

Bioinformatics: Scope of Intellectual Property Protection

Raguvaran Gopalan[†]

National Law School of India University, 7201 Nagarbhavi, Bangalore 560 072

Received 23 July 2008

Bioinformatics is a new field of science which marks amalgamation of one of the oldest areas of research and deliberation in human civilization, life sciences, with one of the latest and still largely developing areas, information technology. This paper is a study of this field of bioinformatics and scope and application of intellectual property rights (IPR) to this area. The paper seeks to understand basics of bioinformatics, through an examination of varied definitions on offer. Various streams of research, which constitute, or rather create a need, for bioinformatics are looked upon. The paper examines trend of growth in this field, and analyses factors which necessitate contemplation of applying IPR. Objections that have been raised to the application of IPR to the innovations and inventions in this field are also covered. Finally, the paper makes an assessment of the exact nature of IPR which will best suit the field of bioinformatics.

Keywords: Bioinformatics, database, software, copyright, patent, trade secret, genes, molecular biology, computational biology, data mining

The most rudimentary definition of Bioinformatics is any use of computers and computer based technology to handle biological information.¹ Most other definitions tend to be more or less synonymous. The National Centre for Biotechnology Information defines it as field of study in which biology, information technology and computer science merge together to form a single discipline.² Most other definitions are on the line of use of computer technology to analyse or understand biological information.³ It is the use of computer programs and other attributes of computer science to manage, catalogue and access the vast realms of biological information available. A look at the historical background of this field and its evolutionary pattern, and the circumstances which led to its genesis might aid in better comprehension of the exact nature and importance of this area of study.

With the analysis of the structure of DNA by Watson and Crick in 1953, the field of molecular biology, i.e., the study of the internal makeup of cells and the components therein, namely nucleus, DNA, RNA, genes and the other basic components, has been developing and burgeoning at an exponential rate. The amount of information created from these forays into the world of molecular biology has also been increasing simultaneously. There arrived a point in time when it became important to devise methods to

manage all the information gleaned so far and to arrange and catalogue them in a manner that may facilitate their use. These research forays produced such great amount of information and there were circumstances when the actual utility of most information was unknown, but needed to be preserved for future and possible use. It is at such a juncture, the use of computer technology to store and catalogue the data began. Initially, the extent of computer science and its involvement in the nascent area of molecular biology research was merely to the extent of data management and cataloguing. Later, the need to access information from these databases led to the development of mining software.⁴

It is this explosion of information resulting from research in the area of molecular biology and genetic sciences that led to the creation of bioinformatics. It is precisely these historical circumstances which have led to a profusion of definitions that tend to term bioinformatics as a convenient marriage between life sciences and computer technology. Bioinformatics, however, has evolved with time into something more than a mere library catalogue of sorts.

Bioinformatics, Definition and Need For

It has already been mentioned that the explosion of raw data made available through decades of laboratory experiments on genetic structure has necessitated the need for genesis and development of bioinformatics. However, to present a clear and

[†]Email: raguvarangopalan@gmail.com

unambiguous picture of various fields of bioinformatics, it is imperative to state exact nature of these discoveries and multiple avenues of exploitation of the same. On a broader level, it is hoped that information will enable discovery of cures for various diseases and also help understand reasons for incidence of diseases at a genetic level.⁵ A short diversion into the field of genetics and molecular biology is required to comprehend the nature and significance of discoveries.

The most fundamental building block of human body is the DNA which is the material whereby genetic traits are transmitted from one generation to the next. DNA is composed of repeating units, nucleotides. Genes are comprised of DNA.⁶ DNA is expressed through the arrangement of a particular subunit called the nucleotide. There are four different nucleotides which come together to form the DNA. These nucleotides are expressed in various combinations throughout the structure of the DNA molecule. Genes, being comprised of DNA are also expressions of various combinations of these four nucleotides. The tremendous variety of genes which determine wide range of traits of species is through various permutations and combination of these four basic nucleotides.⁵ In effect, all genes are repetitions of the four basic nucleotides, adenine (A), guanine (G), cytosine (C), and thymine (T) in umpteen combinations. These genes instruct the body on the creation of proteins that maintain cellular structure of the organism and direct function of the cell.⁵ Many diseases which plague the human body are due to defective enzymes or due to some other defective functioning of the body systems. Proteins are the constituent elements of enzymes.⁶ Therefore, understanding the nature and composition of a protein would help in understanding the nature and the molecular reasons for incidence of diseases. Proteins in turn are determined by various genes. Therefore, at the most fundamental level, an understanding of genes might help further understanding of human diseases and possible cures.

The true extent of discoveries in this field is mind boggling. The human cell is estimated to have close to 100,000 genes and each gene is specific and unique sequence of DNA. These genes produce proteins which control all major functions of life. Thus, there is a definite need to have some technique or application which would primarily aid in the cataloguing and sequencing of all this abundant

information. At a secondary level there is a need for applications and technologies that would help in the study of individual genes and their functions.⁵

The primary need mentioned above is the reason for genesis of bioinformatics, defined in the strictest sense. The secondary need has led to the development and proliferation of applications and software that aids in the analysis of the data collected thence far. This field, called computational biology tends to be treated synonymous with bioinformatics. Most writers on the subject use the term bioinformatics to cover applications, both for the basic cataloguing and management of data; and for the analysis and comparison of data. In the latter sense, bioinformatics has grown to include analysis and interpretation of various types of gene related data, including nucleotide and amino acid sequences, protein domains, and protein structures, though this is the original mandate of computational biology.^{2, 5, 6}

Bioinformatics: Is there a Need to Protect IT under the IPR Regime?

The employment of these new technologies has led to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases.² The possibility of the discovery of drugs and cures based on genetic studies that may have the potential to treat diseases hereto considered incurable, has meant investment of huge amounts in the research and development of bioinformatics tools of a more sophisticated, nature which would aid in the field of comparative genomics and such.^{2, 5} In this scenario there has already been a copious amount of funding into the research and development of bioinformatics tools and applications, by both public institutions and private organisations (pharmaceutical and software companies).⁵ It is only natural to expect some form of legal framework of protection for the innovations in terms of new bioinformatics tools, which would ensure a return on the investments secured from marauding interests. This is the argument for the application of IPR to the field of Bioinformatics, at the most basic level. To put it in the words of Abraham Lincoln, 'the patent system added the fuel of interest to the fire of genius'.⁷ IPR are seen as the catalyst of scientific, technological and economic development. In terms of grant of patents to an innovation, the creation and grant of exclusive and legally enforceable rights ensure that the innovator is awarded. These rights also vest the innovator with

authority to market his product, and exclusivity gives him a natural market advantage. Therefore, the patent system ensures that the innovator has the opportunity to gain revenue at the level of basic marketing and selling of the product, the profits which can possibly be accrued through the market advantage, and further ability to gain revenue through licensing of technology or product. This framework of exclusive rights of exploitation coupled with economic gain acts as a motivation to the innovator and others to further innovate. This is the fundamental economic argument for the perpetuation of IPR. The same can be translated into other streams of IP such as, copyright, in terms of protection of creative and original expressions, trademark, in terms of protection of marks which reflect the consumer trust and quality of a product.

This argument stands true for any kind of innovation in the area of research and therefore can be extended to the field of bioinformatics. Technology in bioinformatics comprises mostly of programs and software which would aid in the compilation and updating of extensive databases of information, databases themselves, and also includes software which aid in the retrieval, analysis or comparison of relevant data. Seen in this light there is no reason whatsoever to not extend the protection accorded by IPR regime to the field of bioinformatics. The protection of innovation would incentivise further innovations, which would only aid the pharmaceutical industry further in its pursuit to discover new cures for diseases.

However, there have been objections raised against the extension of IPR protection to the field of bioinformatics on the grounds that the protection would act to enclose the commons in the very ethically sensitive realm of human genome and gene related studies. The introduction of a profit motive in this realm of study, it is felt would tip the ethical balance in an unfavourable manner. The realm of human genomic science, it is contended is one that should be common and accessible to all humanity and should not be truncated and severed and owned absolutely by a few individuals by virtue of being the first claimants. This strain of argument is further developed wherein it is contended that the main motive behind innovations in bioinformatics is to further medical treatment and rule against patenting of other medical procedure should also apply here.⁶

To arrive at a well thought out and reasoned conclusion about the relative merits of the arguments

advanced by the pro-intellectual property regime protection and the pro-commons camp, it is necessary to understand the actual scope and application of the current IP regime to the field of bioinformatics and efficacy of protection that can be accorded.

Scope of Intellectual Property Protection

Scope of IP protection for bioinformatics can be split into major components: bioinformatics tools in the form of databases and compilations of raw data, and bioinformatics tools in the form of particular software which aid in the retrieval and analysis of data.² Some theorists tend to list DNA sequencing or isolation of DNA sequences as a component of bioinformatics.⁶ The researcher however disagrees with this latter grouping. Bioinformatics by definition is the application of computer technology such as software and programs for the compilation and cataloguing of data. The isolation of specific sequences of DNA or the human genome is not a part of bioinformatics proper. It is in fact a part of molecular biology research. It is a research which provides raw data, to tackle which bioinformatics tools are required. The mere coincidence that computer based technology is used in the isolation of DNA sequences does not operate to make it a part of bioinformatics. It is in fact a part of computational biology.⁸ This paper does not address to the question whether there is scope for the protection of IP involved in the isolation of DNA sequences, rather it addresses to the possible protection of innovation in the field of computer databases; and data-mining computer programs and software.

Database

A database in the context of bioinformatics is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. The main objective behind these databases is to provide easy access to information and facilitate retrieval of data for analysis and comparative studies.^{2, 6} The scope of protection under IP regime that can be extended to databases will be examined under the three streams of patent law, copyright law and the law of trade secrets.

A bioinformatics database is a mere compilation of information and is therefore not patentable, because the information is of an abstract nature and describes composition of DNA and RNA molecules, and is not the isolated DNA or RNA molecules themselves which when isolated are patentable.⁹

Therefore, a database might not constitute patentable subject matter as it is mere presentation of information and is not a composition of matter that involves any modicum of innovation. However, a database may be extended patent protection if it is not a mere catalogue, but is more along the lines of data processing system that has the ability to convert the raw data into a tangible result.⁶ It has been held by the Federal Circuit Court of United States that a data processing system is patentable subject matter as it involves the practical application of a mathematical algorithm, formula, or calculation, leading to a useful, concrete and tangible result.¹⁰ This interpretation can be applied to bioinformatics databases in the present instance as they are not mere compilations of data. Bioinformatics database are more often computer generated and built on basic software and provide facilities such as retrieval of data under particular specifications, operation of queries and methods to update. To this extent they are data processing systems and may, therefore, fall under the scope of patentability as interpreted in the abovementioned judgement. It can be argued that there is a tangible result in the use of these databases i.e., these databases have numerous applications; then the process of creating the database may be patentable.¹¹

Such a manipulation may be possible also under Indian law which does not allow patents in mere presentation of information or computer programs *per se*.¹² A bioinformatic database can be argued, deserves patent protection because it is neither mere presentation of information nor a mere computer program, but both combined with other operations which can be used in a number of applications.

However, such a patent is merely on the process of compiling and operating the database, namely, software or computer program. It does not extend to the data within the database, which still remains mere information and is therefore not patentable subject matter. Therefore, protection accorded to bioinformatics database by patent law is a mere token protection and does not ensure total exclusivity of the data compiled. An infringer may still utilize the compiled data to create an independent database by merely modifying the algorithm or software used to generate the database. Also, compilation of bioinformatics databases has been happening since the genesis of the field of study in early 80's. Therefore, there might be problems in receiving patent also on the grounds of prior art and non-obviousness of the claim.⁶

Copyright seems to be the most efficacious mode of protection of the bioinformatics databases. Under the United States law on copyright, protection could be extended to compilation. The US Supreme Court when interpreting the position on copyright protection accorded to compilations,¹³ held that facts are not copyrightable but compilations of facts are, provided there is a sufficient degree of originality in the compilation in terms of the selection and arrangement of terms, in terms of indices employed etc. However, under this law, copyright protection is extended only to the compilation to the extent of its original selection and arrangement and not the contents.

There are other forms of copyright protection which has been accorded to databases in other jurisdictions. The European Union Directive on protection of databases provides for the protection of the content of the database coupled with protection for the database if there is originality in the selection and arrangement of material.¹⁴ The European Union Directive is based on the rationale that a person who has made a substantial investment in obtaining, verifying or presenting the database must have right of exclusivity over it. The Directive protects against unauthorized extraction of the information or utilization of the whole or a substantial part of the database.¹⁵ According to the researcher, the best form of protection which can be accorded to the bioinformatics database would be a combination of the traditional rights bestowed copyright law on compilation coupled with the rights guaranteed under the EU Directive. Such a combination would protect compilation to the extent of its selection and arrangement and also the contents from extraction and re-utilization. The traditional rights, such as right to reproduction, licensing and publication would still apply and would be vested in the person making the compilation. The fact that most of these compilations are computer generated does not in anyway affect the copyright, because even with computer-generated documents, it is generally agreed that the end result product is an expression of author's idea through the medium of computer.¹⁶

Trade secrets are information including formulae, patterns, compilations, programs, devices, methods techniques or processes that drive an independent economic values, actual or potential from not being generally known and not being readily ascertainable by proper means by, other persons who can obtain economic value from its disclosure or use, and is the

subject of efforts that are reasonable under the circumstances to maintain its secrecy.¹⁷ A database can be protected as a trade secret if it fulfils the above two conditions. Trade secret protection, however, by its nature is best suited to a component of product which is to be marketed. For instance, if a company wants to market a particular video game, then the software or program relating to an interface within the game can be protected as trade secret if it fulfils the necessary criterion. However, in the realm of bioinformatics databases, main aim of protection is to imbue the person who made the compilation with some amount of exclusivity, when the database is marketed for access. By doing so, the owner who commercializes the database runs the risk that the information within the database will be disclosed and released to the public domain. Also, once the trade secret protection is breached remedies are in the form of damages for the breach and no injunction can be sought to stultify the actual leak of information.¹⁷

Software

Software now constitutes patentable subject matter under the US law if it produces a useful, concrete and tangible result. The Supreme Court and Federal Circuit have indicated that so long as a software program is more than a mere algorithm, the program may be eligible for patent protection.¹⁸ Bioinformatics software would therefore be eligible for the same protection as the software can be used for the purpose of biological research to produce results which are tangible, concrete and useful. The results acquired from data analyses utilizing these software applications have wide ranges of uses in medical diagnoses, to design drugs, or draw evolutionary conclusions. Under Indian law, there is no patent available for a computer program *per se*.¹⁹ However, the term '*per se*' is open to interpretation, and a computer program coupled with some hardware component may fall under the scope of patentable subject matter, provided the claim is cleverly constructed in such a manner that the patent appears to be for the hardware, but protection is claimed for the software as well, as an integral component.

There is also a trend to build or manufacture customised hardware components with specific bioinformatics software keyed in. Such hardware apparatuses will fall under the scope of patentable subject matter as a machine or apparatus.⁶

In the realm of copyright, the term literary work has been construed to include software and protection

accorded to software under copyright has been extended to human identifiable language, source code and machine readable component and object code.²⁰ Under the Indian law, software is included in the definition of literary work.¹⁶ Also computer program has been defined to include both source code and object code.²¹ Protection is extended to computer program as long as the work is an original expression of the idea of the person creating the program. However, copyright protection for computer software is not the best alternative, as the protection is same as extended to a literary work, and therefore extends only to the original expression of the idea. Therefore, it is eminently possible for a person to merely change some aspects of the object and source code to claim an independent copyright, as long as it does not become a substantial copy of the original.⁶

The definition of trade secrets includes software and therefore, bioinformatics software can be protected through the medium of trade secrets. It is usual practice for code writers to maintain the source code of their programs as a trade secret, releasing only the object code for sale or license. However, in the realm of bioinformatics software, where there is a definite desire to market the product, there is always the possibility that the trade secret may be disclosed by reverse engineering. The object code may be used to reach the source code, and once this is done, protection effectively collapses. The danger of reverse engineering also applies to customised bioinformatics apparatus which might be stripped down and each individual component analysed to understand the protected trade secret.⁶

Conclusion

Bioinformatics is a field that combines cutting edge discoveries and innovations in the field of molecular biology and genetic sciences with nascent information technology and computer sciences. The level of innovation that happens in this hybrid area of study is phenomenal. The implications that this field has for the larger world cannot as yet be completely gauged. But it is quite apparent that it has implications for the health and pharmaceutical industry. By its very nature, it is also a field of study which has the propensity to stir up ethical debates. The economic significance of the field cannot be underplayed either. The ever open chance of a discovery of a miracle drug based on genetic studies that might just be the answer to all human maladies has ensured great interest in this field. This has also made it a hot bed for

investment. Funds, both from public institutions and privately owned pharmaceutical and software giants have been fuelling research. With this high degree of investment, desire to ensure that there is an appropriate mechanism to reap the returns has implied the need to apply the IP regime to this field.

The innovations in this field of study are mostly confined to the twin areas of database creation, and development of better and ever more efficient data mining and analysing software. The best form of protection for databases seems to be along the *sui generis* protection mooted by the European Union. A combination of the traditional rights that accrue to an author under copyright law, coupled with the ban on data-extraction and re-utilization as provided under the EU Directive may be the best solution to protect the IP and monetary investment in databases.

The laws on patent and patentable subject matter have evolved enough in various jurisdictions to allow for the protection of bioinformatics software under the stream of patents, provided the object for which protection is sought satisfies certain criteria, such as, novelty, non-obviousness/inventive step and industrial applicability.

But returning, to the debate on whether or not IP protection must be extended to this field, the researcher feels that protection of databases and data-mining software is not the same as protection and creation of monopoly over surgical and diagnoses methods or inventions in other fields. Also, this field is quite different from molecular biology proper, and protecting these innovations does not amount to creating monopoly over the human genome sequence. To allay the fears about excessive monopoly sufficient requirements with respect to disclosure in this field need to be worked so that IP protection can be extended to this field, but not unequivocally.

References

- 1 http://wiki.bioinformatics.org/Origins_of_bioinformatics (26 August 2007).
- 2 <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html> (26 August 2007).
- 3 Vorndran C *et al*, Bioinformatics: Patenting the bridge between information technology and the life science, *IDEA: The Journal of Law and Technology*, 42 (2002) 93-97.
- 4 Ouzounis C A *et al*, Early bioinformatics: The birth of a discipline, *Bioinformatics Review*, 19 (17) (2003) 2176-2190.
- 5 Fernandez D *et al*, Intellectual property strategy in bioinformatics and biochips, *Journal of the Patent and Trademark Office Society*, 85 (2003) 465-69.
- 6 1 McBride M S, Bioinformatics and intellectual property protection, *Berkley Technology Law Journal*, 17 (2002) 1331-1363.
- 7 Idris K, *Intellectual Property: A Power Tool for Economic Growth* (WIPO, Geneva), 2000, 78-79.
- 8 Computational biology is a hypothesis driven research which uses computer based technology and available data to make further discoveries. For instance, in computational biology, particular software might be used to try and prove the hypothesis that there exists a slimming gene in the human genome. This may be done by using the data related to the protein produced by the similar gene in a rat cell. Here the main aim is to discover the existence of the gene, www.bioinformatics.org (26 August 2007).
- 9 Information *per se* is not patentable under most laws. Section 101 of the United States Patent Act permits patenting of compositions of matter. US Supreme Court in its decision in the case of *Diamond v Diehr* 450 US 175 held that abstract phenomenon, ideas and laws of nature and mere discoveries are not patentable. Section 3(c) of the Patent Act, 1970 states that there will be no patent granted in mere discovery of a scientific principle or a living thing. Section 3(n) states that there will be no patent in a mere presentation of information.
- 10 *State Street Bank v Signature Financial Group* 149 F 3d 1368.
- 11 Chisum D S, *Chisum on Patents(I)* (Lexis Publishing, New York) 2000, 1-78.3 to 1-78.5. McBride M S, Bioinformatics and intellectual property protection, *Berkley Technology Law Journal*, 17 (2002) 1331-1363.
- 12 Section 3 of the Patents Act 1970.
- 13 *Feist Publications Inc v Rural Telephone Service Co* 499 US 340.
- 14 Davison M J, *The Legal Protection of Databases* (Cambridge University Press, Cambridge) 2003, p. 11.
- 15 Directive 96/09/EC of the European Parliament and of the Council on the Legal Protection of Databases, <http://europa.eu.int/ISPO/infosoc/legreg/docs/969ec.html> (27 August 2007). Chalton S, Database right: Stronger than it looks, *European Intellectual Property Review*, 23 (6) (2003) 296-300.
- 16 Section 2(o) of the Copyright Act, 1957 which includes computer databases under the definition of literary works.
- 17 Risch M, Why do we have trade secrets?, *Marquette Intellectual Property Law Review*, 11 (2007) 1-75.
- 18 *Diamond v Diehr* 450 US 175; *State Street Bank v Signature Financial Group* 149 F 3d 1368; *AT & T Corp v Excel Communications* 172 F 3d 1352. The US courts, both Federal circuit and Supreme Court have held that an algorithm by itself may not be patentable but a process wherein an algorithm is a component is patentable if the process produces a tangible product. In the *State Street Bank* the term tangible product was replaced with tangible result and the protection was extended to data processing systems.
- 19 Section 3(k) of the Patents Act, 2005.
- 20 Nimmer, *Nimmer on Copyright (I)* (Lexis Publishing, New York) 2000, p. 2-51.
- 21 Section 2(ff) of the Copyright Act, 1957.