17
Patenting Applied to Genetic Sequence Information

MANUEL DUVAL* AND TZUNG-FU HSIEH

Alpha BioLaboratory, Inc., 1015 Edwards Road, Burlingame, CA 94010, USA

Introduction

Scientific investigators are usually not inclined to consider the intellectual property aspect of their research outcome. They are fuelled by their aspiration of discovery and consistent progress of their discipline. There is an understood code of ethics in the scientific community, not unlike the Hippocratic oath that applies to medicine. One of its terms might be formulated as follows: 'We scientists dedicate ourselves to the pursuit, promotion, and advancement of scientific knowledge'. The author of this review article is genuinely bound to the aforementioned statement, and believes that the vast majority of the readers of *Biotechnology and Genetic Engineering Reviews*, as well as the scientific community – including biologists – adhere to this standard. We strive to unravel the laws of nature, in order both to direct them to the welfare of humankind and to foster our curiosity. On the other hand, we practise our art solely within our community. We are therefore bound by the rules that govern it, too. Indeed, scientific endeavours are not only made possible by the commitment of motivated scientists, they also require monetary investment, provided by either public agencies or private entrepreneurship, or even both.

Regardless of the source of funding, it is reasonable to have a system that, in some instances, shall reward the inventors and the investors for the industrial use of their findings. In return for this incentive, the inventor would wish to disclose the details of his invention to the public rather than keeping it secret, therefore promoting the advancement of his discipline. This transaction is typically what the patent system is about. Its primary objectives are to prevent the concealment of useful scientific knowledge, by creating an incentive for investors to allocate resources in scientific and technological research programmes. The reward for the patent owners is the exclusive right to practise their invention. If such monopoly were granted indefinitely, it would be counter-productive with the foundation of the patent's law, which

*To whom correspondence may be addressed (mduval@alphabiolab.com)

Biotechnology and Genetic Engineering Reviews – Vol. 23, December 2006
0264–8725/07/23/317–330 $20.00 + $0.00 © Lavoisier/Intercept, 14 rue de Provigny, F-94236 Cachan cedex, France

is to promote the progress of science. Therefore, this exclusive right granted to a patent owner expires after a given range of time, depending on the legislation of the issuing country and the industrial sector. In the biotechnology and pharmaceutical industry, a patent typically lasts for twenty years.

While the patent system aims at promoting progress in science and technology, those who drafted the patent law are evidently the legislators. In addition, patent practitioners, who arguably master the idioms and jargons of law, undertake the writing of the patent application. Hence, the patent system involves the conjunction of three players: the sponsor, the lawyer, and the scientist. The sequence of events in which each of these actors intervenes in the process can vary. Typically though, in the chain of events leading to a patent application, the scientific and/or technical input comes first. Evidently, any patent originates from a creative phase. Nevertheless, while scientists are extremely meticulous in the way they communicate their result to their peers in media dedicated to them, and follow very closely the progress of their 'competitors' in their field with respect to the primacy of the publication of new results, they often tend to rely heavily on the patent practitioner when it comes to the management of the intellectual property facet of their research. While the latter is usually a specialist in a given discipline, he, however, cannot substitute the scientific contributor. His main function is to cast the scientific and technological specifications of an invention into the format and language of the patent procedure. A lack of involvement of the scientific contributor in the building of a patent application can, in some cases, lead to documents (the patent per se) that are poorly drafted (Dufresne and Duval, 2004). Indeed, patent applications do not enter the peer-review process of scientific publishing. They are reviewed by Patent Office Examiners, who in turn may not substitute the scientific expert. The bottom line is that, when published, the patent document is what ultimately determines the depth and limits of the intellectual property asset. If it turns out to be poorly written, the chances are that the actual asset could, in the best scenario, become altered, and in the worse, be rendered obsolete.

This brief review aims at providing insight regarding the management of intellectual property specifically for genetic sequences. Its scope is limited to the scientific and technical aspect, and not to the legal one.

A concise outline of the patent system

An issued patent confers ownership of an intangible asset to its owner (e.g. a method to amplify DNA). There are other means by which one can acquire ownership of such assets: copyright and/or trademark. However, these two types of intellectual property protections seldom apply to our biotechnology sector. Indeed, a copyright applies to forms of expression rather than subject matter, while a trademark applies to 'a word, phrase, symbol, or design, or a combination of those that distinguish the source of the goods of one party from those of others'. In the scope of intellectual property inventions pertaining to genetic sequences, a patent is the relevant type of protection that needs to be sought. Inventions based on genetic sequences definitely fall into the category of patentable subject matter: examples of industrial utilization of inventions based on genetic sequences are biomarker probes applied for diagnosis and recombinant protein therapeutics. By and large, there are three broad

categories of patentable subject matter: processes, machines, and articles of manufacture and use. The latter is the most relevant to our sector since it includes compositions of matter. Yet, some inventions featuring genetic sequences fall into the processes category as well (e.g. method of treating patients with a given disease through the use of a particular gene).

Since the patent system is a legislated right, it obviously varies with respect to the country where the patent law has been enacted. The main difference is that the US system awards the patent to the 'first to invent', while in the EU and Japan the 'first to file' is awarded the patent. Fortunately for inventors, the other differences between all the national patent laws are minimal enough that patent rules that apply broadly can be inferred. There are four requirements for an invention to be deemed patentable:

(1) novelty;
(2) non-obviousness (also referred to as inventive step);
(3) usefulness (industrial applicability); and
(4) enablement (written description requirement).

Non-fulfilment of only one of these then disqualifies the invention for a patent. As a result, it is absolutely critical that all the relevant information is retrieved in the course of a patentability search in order to assess the novelty of a candidate-invention for patent application. A patent application must provide a full explanation on how to practise (i.e. make and/or use) the invention(s). In addition, it has to include an explicit written definition of what has been invented: this latter section is what is referred to as the *Claims*. They define the scope of the rights that are granted to the patent's applicant. A typical claim is as follow: 'The subject matter of Claim 1 relates to the purified and isolated polynucleotide encoding the amino acid sequence of seven transmembrane receptors set in SEQ ID NO:XX possessing one ligand binding activity'. The claims section is the most critical part of the patent document: each of its words is considered an 'element' or 'limitation' of the claim. In order to exclude a third party from using an invention granted by a patent, one has to demonstrate that what it is using is identical to the claimed invention. The search for such type of information is what is referred to as 'infringement search'.

Options in managing an intellectual property portfolio

Generally speaking, five initial paths are offered to the investigator with respect to the management of a putative intellectual property asset:

(1) trade secret;
(2) trademark;
(3) copyright;
(4) provisional patent application; and
(5) patent application.

These paths are not mutually exclusive (e.g option No. 1 can be paired with option No. 2, and option No. 2 can be paired with options No. 4 or No. 5). As mentioned previously, the relevant approach to gain intellectual property rights for genetic sequence-based invention is the patent system. The only exception to this rule

might pertain to recombinant proteins used for therapeutic application, where a trademark also could be sought for protecting a brand name in addition to a patent.

Biologics, which refer to the group of therapeutic agents whose active compound is a biomolecule (as opposed to small molecule therapeutic agents), need to attain regulatory approval before entering the market. The composition of matter of such an active compound is assumed to be protected by a patent. Patent on biologics, like those for small therapeutic agents, expire after twenty years of filing. And, as with small molecules, when the patent has expired, third parties are allowed to produce and commercialize the molecules. In the case of small molecules, an organization willing to market an off-patent therapeutic agent must attest that its product is chemically identical to the compound that was initially approved by the regulatory agency. In the case of biologics (e.g. a recombinant protein used for treatment of chronic metabolic disease), the condition for distributing a generic (generic stands for therapeutic agents that are identical to off-patent molecules) goes beyond the identity of the composition of matter (Combe *et al.*, 2005). Regulatory agencies deliver approval for a generic biologic on the condition that the outcome of clinical trials proves it complies with both safety and efficiency requirements.

This regulatory difference between small molecules and biologics stems from the fact that the manufacturing of recombinant proteins is a substantially more complex procedure than that for producing small molecules. In addition, the end product of the manufacturing process differs remarkably between small molecules and biologics. The product of a chemical synthesis procedure leads to a homogeneous molecular entity: in addition to the solvent, the reagent might exist under a small number of forms (isomers) in dynamic equilibrium. For small molecules, the composition of matter disclosed in the patent and in the regulatory agency application realistically describes the actual chemical entity produced and distributed. On the other hand, this situation does not hold true for biologics. A polypeptide is evidently a highly complex system and in reality, never exists under a single entity. Various conformational isomers co-exist when a protein is produced. Also, a plethora of post-translational events occur, some controlled, some not (e.g. partial oxidation or deamination of residues on the surface).

The occurrences of these post-translational modifications depend on the production process. Analytical methods (e.g. mass-spectroscopy) could, in theory, be applied to assess the whole set of modifications that a recombinant protein undergoes upon the completion of the production process. In practice, this is hardly true. Hence, rather than describing the comprehensive set of post-translational modifications, a patent claims for the primary structure of the protein and usually for a limited number of modifications (evidently those known to directly impact the potency of the biomolecule). On the other hand, what is approved by regulatory agencies is the combination of the composition of matter (the actual protein sequence) bound to one given method of production. Hence, in the case of the bio-pharmaceutical industry, the owner of a genetic sequence patent must also design the appropriate production procedure in order to apply her/his right to make an industrial usage of her/his invention. Typically, the production protocol might fulfil the requirements for a patent application. Yet, the patent category in this latter case relates to a process rather than a composition of matter. Whether it is relevant to protect such invention by a patent rather than keep it confidential (trade secret) evidently depends on the

business context and long-term objectives. In all situations, there should be an option that would provide the optimum benefit over risk ratio.

When one considers the intellectual property implications with dealing with genetic sequence-based invention, the following use case scenarios can be drawn:

(1) The composition is kept secret (trade secret):
Pros: (i) no filing cost; (ii) the information is not disclosed;
Contras: if a third party applies for a patent regarding the same invention, and this patent is issued, the initial inventor, who did not protect her/his invention, is infringing this patent and must either cease the use of her/his own invention or pay licence fees to the patent owner.

(2) The composition is published without patent protection:
Pros: (i) no filing cost;
Contras: any third party can make use of the invention and market it, the exclusivity is lost.

(3) Patent protection:
Pros: up to twenty years of exclusivity to exercise industrial and commercial use of the invention;
Contras: (i) cost of patent submission; (ii) the information is disclosed to the public eighteen months after the filing date.

For genetic sequence-based invention, the first option is obviously a poor choice if the invention's scope is to commercialize a product whose component is a recombinant protein or a nucleic acid reagent used as a probe for a diagnostic kit. If the invention pertains to a reagent designed through recombinant DNA technology used in the manufacturing of a product, the object of the invention itself is not commercialized. Rather, it is used internally as a component of a production process. Publishing its composition without seeking intellectual protection evidently would be, in this case, a poor choice. A trade secret might be advantageous only if one is not willing to get revenues by licensing its invention to third parties and/or is willing to keep a competitive advantage over the competitors over a period of time exceeding twenty years. Implicit in the last statement is that the invention is very unlikely to be re-discovered by a third party. This situation also relates to the aforementioned case regarding the process of producing recombinant polypeptides for therapeutic applications. Keeping the process as a trade secret would allow preserving the exclusivity of a biologic once the patent for the composition of the matter has expired. Competitors willing to commercialize generics would have to design their own production process.

The path to option No. 1, No. 2, or No. 3 is dictated not only by a commercial prospect. In biotechnology R&D laboratories, it is not feasible to ignore what is happening to intellectual property. Even in the event that a given organization is not willing to seek for patent protection for its ongoing discoveries, it may be unwittingly infringing a third party's patents. The technology industry as a whole is seeking more patents, particularly in the IT and biotech sectors. In Europe alone, more than 5000 patents in the field of biochemistry and biotechnology were published in 2004 (see *Figure 17.1*). More than 1.4 million DNA sequence records and 800 000 polypeptide entries are featured in patent documents (Xu *et al.*, 2002; Rouse *et al.*, 2005; Verbeure *et al.*, 2006). Meanwhile, the volume of patent

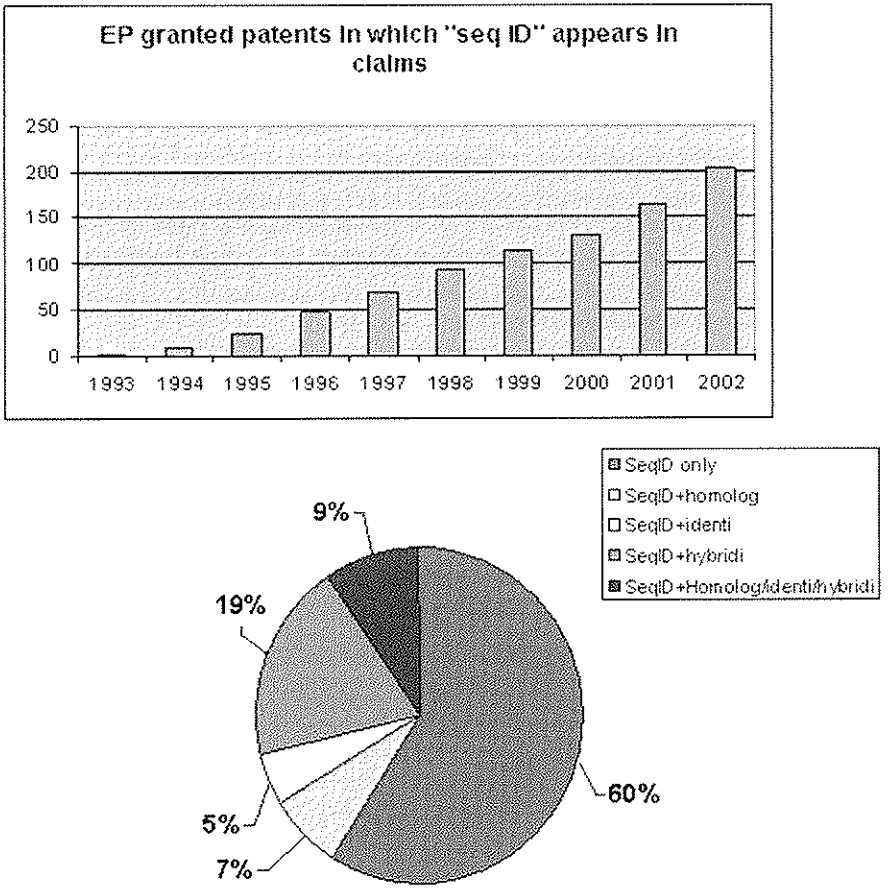

Figure 17.1. The proliferation of patents claiming genetic sequence-based inventions. Between the years 1993 and 2002, the USPTO (US Patent and Trademark Office) and the EPO (European Patent Office) issued more than 14 000 and 850 patents, respectively, with one or more claims referring to SEQ ID numbers. These patents include claims that define the covered sequences in terms of, inter alia: per cent identity to an enumerated protein or nucleic acid sequence (a 'working sequence') or a nucleic acid coding sequence (or protein encoded by a sequence) that hybridizes under particular conditions to a working sequence.

applications submitted to the patent offices and comprising genetic sequence information steadily increases, while showing no signs of decline. Hence, it makes sense that in any biotechnology R&D organization, the status of every genetic sequence involved in ongoing research programmes needs to be assessed and tracked during the advancement of the programme. No matter what the business prospects and strategy a given institution or corporation wants to achieve regarding gene patenting, there is a primary issue that needs to be addressed: the access to relevant information.

Patent system and Genomic both rely on complex concepts and inference workflows. Genomic is still an emerging scientific discipline and has had a difficult time in disciplining itself regarding standards and data management systems. Progress has been made in achieving a community-wide information system. When intellectual property (IP) investigators conduct their studies, they reap the benefits from

technologies implemented for the scientific community. Indeed, both sectors have common needs. However, IP research has its own requirements, meaning that methods used for *in silico* biology research cannot be translated *in extenso* to *in silico* IP research. Even though the primary data source could be the same, IP investigations rely on attributes that do not always overlap with scientific attributes. Additionally, IP inference relies on its specific workflow. On the other hand, standards that have been designed recently by the genomics community could benefit IP investigators. To mention only one, the Gene Ontology system provides a powerful method for assigning functions to genes. This classification represents the foundation upon which information systems are designed to map and exchange data across genomic databases. The last section describes the rationale of a methodology for prior art search regarding genetic sequence disclosed in issued and granted patents.

Claims on DNA and polypeptide sequences

A system featuring some genetic sequence information is patentable under the condition that a function has been assigned to it. A gene patent is built, therefore, on two types of information: (1) a sequence; (2) a function. Both of these attributes are required for the validity of the claims. The prosecution of patent genes would have been an easy task if nature did not bring elements of complexity to this matter. These are of three types: (1) a given biological property can be achieved by two non-related genes; (2) on the contrary, a given gene can produce two distinct biological effects; (3) the genetic code is degenerate and some mutations at the amino acid level are conservative, meaning they have either no or little impact on the biochemical function. As a result, there is not a simple one-to-one relationship between a given function and a genetic sequence. To ensure the enforceability of issued patent genes, not only a prototypical sequence is claimed, but also a set of related sequences linked to the former one. Taking into account that derivatives of the sequence disclosed in the patent application can achieve the same function, gene patent applicants claim for a virtual collection of closely linked sequences. Typically, the link is based on a given threshold of identity value (Dufresne *et al.*, 2002; Dufresne and Duval, 2004). Such a general approach to describe the set of genetic sequences able to achieve the same biochemical function reflects our current lack of method to predict which derivatives of a given polypeptide are associated with the same function. Let us say that an investigator has designed a new recombinant DNA construct coding for a polypeptide able to synthesize RNA ten times faster than any naturally occurring DNA-dependent RNA polymerase. This is evidently new, the approach to the design was non-obvious and the utilities are numerous. The investigator then fills out a patent application where she/he discloses the sequence of the new, engineered RNA polymerase. The same investigator, being a protein chemist, is perfectly aware that a similar catalytic activity could be achieved by modifying the composition of the protein she/he claims on her/his patent application. Theoretically, she/he should disclose the sequences of all the isoforms having the same biochemical function within a certain range of kinetic values. However, it is currently beyond our reach to achieve this kind of prediction. In addition, even in the hypothetical scenario where such prediction could be made, it is likely that the number of sequences in the prediction set would be fairly large, hence unpractical to disclose, as such.

To summarize this last point, the ideal scenario for disclosing genetic sequence-based invention would be as follows:

(1) write an instance of the set of the sequences;
(2) write the associated biochemical and/or biological function;
(3) write the range of parameters' value assigned to the function;
(4) write the algorithm that enables to retrieve all instances of the genetic sequence capable of achieving the function aforementioned.

One could see this scenario as a common objective for the patent offices and the biotechnology industry. However, current and legacy genetic sequence data featured in patents are mainly defined according to a percentage of identity with the sequence specified in the claim. This, in turn, has some implication with respect to prior art search (prior art search refers to the process of assessing the novelty of an invention) (Dufresne *et al.*, 2002).

The databases of genetic sequences contained in issued and applied patents are populated with sequence records to which are assigned an implicit domain of connected sequences. This last information can be viewed as one meta-data attribute for each record. Yet, in order to carry out patentability and/or freedom-of-use and/or infringement searches, these databases have to be interrogated with a search engine capable of retrieving all instances that reach the identity level with a given queried sequence. This search engine has to rely on an algorithm, which would compare the whole query sequence with the database records and find matches based on the criteria of identity. Approximate string matching algorithms provide the means to address this question (Navarro, 2001). DNA and polypeptide sequences are simply string literals. Sequence databases are large amounts of string characters data stored in a linear form. The query itself, the DNA template and/or the polypeptide, can be seen as a word or a phrase. Hence, why not run a straightforward literal search to uncover elements of the database that could match the query, and so assess the uniqueness of the invention? For the very reason, as mentioned earlier, that the patents featuring genetic sequences are not only claiming one sequence, but also a set of closely related ones. In addition to claiming the exact disclosed sequence, a set of neighbour sequences are also claimed, based on the principle that variations on one given genetic template can lead to the same function. Nevertheless, from a data administration standpoint, databases of patented genetic sequences are only filed with the actual sequences specified in the patents. The additional claimed variant sequences represent a virtual data set, which is not literally recorded in the database. This is only an attribute of the sequence records, therefore rendering a literal search method inadequate for querying these databases.

Querying databases of applied and issued genetic sequence means retrieving occurrences of the database that either completely match the query or that are related to it to some extent. This problem is reminiscent of the issues of effective retrieval of information in online databases and of text processing. The way to address such questions is by implementing string-matching algorithms (Navarro, 2001). String searching consists in finding one or, more generally, all the occurrences of a string (called a pattern) of length m in a text. Moreover, at least two additional specifications are needed in the context of genetic sequences: (1) gaps have to be allowed: they represent insertion and deletion events; and (2) permutations of residues are not

allowed because they obliterate the significance of a sequence. Approximate string matching involves searching a textual database for strings that are similar, but not necessarily exactly identical, to a given pattern string. The definition given by the National Institute of Standards and Technology is as follows: 'searching for approximate (e.g. up to a predefined number of symbol mismatches, insertions, and deletions) occurrences of a pattern string in a text string'. The use of the term 'approximate' simply emphasizes the fact that a perfect match may not be achievable and that imperfections such as missing and extraneous symbols have to be considered. The primary principle is the concept of string edits distance, a measure for quantifying the similarity between two strings.

Approximate string matching traditionally has been used to help with the problem of retrieving information while spelling variants, misspellings, and transliteration differences are populating databases. In addition, it has been applied already in the field of computational biology to address the need of finding the optimal solution in pair wise DNA sequence alignments (Landeau *et al.*, 1986). A program, referred to as KERR, has also been proposed to implement efficient approximate string matching (Dufresne *et al.*, 2002). Given two sequence data, KERR performs a pair wise alignment to return the maximum match between the two compared sequences. Typically, two kinds of optimal alignments can be deduced from any given pair of sequences, keeping in mind that one of the sequences from the pair is the subject sequence, the second one being the query. One method will find a sub-string out of the query sequence that gives the best fit with a subject sequence: this is referred to as local alignment (e.g. BLAST, Altschul *et al.*, 1990), as opposed to a global alignment algorithm whose purpose is to align the whole query sequence to the subject (e.g. Needleman and Wunsch, 1970). Intellectual property relevant sequence queries rely on a global alignment algorithm output: for every queried sequence, the algorithm retrieves any database records that (1) either entirely contain the whole query; or (2) are contained in the whole query; or (3) have one overlap with the whole query. In other words, the query object is indivisible for the outcome of the search, as the question is not to find portions of the query that may fulfil the identity threshold criteria with one database subject, while the global alignment of the query sequence with this subject database would return an identity value smaller than the threshold. The schema in *Figure 17.2* illustrates this concept. KERR performs a global alignment with respect to the query sequence. Database records that are retrieved are those that align with the query sequence with an identity greater than the input value threshold.

In addition to retrieving all subject databases that fulfil the identity threshold requirement based on the whole query sequence, KERR returns the value of per cent identity between the query and the subject with respect to the whole sequences of both. This functionality is intended to assist in decision making for deciphering specific cases. To mention only one, eukaryotic genes have the tendency to encode polypeptides longer than prokaryotic genes do (Brocchieri and Karlin, 2005). Some eukaryotic proteins include several domains, linked together in a single polypeptide. For instance, the human liver acetylCoA carboxylase enzyme has two catalytic domains plus a domain bound to biotin. On the other hand, the *Escherichia coli* acetylCoA carboxylase is a multi-component enzyme containing a biotin carboxylase subunit, a biotin binding subunit, and a carboxyltransferase

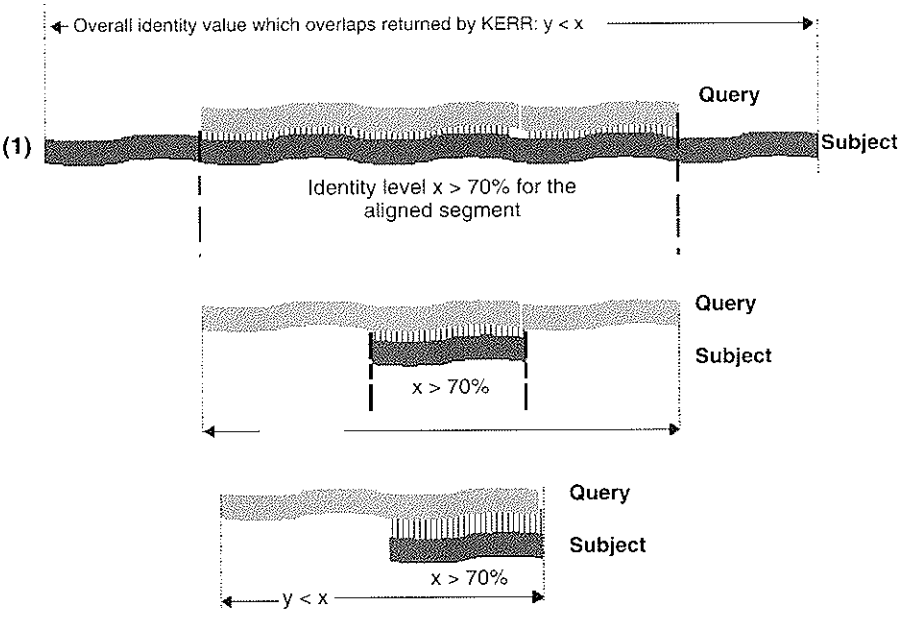

Figure 17.2. Schematic representation of the three possible KERR outcomes. KERR first retrieves all occurrences in the database that align globally to the query and from which a per cent identity value is greater than the threshold (e.g. >80%). In addition, KERR returns a second identity value, referred to as *y* in this schema, that takes into account overlaps either for the query or for the subject. Gaps are allowed in the aligned segment, as long as it does not lower the identity value below the threshold.

(transcarboxyłase) subunit. Querying with one *E. coli* gene (e.g. with the biotin binding subunit) with KERR will retrieve the mammalian acetylCoA carboxylase gene, as its biotin binding domain shares more than 70% identity with the *E. coli* protein. KERR will also return the overall per cent identity, taking into account the totality of the subject sequence, which evidently lowers the per cent identity score. Both values are relevant to the investigator to decipher, in the light of the patent claims, the status of her/his invention. The central point is that KERR retrieves the relevant database subject entries in the context of a patent search, and in addition, delivers values that pertain to each case.

It is worth mentioning that BLAST offers different options for running a search, most notably:

(1) the scoring matrix;
(2) the cost to open a gap;
(3) the cost to extend a gap;
(4) a word size for running the indexation of the database;
(5) cut-off scores for pursuing an ongoing alignment;
(6) filters for low complexity regions.

The reason BLAST features so many different options is that searching for homologues is dependent on the context. Whether the search is conducted against either a genome sequence database, a cDNA sequence database, or a protein database; whether the database features all representatives of all major phyla and one is aiming

to detect as many homologues from various species, a set of different parameters will be picked for maximizing the chance to retrieve candidate homologues. BLAST runs its search using a scoring model and eventually returns a per cent identity value afterwards, once the local alignment with the highest score has been obtained given a set of parameter values used for carrying out the run. Hence, for each set of parameters used, a different outcome is returned by BLAST. The 'take home message' is that BLAST, as well as its related algorithms, namely Smith–Waterman and Needleman and Wunsch, are designed to query a database for homologues, and more importantly, for inferring functions to genes whose only data available are DNA sequences. These algorithms were not intended to be applied for patentability searches for which the only characteristic of identity threshold is the criterion for finding matches in genetic sequence databases.

Conclusion

What do companies like IBM, Bayer, Genentech and/or Agilent have in common? They all rely on their intellectual property asset for doing business. The biotechnology industry does not differ from its high-tech industry counterparts in this respect. It belongs to what is referred to as the 'knowledge-based economy'. This latter economy has been encouraged by the statement of the European leadership during the 2004 European Union Summit in Lisbon that it is a 'key factor in growth, the competitiveness of companies, and employment' (Communication from the Commission, 2004). Knowledge-based economy is often also referred to as the 'new economy'. What these denominations intend to convey is the shift to knowledge as the primary source of value. Intellectual asset is the major driver of business in our modern economy.

Hence, this applies to the biotechnology sector. This sector seeks protection mainly, though not exclusively, for composition of matter. In turn, this latter set mainly coincides with genetic sequence-based products. Concomitantly with the advent of the genomics era, concerns and ethical debates were raised, by both the scientific and industrial communities, regarding the proper way of assigning intellectual ownership to genetic sequence information (Davis *et al.*, 2005). Granting patents on genes and gene products is not a recent development. In 1911, Takamine obtained a US patent for adrenaline. Genes and polypeptide objects are regarded for patent purposes as biologically active substances and are, therefore, eligible for patenting on the same basis as other chemical compounds. The European Patent Office and the US Patent and Trademark Office released gene patent guidelines (Directive No. 98/44, 1998; US Patent Trademark Office, 2001) demonstrating that there are no objections to the patentability of genes based on the very strict criteria of utility, novelty, and non-obviousness.

In our industry, recent technical and scientific breakthroughs offer unequal business opportunities, as well as unprecedented tools to, for example, progress in designing new and better therapies for critical human health disorders. Intellectual property management entails assessing the novelty of research outcomes, as well as deciphering the extent to which the use of a given method and/or composition of matter does not infringe a patent. Data management is, therefore, critical. Prior art investigation in the field of genetic sequence information currently involves the

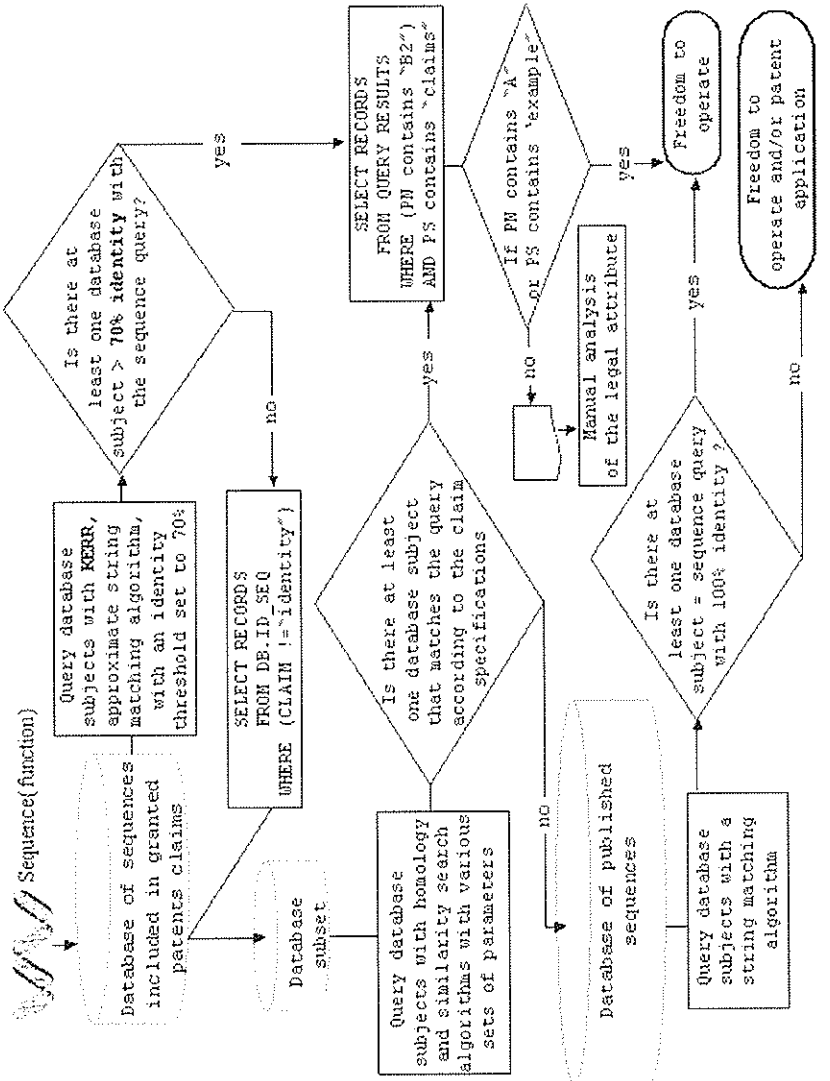

Figure 17.3. Potential workflow for assessing the intellectual property status of genetics sequence data. The PN and PS attributes point legal annotations, and stand for Patent Number and Patent Sequence, respectively.

query of several databases. A characteristic of this field is the number and diversity of players involved: academic institutions, colleges, patent offices, regulatory and governmental agencies, patent consulting companies, biotechnology companies, pharmaceutical companies, software companies, and database providers (both private and publicly funded). All these organizations include legal department managers, patent lawyers, scientists, computer scientists, bioinformaticians, information managers, managers, etc... The enumeration of these groups and skills only suffices in translating the complexity of the field, and therefore the great need to design and implement common standards and protocols.

The bottom line is that gene patenting involves a claim on a genetic sequence assigned to a given function. This, in turn, impacts the means by which patent search with respect to gene patent is conducted. A gene patent can be viewed as a data structure known as key–value pair. The key is represented by the genetic sequence, while the value is represented by the function. Many applications lend themselves to this type of data structure. Hence, this may suggest that retrieving patent relevant information for genetic sequence-based invention is an easy task. On the contrary, as discussed earlier in this review, despite some ongoing efforts within the genomic community to set a controlled vocabulary for designating gene function (e.g. the Gene Ontology), searching a database of genetic sequences featured in patent by querying the function value generates mass data and involves a significant extent of non-automated data analysis. On the other hand, searching the same database through the key, in other words the sequence data, allows the retrieval of a more accurate and comprehensive data set, which in turn can be filtered through the function value (*Figure 17.3*). Ultimately, the management of intellectual property assets for genetic sequence-based inventions depends on the combination of the relevant database search algorithms and the maintenance of biological and legal annotations.

Acknowledgements

The authors would like to thank Prof. Stephen Harding for his useful comments during the preparation of this article.

Disclaimer: the contents of this review are informational only and should not be substituted for legal advice. The views expressed herein are those of the author.

References

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. AND LIPMAN, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

BROCCHIERI, L. AND KARLIN, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33**, 3390–3400.

COMBE, C., TREDREE, R.L. AND SCHELLEKENS, H. (2005). Biosimilar epoetins: an analysis based on recently implemented European medicines evaluation agency guidelines on comparability of biopharmaceutical proteins. *Pharmacotherapy* **25**, 954–962.

DAVIS, P.K., KELLEY, J.J., CALTRIDER, S.P. AND HEINIG, S.J. (2005). ESTs stumble at the utility threshold. *Nature Biotechnology* **23**, 1227–1229.

DIRECTIVE NO. 98/44 (1998). Legal protection of biotechnological inventions. *Official Journal of the European Union* No. L213.

DUFRESNE, G. AND DUVAL, M. (2004). Genetic sequences: how are they patented? *Nature Biotechnology* **22**, 231–232.

DUFRESNE, G., TAKACS, L., HEUS, H., CODANI, J.-J. AND DUVAL, M. (2002). Patent searches for genetic sequences: how to retrieve relevant records from patented sequence databases. *Nature Biotechnology* **20**, 1269–1271.

LANDAU, G.M., VISHKIN, U. AND NUSSINOV, R. (1986). An efficient string matching algorithm with k differences for nucleotide and amino acid sequences. *Nucleic Acids Research* **10**, 31–46.

NAVARRO, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys* **33**, 31–88.

NEEDLEMAN, S.B. AND WUNSCH, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.

ROUSE, R.J.D., CASTAGNETTO, J. AND NIEDNER, R.H. (2005). PatGen – a consolidated resource for searching genetic patent sequences. *Bioinformatics* **21**, 1707–1708.

US PATENT TRADEMARK OFFICE (2001). Utility examination guidelines. *Federal Register*, January 5, pp 1092–1099. Washington, DC: GPO.

VERBEURE, B., MATTHIJS, G. AND VAN OVERWALLE, G. (2006). Analysing DNA patents in relation with diagnostic genetic testing. *European Journal of Human Genetics* **14**, 26–33.

XU, G.G., WEBSTER, A. AND DORAN, E. (2002). Patent sequence databases. *World Patent Information* **24**, 95–101.