medical School. Gilman won his Nobel Prize for his work on the role of G proteins in cell signaling, and the goal of the project is to map complex signaling networks. While cell biologists once believed that signals, such as a chemical ligand binding to a cell receptor, initiated only one pathway, it is now clear that a chemical stimulus can excite different networks that interact in complex ways. Conversely, different chemical stimuli can excite the same pathway. The ultimate goal of the experimental work within AFCS is three-fold: first, to catalogue the “parts” – that is,

---

<sup>32</sup> NIH Guide: Large-Scale Collaborative Project Awards, Plan for Handling Intellectual Property and Sharing of Research Resources (emphasis added)

the stimuli and signaling proteins; second, to identify the interactions between these parts; and third, to generate a computational model of signaling within the cell.<sup>33</sup> A separate component of AFCS is responsible for maintaining and updating a collection of Web pages, each devoted to a particular signaling protein, that collects all public information on those proteins.

AFCS comprises seven “wet labs” and one bioinformatics laboratory that is responsible for integrating the data produced by the eight wet labs.<sup>34</sup> AFCS initially studied B-lymphocyte and cardiac myocyte cells. Because of technical limitations that have arising in using the B lymphocyte (including difficulties in apply RNA interference technology to the B lymphocytes), AFCS has now switched to using the macrophage as its model cell.

My research on AFCS included interviews with Alfred Gilman; 3 lab directors (Shankar Subramanian, Director of Laboratory Bioinformatics and Data Coordination; William Seaman, UCSF, Director of Development of Signaling Assays; and Alex Brown, Director, Lipidomics Laboratory); Steering Committee member Henry Bourne (UCSF) and Editorial Committee Chair Patrick Casey (Duke).

All interviewees noted the novel nature of the collaboration in which they were engaging; indeed, AFCS is seen by its participants as an “experiment on how to do

---

<sup>33</sup> Shankar Subramanian.

<sup>34</sup> The seven wet labs are as follows: Cell Preparation and Analysis Laboratory, UT Southwestern Medical Center (Paul Sternweis, Director and Richard Scheuermann, Associate Director); Laboratory for Development of Signaling Assays, UCSF (William E. Seaman, Director); Molecular Biology Laboratory, California Institute of Technology (Melvin I. Simon, Director); Protein Chemistry Laboratory, UT Southwestern Medical Center (Marc C. Mumby, Director); Microscopy Laboratory, Stanford University (Tobias Meyes, Director); Antibody Laboratory, UT Southwestern Medical Center (Susanne M. Mumby, Director); Lipidomics Laboratory, Vanderbilt University (Alex Brown, Director). The director of the UCSD bioinformatics and data coordination laboratory is Shankar Subramanian.

experiments.”<sup>35</sup> Several noted that there is significant opposition within the scientific community to any move away from the R01 individual laboratory model and towards a model that resembles “big biology.” The interviewees also agreed, however, that it would be difficult to imagine a single laboratory having all of the necessary expertise for the work required. According to the interviewees, significant effort is made to coordinate and standardize the wet lab biology being done by the different labs. For example, much work has gone into standardizing the cell lines that are used for experiments in different labs.

Communication among the different laboratories and Alfred Gilman’s central office at UT Southwestern Medical School occurs along multiple different pathways. First, and perhaps most importantly, all laboratories participate twice a month in video conferences. According to one participant, Alex Brown, videoconferencing – which often includes simultaneous PowerPoint presentations – was a sociological experiment for AFCS members, but it now provides “85%” of the value of being in a real meeting. Second, the data analysis committee, which is a separate component of AFCS but, for the most part, comprises investigators from the universities participating in the eight laboratories, also meets once a month. Third, each laboratory meets with the AFCS steering committee once a month. Finally, AFCS has an annual meeting each year.

Another novel aspect of AFCS involves its lack of emphasis, at least thus far, on conventional publication through scientific journals. Rather, after limited internal review, data publication takes place expeditiously on the Web. Moreover, AFCS investigators have no “head start” in terms of analysis of this data. In this respect, as in many others, AFCS is explicitly modeled on the Human Genome Project.

---

<sup>35</sup> William Seaman

The lack of emphasis on conventional publication coheres with the organizational structure of AFCS. While most lab directors are senior tenured professors who have advanced through the conventional career track for academic scientists, many of the individuals who work in AFCS laboratories are on a different, and in some respects novel career track: they are postdoctoral scientists who are not planning on tenure-track appointments. Many of these individuals plan to go into private sector research.

Finally, and perhaps most unusually, all participants in AFCS have agreed to disavow intellectual property rights in their research. Indeed, Shankar Subramanian, the lead bioinformatician on the project, releases most AFCS-related software under the “copyleft” version of the open source license. This agreement to disavow conventional property rights is, quite obviously, somewhat contrary to the trends in patenting that we have witnessed since passage of the Bayh-Dole Act. Moreover, many of the institutions participating in AFCS – perhaps most notably the University of California system but also the University of Texas and the California Institute of Technology – have substantial numbers of patents. Every AFCS participant with whom I spoke attributed the decision to disavow intellectual property rights to Al Gilman. Moreover, Gilman’s Nobel Prize, as well as his stature in the area of cell signaling enabled him to convince recalcitrant university administrators, particularly at the University of California and the California Institute of Technology, not to interfere in the disavowal of property rights. But even someone of Gilman’s stature and “force of personality” found the task difficult.

Interviewees concurred that they were not sure how many directors of large-scale projects could “pull off” what Gilman had done. Gilman himself believes that his stature played a role, as did the stature of his faculty collaborators. According to Gilman, he was also

able to convince technology transfer officers that what AFCS was doing would not have commercial value.

Gilman had multiple motivations for his decisions to require disavowal of intellectual property rights. First, he sees the AFCS as a data “infrastructure” project similar to the Human Genome Project. Its value lies in the ability of others to build upon the raw data. Second, he wanted to avoid the complications and distractions associated with the filing of intellectual property rights in situations where there might be multiple discoverers.

**\*\*Other glue grant projects\*\***

#### E. University Licensing Policies

Biomedical Research: Many universities have reasonably well developed policies on the licensing of biomedical research. Although they often patent biomedical research, various major research universities, including MIT and Stanford, report awareness of, and some agreement with, licensing recommendations made by NIH in 1999. The NIH recommendations note that universities should consider carefully whether to patent broadly enabling research tools. Even if these tools are patented, they should be licensed nonexclusively. The extent to which universities are actually patenting more selectively or licensing nonexclusively is less clear. As noted earlier, AFCS members report resistance at key universities to their efforts.

In addition, at least some university technology licensing officers within the University of California, most notably Joel Kirschbaum of the University of California, explicitly view nonexclusive licensing as a drain on technology transfer office resources

that bring little compensating gain back to the university. (\*\*Other available data\*\* on licensing)

Software: The area of software is quite different from biomedical research. Universities are much less knowledgeable about software, including open source software, than they are about biomedical research. Generally, however, software is not patented because the costs of the application process outweigh any licensing revenues that might be derived. A few universities have well developed policies on open source software (University of Texas, University of Washington, Georgia State). For software that is not commercially valuable, these universities allow open sourcing at no charge. For software that is commercially valuable, these universities may recommend a “forked” open source license: paying for commercial users and non-paying for non-commercial users.

MIT/Stanford claim to be very “open” to open source

Other universities less clear (Cornell, parts of UC system (in particular UCSF, which explicitly rejects open source model for software), Harvard, Yale)

#### Part IV: Discussion

##### Advantages

- 1) Good for computational biology, databases; may be particularly good for complex “systems biology” (with many eyes, all bugs are shallow)

- 2) Reduces transaction costs associated with follow-on R&D
- 3) In collaboration between different institutions – the types of collaborations particularly important for systems biology – avoids costs associated with hammering out agreements on who gets the IP
- 4) Coordination is an improvement over somewhat uneven cooperation, openness of traditional molecular biology, particular traditional human genetics

#### Disadvantages

- 1) Could undermine development of upstream R&D towards drug
- 2) Probably does undermine small biotech (is that a problem)
- 3) Collaboration is generally good, but perhaps some competition can incentivize scientists (Human Genome Project)
- 4) In some cases (AFCS), need to find way to reconcile immediate data release with later analysis and publication of data. Relatedly, is alternative career track of “data producers” who don’t publish (AFCS) a good thing?
- 5) Restrictive, “copyleft” licensing may be problematic

Observation: Outside of software, will need public funding; little prospect of significant licensing related to “services”

#### Practical Problems

- 1) Universities that are heavily in biomedical sciences may be resistant to open source models, including open source software (UCSF)

2)Universities that have historically had a greater presence in software may be more amenable to open source models (University of Washington). Those who do software sometimes come from a different background (library, information services) than those who do biomedical research.

**\*\*TO BE DONE\*\***