Michael Brudno
SIMS 296
5/15/99

Protection of Genomic Databases

1. Introduction

The emerging bio-tech industry has caused many problems for legal scholars. The bio-tech companies, which often do not make any product protectable under one of the standard intellectual property regimes, are nevertheless extremely important in the development of data that helps pharmaceutical companies (pharmas) develop new drugs faster and cheaper. This paper will look at one such product: genomic databases. Genomic databases produce an especially interesting subject for study because databases currently are not protected by any law. Companies that make these databases usually license these to their clients, both because this provides the database with some form of protection, and because the large amount of money at stake usually makes the cost of negotiating a license small as a percentage of the deal. This paper will look at licensing, and also at the main alternatives to licensing, either proposed or available now, and whether any of these may prove more advantageous to the companies or the general public.

II. Protection of Underlying Information

Before discussing genomic databases, it is necessary to look at the level of protection that the underlying data has. The patentability of genes is a serious problem, but generally beyond the scope of this paper. It suffices to state the current situation. The granting of patents on genes with a well understood function is now taken for granted.[6] A more complicated problem is the patentability of expressed sequence tags (ESTs). The ESTs correspond to parts of the genome which we know to be useful (only 10% is), but the exact function of which we do not yet know. Since of the about 100,000 genes which together make up the human genome the function of very few is actually known, the bulk of the data is contained in unknown ESTs [9]. The Patent and Trademark Office (PTO) has started giving patents on ESTs, but there has not yet been a challenge of such a patent in court.

III Types of Databases

Currently, genomic databases come in three main types: public, proprietary, and differentially priced. These terms are meant to refer to the database itself, and not to the underlying data.

III.1 Public Databases

A public database is free for anyone to use, whether the user is a scientist in a non-profit institution or a large pharmaceutical company. The best known of these is GenBank, developed by the National Institutes of Health (NIH) Some others are DDBJ (DNA Database of Japan), EMBL (European Molecular Biology Laboratory), and Swiss-prot, a protein database from Switzerland. All of these databases suffer from the same problems: the funding for their support is small, and because of the large amount of data which is generated on a daily basis they have all become archival rather than curated. That is to say that they contain large amounts of information with very little annotation or validation of correctness. These databases mainly depend on submissions by the scientists, who in turn are required to submit as a condition for publication in most journals. About 100,000 new base pairs per month are logged in these databases.[7] There are several problems with GenBank and similar databases. The main one is that because of poor curation it is very hard to automate any application. The fields in GenBank entries do not have standardized values, therefore almost every search, no matter how simple, requires that a human look over the results. Also since these databases depend on submissions, they can not guarantee correct descriptions or proper documentation in the entries.[8]

III.2 Proprietary Databases

Proprietary databases are usually much better curated then public ones. These are developed by bio-tech companies, usually for large pharmas. Although they come in all types, in general they share several features. These databases benefit from being as large as possible. Since the buyer will be looking in it for an EST which corresponds to a specific illness, he wants as large a collection as possible. The second requirement is that these databases need to be on the cutting edge of the research. This is best categorized by **Celera**, a subsidiary of **Perkin-Elmer**. Celera's main product is the Human Genome Sequencing Database, which contains over 1.2 million ESTs and is licensed to pharmas. Most of the

ESTs are generated in-house, and are considered proprietary by the company. Nevertheless, they release all of the data which they find after a three month period. This indicates the confidence that given a 3 month head start on using the new sequences, they will be able to be the first to patent a useful product deriving from them.[14] Thus, to be useful, the databases need to be up to date. The most important feature which all of these databases share is that they are expensive. **Incyte** charges $15-20 million for non-exclusive access to its databases for a three year period. **Human Genome Sciences** (HGS) sold exclusive access to its database for $125 million. These high prices are a combination of the large amount of effort required to collect and annotate the information, and the large R&D budgets of pharmaceutical companies.[7] These prices also put the products well out of the range of academic users.

III.3 Differentially Priced Databases

These databases give free (or at least reduced cost) access to academia, while industry still has to pay for access. A good example of this is EcoCyc, a database of the DNA of the E. Coli bacteria developed by **Pangea Systems**. This database is freely available to researchers working for non-profit institutions and government labs, while for-profit companies have to pay a fee to use it. The database is not protected through any means and depends on the companies' honesty.[11] Another example is the forthcoming database of the *Drosophila melanogaster* (Fruit Fly) by the Berkeley Drosophila Genome Project in Lawrence Berkeley Laboratories (LBL) and Celera.[10] The common feature of differentially priced databases is that they are generally not directly useful in the development of drugs. The reasons for this will be discussed subsequently.

IV Conflicting Interests

The different interests involved in the development and use of genetic databases make this a much harder problem than the classical inventor versus the general public. The bio-tech companies which make these databases, the large pharma which use them, the academia which needs access to the raw data to conduct their own research, and the general public all have their own conflicting needs. These are rarely in agreement and are hard to coordinate. Let us look at each one in more detail.

IV.1 Interests of Bio-tech

The interests of the bio-tech companies should be given the primary place in our discussion because they make these databases, and it is unclear who would fill the void should these companies disappear. These companies expend a lot of money towards creating a well annotated, useful product, and should certainly reap the fruits of their efforts. However this does not mean that they would necessarily benefit from the strongest possible protection. Although some of the industry leaders do their own sequencing in order to increase the value of the product (of the 3 million ESTs in Incyte's *LifeSeq Gold* package, 2.3 million are proprietary), all include the publicly available ESTs in their packages, and some include only these.[12] These sequences are collected from GenBank and similar databases. If these databases to become proprietary, many of the companies which rely on this free source of data would be unable to continue. Therefore the companies would most benefit from a system where their product (the databases custom-made for large pharma) are protected, while the public databases remain open to everyone.

IV.2 Interests of Large Pharma

For the large pharmaceutical companies the genomic database is only a step toward their final product, which is a marketable drug. Thus it is only seen as an expense, one which should be minimized, if at all possible. If they could get this data for free they would be perfectly happy, but this is not a plausible notion. Although there have been some efforts toward the creation of free, well curated databases through the funds of contributing pharma, of which the SNP consortium is the best known example,[8] currently the in-house abilities of the large pharma are not capable of meeting the demand of the R&D departments. Also because of the high overhead of developing curatorial tools it is much more efficient to have an outside provider of this data. Since the pharmas are stuck, at least for now, with these outside providers, it should be in their best interest to reduce the underlying costs of this provider, thereby reducing the price. Another aspect which the pharma are interested in is relative exclusivity. The search for a cure to a disease in the genome has been described as "finding a needle in a haystack, [except for] finding the gene is even more difficult, because even close up, the gene sill looks like just another piece of hay." [9] Since there are only so many needles in this haystack, a company wants to be reasonably sure that as few people as possible are looking in their part of the haystack.

They want exclusivity, and the fewer other groups that are looking in the same database, the higher the value of this database for the pharma. This stands in direct contrast with the usual experience of positive network effects with most information goods.[5] Although lately the demand for exclusivity has become somewhat less stringent, with pharmas willing to accept non-exclusive contracts and even collaborating with each other to reduce cost, this remains a major factor.

IV.3 Interests of the Academia

These interests are perhaps the simplest of the four groups. The academics want access to lots of free information. They are usually not concerned with the economic value of the product they are using, and are not a competitor of the big pharmas. Since they were responsible for so much of the basic science that underlies the fancy databases created by bio-tech companies, they do not understand why they should not get complete freedom to use this databases in their research. While they could take advantage of some database protection to market their own databases, creating and maintaining a database involves much non-scientific work, and most scientists would not be interested in this.

IV.4 General Public

The general public is interested in more drugs getting to market faster. They are not interested in how these drugs are developed, or where the basic knowledge which underlies this new drug comes from. Thus the interests of the general public closely correspond with the interests of the big pharma, since it is the pharma who bring the final product to market.

IV.5 Observations

Initially it seems that in this case the interests of tree of the parties (big pharma, bio-tech, and general public) are in agreement, and as such should outweigh the interests of the academia. This, however, is not so clear. The academia created the underlying science on which the current technology rests, and there is a real danger that if the needs of this sector are ignored that important discoveries which may fundamentally change our understanding of biological processes will never be made. Since for the companies the pocket book is much more important than the advancement of science, it is unreasonable to expect them to invest much money into research which has no foreseeable monetary return.

V. Private versus Public

Another aspect which makes the task of deciding what an appropriate protection policy would be is the blurry line between public and private institutions. Some companies have sponsored the creation of labs within universities. These labs, although nominally independent of the sponsoring companies, do research in fields which interest the companies and usually license their discoveries back to these companies. It is hard to label a lab which is doing this "public", though all the researchers may be on the staff of a university or a major research lab. A good example of this is the $1.3 million donation Phytotech made to Rutgers University in order to support phytormediation research, in return getting preferred licensing terms on any valuable discovery.[1] Another barrier which must be overcome in creating a viable public v. private distinction is the proliferation of start-ups, many of them founded by researchers from major universities. An example of this is Acacia Biosciences, founded by Jasper Rine, a professor of Molecular Biology at UC Berkeley.[13] If a company were to distribute its products for a reduced fee, or for free to a researcher who later created a start-up, the company would reduce the value of its product to other customers who would prefer exclusive or selective access. The researcher can effectively use his academic status as a wall until he is ready to bring his product to market, incurring few of the costs of the for-profit companies, but effectively competing with them. As a result, the differentially priced products (described in III.3) are usually those with smaller value to the industry, and hence cheaper ones. While a major pharma which spends many millions of dollars on research would not hesitate to buy access to a differentially priced database for a small to medium amount, they might find it much wiser to set up a lab at a university in order to get a discount on Incyte's hefty $15-20 million rate for its *LifeSeq* database. The blurriness of the division between public and private in the bio-tech and pharmaceutical industry makes it hard to create differential pricing regime based exclusively on up-front payments. One possible way to go around this is using royalties on future products and reach through license agreements (RTLAs). These methods will be covered subsequently.

VI Protection Schemes

This section will discuss the various ways of dealing with the problems outlined above. Various legal protection schemes will be examined and evaluated based on how well they satisfy the interests of different parties, as outlined above. The protection schemes to be discussed will be copyright, a federal database protection law, and licenses.

VI.1 Copyright

Section 103 of the Copyright act of 1976 provides for copyright protection for compilations and derivative works. The act is clear that, "the copyright in a compilation or derivative work extends only to the material contributed by the author of such work, ... and does not imply any exclusive right in the pre-existing material."

VI.1.1 Overview

The 1993 Feist decision made it clear that in order for a compilation to be protected by the Copyright Act its creation must have entailed "some minimum modicum of creativity." No matter how hard the researcher may have worked at collecting the data, unless he made some creative choices as to the selection and arrangement of material, he is not granted protection. The amount of creativity necessary is quite minimal. Some cases have held that as long as the creator uses some form of "discretion" or "professional judgment or expertise" in the creation of the compilation, he has a protectable product.[7]

Based on this information it is easy to conclude that most archival databases such as GenBank are not protected under copyright. The compilers of these databases do not choose what actually goes into them: they are completely dependent on the individual scientists for submissions. Although they make some effort to prevent bad data from getting in, this effort is minimal. The are no set criteria for the contents of GenBank (and most other databases): GenBank contains all submitted ESTs with no selection process. Other concetrate on peculiar phenomenon, like the Alternative Splicing DataBase (ASDB) developed in LBL. The little bit of curating work which is done for each entry of GenBank, is done by the contributing scientist in deciding which journal articles to link the entry to, and what annotations to include. Thus the large public databases like GenBank are not copyrightable.

The matter is quite different in relation to proprietary databases developed by bio-tech companies. The main advantage which these databases have over public ones is that they are well maintained and well annotated, to the extent that some of the products currently available from the bio-tech sector are just collections of the same sequences which are available in the public domain, but with better annotation.[12] In creation of an entry in one of this proprietary databases, the creator has to choose which journal articles are relevant, and what other genes are related to this entry. Further creativity is required in the decisions of what to do with conflicting data: because different researchers call the same things with different names, GenBank contains some number of entries which appear different, but in reality are the same. Resolving which one to follow, especially in cases where due to experimental error there are slightly different sequences is a task which requires "professional judgment and expertise," and is almost certainly within the realm of copyright.

VI.1.2 Advantages of Copyright

Perhaps the main advantage which copyright has over the other methods suggested is its age. The first copyright statute was passed in 1790, and since then the law has become very uniform, making it easy to predict the decisions of the courts in the great majority of cases. Furthermore, by not protecting the large archival databases, we would guarantee the continuation of this source of free data, for both non-profit institutions and the industry. The databases which the bio-tech firms create would carry some form of protection against direct copying. The actual extent of this protection will be quite minimal, and is discussed lower.

Another significant advantage is the flexibility that the fair use exception provides for academia. According to the 1976 act the four main considerations in deciding whether a use is fair are

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.

Let us assume that a scientist has created an infringing copy of a proprietary database. In any trial he will most likely prevail on point 2, because compilations have generally low levels of copyright protection. He will also lose on point 3, since the main value of the genomic databases are their size. This will leave the trial turning on points 1 and 4. If the scientist is truly not interested in creating a marketable work by using the database, then he will definitely prevail on point 1, and most likely on point 4 as well. Although the author could argue that there has been harm to market just because he was not able to sell a copy of the database to the scientist, it is very unlikely that the scientist would have bought one, and the courts have often dismissed such arguments off-hand. On the other hand if the scientist works for a lab which is sponsored by a pharma, or goes on to start his own company, the decisions on points 1 and 4 would most likely reverse. If every single such lab, even if it is ostensibly a part of a university, were to get a fair use exception, the market would disappear since the pharmas would be sponsoring research in labs rather than doing it in-house. A researcher who starts his own company would quickly lose the non-profit protection, and lose on point 1, with point 4 following soon thereafter. Thus the fair use exception to copyright would provide the necessary protection for true academia.

VI.1.3 Drawbacks of Copyright

There are many reasons why copyright, a set of rules originally developed for print media three hundred years ago, is not the ideal protection for the information media of today. Since the underlying information, the actual genes which are in the database, is usually not protected by copyright or any other law, an infringer who was interested in creating a competing product could modify the database in some ways and legally resell it. The actual amount of change required to create a non-infringing copy is not clear, but the amount of effort which would be involved will certainly be smaller than the effort of the original author. Because genomic databases require frequent updates to keep current, if it were possible to deny access to those updates to the competitor, he would quickly be out of business. However if copyright is the only form or protection used, this is hard. The harm to market would be significant, and this fact alone almost makes copyright unusable.

Further, the reliance on fair use for the protection of academia, although a viable idea, has its drawbacks. Although the fair use law is generally well-settled, it is still

dynamic, and until a clear line is drawn between true academia and the researchers just masquerading as such, no scientist would feel completely secure using one of these products, for fear he were prosecuted as an infringer.

Finally, the first sale doctrine of copyright law makes it hard to try to reap profits from some later use of your product. It also makes it hard to create differentially priced products for different markets, since there always exists the possibility of re-sale to another market. Such sales, according to the L'Anza case, are not infringing.

VI.1.4 Summary

Although copyright has some of the basic principles needed to protect both the users and the authors of genomic databases, any advantage gained by using this protection is outweighed by the many disadvantages. It would probably be preferable to take the positive aspects found in copyright and export them into another regime, rather than try to rework copyright from within.

VI.2 Federal Database Protection

Currently there is no database protection available from the federal government. This, however, could change. During the past three congresses a database protection bill has been introduced; there is one currently pending. This section will explore the effects such legislation would have on our topic.

VI.2.1 Overview

HR 354, introduced during the current session, provides for prohibitions against the misappropriation of data from a "collection of information." If passed, it would provide databases with much stronger protection than what is currently available from copyright law. It would prohibit any person from using

> in commerce, all or a substantial part, measured either quantitatively or qualitatively, of a collection of information gathered, organized, or maintained by another person through the investment of substantial monetary or other resources, so as to cause harm to the actual or potential market of that other person

At the same time, the underlying data is not protected: anyone is free to generate it through some other means.

VI.2.2 Advantages of Federal Database Protection

Although such database protection would lack the history of copyright law, it would incorporate many of the positive features found in the 1976 Copyright Act. There is a scientific exception, allowing the use of such collections of information for non-profit or research uses, as long as there is no harm to the market. Unlike copyright law, it would stop a competitor who uses a company's database to create his own competing product while introducing enough changes to escape the requirements of the 1976 Act. In this way HR 354 would eliminate the most significant of the problems with copyright law. It would also have the advantage of uniformity across the many jurisdictions, also found in copyright law, but lacking in the following solution. All of this combined make this an attractive option.

VI.2.3 Drawbacks of Federal Database Protection

The main drawback of stringent database protection is that it could easily impede the flow of free data within the scientific community.[10] A cash-strapped scientist might sell the genetic sequences he finds in the open market, instead of depositing them with a public database. This would lead to a drop in the number of submissions to GenBank and similar databases, making them less useful, and thereby causing even fewer submissions. Potentially these free providers of basic information may completely disappear. At the same time, the individual researchers will not be able to make enough money to subscribe to the large proprietary databases which they are helping create, possibly leaving the scientists without any source for genomic data.

Another drawback is the newness of the legislation: it is unclear how broadly the courts will interpret the scientific and the fair use exemptions provided for in section 1403 of HR 354. Until these are decided, there will be an unclear area which could serve as a disincentive to the use of certain tools by researchers.

VI.2.4 Summary

Federal database protection is an interesting solution to the problems currently faced by the bio-tech industry. Its main drawback is the potential for the destruction of public sources of information. Thus if it were possible to add additional provisos, making sure that research done at non-profit institutions stays in the public domain, it would be a solution which would satisfy almost all interests: there would be large amounts of public data, which the

bio-tech companies could curate and annotate to create their product. These could be given back to the academia under the scientific exemption, while being sold to the industry.

This rosy world, however is hard to get to. The latest trends have been to encourage non-profit institutions to commercialize their products. If all of the products of the research done in the public institutions stayed in the public domain, the large pharmas would see donations to the universities as less worthwhile. Since government scientific funding in the post Cold War era has decreased substantially, the lack of industry funds may be a heavy blow. Thus the database protection legislation, while solving some problems, causes others. Whether to adopt it is a very difficult question, well beyond the scope of this paper, but in the case of genomic databases the risk created to the free databases currently available make the current version of this legislation unacceptable.

VI.3 Licenses

While the previous two solutions described depend on Federal law, those covered in this section mainly use state law. This, as will be shown, has both some significant advantages and some major disadvantages. The methods covered in this section will be licenses with an up-front fee and reach-through licenses. It is under licenses that genomic databases are most commonly protected today.

VI.3.1 Licensing with up-front fees

Under this licensing scheme the user pays a set amount to the provider of the database. The provide, in turn, does not concern itself with how the user actually exploits this database, as long as he does not violate his license. Any product which the user creates by using this database belongs solely to the user. While this is a method which is very satisfactory for large pharma which have immediate cash to pour into such a license deal, a start-up is much less likely to have the funding to create such an agreement. The same is true of research institutions. At the same time the provider of the database may not be willing to license the database to non-profit institutions for reduced fees because the database may eventually find its way into a commercial use through one of the ways discussed in section V. The maker of the database would be a client short, while at the same time he has reduced the value of his product to other customers. Thus licenses with up-front fees are

maleficent for the academia, while good for the industry. This is generally the current state of affairs in the industry.

VI.3.2 Reach-Through License Agreements (RTLAs)

These agreements involve the party which receives the database to sign over some of the rights or royalties in any product developed using it. By giving away those rights the company can significantly reduce its immediate costs, allowing a product to pay for itself. RTLAs also are much more attuned to the interests of academia: if the product is being used in a research project with no foreseeable commercial value, the scientist has nothing to fear from such a license, since he will never have to pay a penny more than the initial fee. At the same time the provider of the database has a guarantee that if his database is later used for a commercial purpose, he will be duly compensated. RTLAs, however, face a major problem in their enforcement. It is very hard to prove that a specific database was used in order to help in the discovery of a certain drug. Another problem is the potential emergence of an anti-commons, where no one can develop a certain product because too many different entities own different small chunks of it.

VI.3.3 Advantages of Licenses

Licenses have the greatest advantage in their flexibility: the license could, potentially, contain the precise wording that the two parties involved want it to contain. Thus if one party is concerned that the other will do something that will decrease the future market value of its product, it can add specific language clearly specifying that certain actions are prohibited. Further, by choosing the exact type of license to get (up-front fee or RTLA) a cash-strapped institution can obtain access to good databases.

VI.3.4 Disadvantages of Licenses

The major pitfalls of licenses has the same origin as the main advantages. Because the licenses come in so many shapes and sizes, they are hard to negotiate. While a large pharma would not mind using a staff lawyer to negotiate for a good contract with a bio-tech database provider, a small non-profit institution would not have the resources to do this. As a result it will end up getting a worse deal than it could have. In a large university this would not be a problem, since it would have on-staff lawyers as well. These universities, however, have their own agenda. They are interested in maintaining their endowment, and

may not be as willing to assign all rights in a future invention as the scientist who is almost sure that no such invention will ever take place. As a result, the scientist may not see the database delivered until he has started working on another project, and it is no longer of interest to him.

VI.3.4 Summary

Licenses tend to be a reasonably efficient way to manage the distribution of rights in genetic databases. Although they do not respond completely to the demands of academia, it is possible that the process of getting a license could be significantly simplified through the creation of a uniform license for non-profit institutions. Although such licenses have been developed, they are not currently in use.

VII Conclusion

The design of the best legal scheme for the protection of genomic databases is a hard problem with no clear solutions. This problem is further complicated by the haziness of the public-private divide. If it were easier to separate the true academic users from the industry, it might be possible to draft better-phrased licenses. It would also be possible to let the scientists rely on fair use or scientific exceptions in case of the passage of the Database Protection Act. Since this is not possible, it seems most prudent to continue using licenses. Although clearly sub-optimal, they provide the bio-tech companies with enough flexibility to be able both to sell their product to the industry and to satisfy some of the needs of academia.

References

1. Cohen, Jon: *The Genomics Gamble*. Science Magazine 1997 275: 767-772

2. Dubchak, Inna, Research Scientist, Center for Bioinformatics and Computational Genomics, Lawrence Berkeley National Laboratory. Personal interview

3. Heller, Michael A. and Rebecca S. Eisenberg: *Can Patents Deter innovation? The Anticommons in Biomedical Research.* Science, 280: 698-701.

4. *Intellectual Property Rights and Plant Biotechnology: Proceedings of a Forum held at the National Academy of Sciences.* National Academy Press, 1997

5. Lemley, Mark A. and David McGowan: *Legal Implications of Network Economic Effects*. 86 California Law Review 479 (1998)

6. Marshall, Eliot: *Intellectual Property: Companies Rush to Patent DNA*. Science Magazine, 1997 275: 777-780.

7. Maurer, Stephen: *Raw Knowledge: Protecting Technical Databases for Science and Industry*. Report for a National Reasearch Council's Workshop.

8. Petrov, Sergey, Leader of the Bioinformaion Systems Group, Life Sciences Division, Oakridge National Laboratory. Phone interview.

9. *Primer on Moleclar Genetics* US Dept. of Energy, Office of Energy Research, 1992

10. Reichman, J and P. Samuelson, *Intellectual Property Rights in Data?* Vanderbilt Law Review Vol. 50 p. 51.

Websites

11. Pangea Systems Inc http://www.panbio.com

12. Incyte Pharmaceuticals, Inc. http://www.incyte.com

13. Acacia Biosciences, Inc http://www.acaciabio.com

14. Celera Genomics Corporation http://www.celera.com