Bioinformatics and Copyrights

Knowledge needs Open Codes[1]

by Andrea Monti - amonti@unich.it

[1] This article is a reviewed summary of the lecture on *Copyrights and Bioinformatics: Knowledge needs Open Codes* the author delivered in the Annual Conference of the *License Executive Society of Britain and Ireland* in Bristol (UK) on 24 and 25 June 2004. The author thanks Dr. Andrea Cocito and Dr. Stefano Confalonieri of the FIRC Institute of Molecular Oncology Foundation for the technical review of this article, Prof. Enrico Dainese, a professor of Biochemistry at the Comparative Biomedical Sciences Department of the University of Teramo, Dr. Marcella Attimonelli, Professor at the Biochemistry and Molecular Biology fo the University of Bari, and Dr. Paolo Vezzoni of the Institute of Advanced Biomedical Technologies of the National Research Council (CNR) of Milan, Italy.

The author can be contacted at the following addresses: Andrea Monti – studio legale Monti, 96, via Paolini – 65124 Pescara (IT) tel. 085 294255 e-mail: lawfirm@andraemonti.net

Introduction

Although the Information and Communication Technologies (ICT) are now omnipresent among the Life Science community, there is not yet a clear perception of the legal impact caused by the use of (massive) processing power in conducting genetic research. The – so to speak – traditional fields in which the ICT play a fundamental role range from sequence analysis and pair-wise alignment of proteins sequences[2] and DNA to the planning and maintenance of genetic databases, and yet to the visualization (and prediction) of the structure of proteins. Indeed, the advent of the Internet and of the open source systems has unlocked new frontiers to researchers, allowing them to share and update these tools in real time. Detailed descriptions of these techniques are beyond the scope of this article, but if we are to set bioinformatics in the adequate legal frame, we need at least a basic understanding of what bioinformatics is about and which are the difficult aspects concerning the development and management of the software used by researchers.

What bioinformatics is about

Briefly summarizing the evolution of bioinformatics, we identify a prehistoric phase (between 1950 and 1970), when research was conducted with "analogical" tools, and a historic phase that, in turn, had two stages. The first stage (1970 - 1990) was characterized by the use of computers for the automation of analysis processes, and the second stage (from 1990 on) that has been characterized by the use of the Internet as a tool for distributed management of the information (from the creation of centralized databases that permit remote interrogation and adding up of information, to the availability of analysis software operated through a web-based interface).

Bioinformatics – according to a rather laconic but efficient definition – may be defined as the computational branch of molecular biology;[3] or, from a different perspective, as the branch of information science applied to biology-related information. [4] Leaving aside for the moment the epistemological aspects (which are extremely interesting, by the way) raised by these definitions [5] we only have to understand "what bioinformatics is about", for which purpose we need to identify the operative areas of bioinformatics. Bioinformatics:

- Creates algorithms capable of analyzing biological data,
- Implements such algorithms in suitable software,
- Plans and produces databases that accumulate, analyze and share biological data,
- Identifies, plans and produces mathematical and statistical methods for simulating and predicting the behavior of biological systems.

The following examples provide a practical picture of how the research activity is carried out and – stressing the role played by the information technologies – provide a glimpse of the nature and the significance of the legal implications of this particular branch of scientific research.

---

[2] Proteins are macro-molecules composed of twenty different amino acids (in turn composed of carbon, hydrogen, oxygen, nitrogen, sulfur) held together by the so-called "peptide bond".

[3] CLAVERIE, J.M., NOTREDAME, C. *Bioinformatics for dummies* Wiley Publishing 2003, p.10.

[4] NIH working definition of bioinformatics and computational biology, July 17, 2000 - http://www.bisti.nih.gov/CompuBioDef.pdf.

[5] The same as for other subjects, (criminology, for instance, to stay in the legal area) an open epistemological debate exists as to whether bioinformatics should be considered a discipline of its own or a mere "technological" support to other disciplines normally accepted as such, as in the case of computational biology. On this subject, please see VALLE, G., HELMER CITTERICH, M., ATTIMONELLI, M., PESOLE, G., *Introduzione alla bioinformatica*, Zanichelli, 2003, p. 2.

DNA sequencing, proteins and structural bioinformatics

As mentioned in the introduction, the biological data processed by bioinformatics has mainly originated from the analysis of proteins. Since the 1950's with the discovery of insulin, it is well known that besides the specific amino acids of which they are composed, the structure of proteins depends on the **exact order** in which such amino acids are bonded.

To facilitate their research work, scientists have designed a system that associates the names of the biological data to the letters of the alphabet, thus transforming the structure of proteins into a sequence of letters, and the consequent research activity into some sort of "deciphering" an unknown language.[6] Before the advent of informatics to molecular biology, operations of comparison and – in general – sequence analysis were carried out by writing and hanging endless paper lists of letters on laboratory walls, to then match them all. Further complexity was added a few years later, in 1958, when John Kendrew and Max Perutz identified the first tri-dimensional structure of a protein and confirmed the hypothesis that while the **identification** of a protein depends on the **sequence** of the amino acids that compose it, the **function** (that is, the "task to accomplish" in the "factory of life") depends on the **shape** of the protein.

The revolutionary effect of the irruption of informatics in this setting is quite evident: as we may easily sense when looking at the "recipe" of insulin cited in footnote 6 bellow, as the characters that composed the sequences increased in number, so did the time needed for analyzing them, and the margins of mistake grew accordingly. But with the arrival of the first computers, even with processing powers that seem ridiculous today, the sequencing activity became incredibly faster and more reliable. As regards the shape of the amino acid chains as taught by structural bioinformatics, playing with the digital patterns of the proteins on computer screens became obviously much easier than handling a 3-D puzzle made up by some thousand plastic pieces.[7] Although it is true, as said before, that it is the shape of a protein that determines the function of the protein, it is also true that such shape in turn is determined by the sequence of the amino acids that compose it. In effect, two proteins with similar amino acid sequences have, in most cases, similar shapes. Since these discoveries, in recent years we have witnessed the flourishing of algorithms and computer programs attempting to foresee the structures or shapes of the proteins starting from their amino acid sequences using only computer tools (most of which are open-source and web-based).

In roughly the same period when proteins were being studied, other paths led scientists to understand the structure of the deoxyribonucleic acid or DNA, which was found to consist of four "bricks" (technically called "nucleotides"). They also discovered that it is the sequence of the nucleotides of some parts of the DNA (the genes) that dictates the sequence of amino acids that a cell assembles to form the protein that corresponds to a certain gene[8].

---

[6] The sequence of human insulin, for example, looks like this:
MALWMRLLPLLALLAWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVE LGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

[7] CLAVERIE, J.M., NOTREDAME, C., *op. cit.*, p. 15.

[8] In other words, *si parva licet*, proteins are comparable to a building, amino acids are the bricks and DNA, the building project.

Also in the case of the nucleotides, scientists associated a letter of the alphabet to each component (thus, A - Adenine, C - Cytosine, G - Guanine, T - Thymin)[9] and followed the same approach that they had when they analyzed proteins. The advantage lied in the fact that they could do it much faster because they had to "handle" only four instead of twenty letters (that identified the amino acids that composed the proteins). [10]

The point of contact between the study of DNA and the study of proteins became evident with the discovery, as mentioned, that the "assembly instructions " of the amino acid sequence that compose the proteins are contained in DNA. Then, the scientists thought, if a way could be found to link the four DNA nucleotides with the twenty amino acids that make up the fundamental components of the proteins, it would be possible to simplify the identification of these proteins. This, in fact, was the goal of biology research during the 1960's, which culminated with the discovery that each gene nucleotide "triplet" corresponds with an amino acid to be inserted in the corresponding protein according to a very precise *translation* map (the genetic code).

Ever since then, the world scientific community has endeavored to reconstruct the complete sequence of human chromosomes (the genome project), to identify the over thirty thousand DNA "fragments" that are translated into proteins (the genes) and to understand the structure, the shape, and mainly the function of these proteins. As a result, they have gathered an enormous quantity of structurally heterogeneous additional information about every gene (named "annotation").

In this process of meticulous analysis scientists concluded that the behavior of cells is characterized not just by the sequences of DNA and of proteins (let's consider that a skin cell and a brain cell of the same individual have the same DNA but are completely different among themselves) but that also the balance of delicate systems of control determine if, when and how much a certain gene, in a certain cell, should be *expressed*, that is, translated into the corresponding protein, and the subsequent behavior of these proteins. The need to understand these mechanisms (called *epigenetics*) has led scientists to develop technologies and instruments capable of measuring, for instance, the quantity of the "*expressions*" in each known gene in a sample of cells, the presence of certain proteins in a cell, their *localization* (for example, in the nucleus, at the center of the cell rather than near the outer membrane), or whether the proteins in a certain cell interact among themselves (that is, the creation, within the cell, of *composites* in which a number of proteins are chained to one another). These technologies and instruments are often capable of carrying out this type of analysis simultaneously on thousands or tens of thousands genes, producing enormous quantities of information in relatively short time, and are defined as *high-throughput* platforms.

In addition, many of these "properties" of proteins (such as localization, interaction, etc.) may be predicted by analyzing their amino acid sequences. The enormous amount of experimental data produced in the last 30 years and the analysis and the management of databases have allowed developing computer programs that predict countless biological properties of the proteins. A good example is provided in the website of ExPASy (http://www.expasy.ch/) (Expert Protein Analysis System), the proteomics server of the Swiss Institute of Bioinformatics (http://www.isb-sib.ch/),

---

[9] The codification used at present is called "IUPAC code" and was defined by the **International Union of Pure and Applied Chemistry**.

10 This is also true of another component of the family of the nucleic acids: the ribonucleic acid, or RNA, which, from a purely bioinformatics perspective, differs from DNA in that it contains a different nucleotide and in that it is composed of only one filament instead of the "double helix". However, the same as DNA and proteins, RNA has a 3-D structure.

where dozens of programs are made available to the International scientific community via the web.

Epigenetics probably constitutes molecular biology's utmost challenge today. Accordingly, acquiring the ability of manipulating the various types of data furnished by these tools and correlating them until simulating complex systems will constitute bioinformatics' challenge for the coming years.

Databases and the Internet

The information produced with the aid of the mentioned technologies was soon to become so copious that it became almost unmanageable (and therefore, useless) without the assistance of computer databases. If just the sequence of the human genome would occupy hundreds of thousand pages in print, this would be still little if compared to the mountains of annotated information and the data produced by the *high-throughput* platforms. In effect, the planning and implementation of biological data archives has become such a relevant activity that one could think that bioinformatics is almost exclusively about this. [11] Before the proliferation of the Internet, however, the use of the databases had some heavy limitations, in particular the difficulty and slowness of keeping updated. There were not enough links to networks that would allow filling and aligning the genetic databases, and the only alternative was that of transferring the data by circulating data tapes or other high storage memories among the research centers from time to time. Thanks to the spread of the Internet, research structures have overcome this limitation and now the genetic databases may be aligned more easily.

Genetic databases are of various types[12]. Some, such as is the case of GenBank[13], keep the data as faithfully and untouched as possible in order to enable researchers to analyze them with new techniques. On the other hand, some frequently used data are calculated and kept in turn inside other databases – called *derivative*[14]- in order to prevent unnecessary repetition of one same analysis.[15]

Databases of this type contain both pure genetic data and "annotations". The annotations are copious additional technical information that databases managers add to "pure data " that is extremely helpful for those who then access this information. Usually, the information memorized in databases is found in flat-file format, that is, a continuous sequence of characters presented in one or more lines, each line identified by a control code placed at the beginning of the line. But the need to cross primary and secondary information, besides the need to de-localize such information, has

---

[11] Obviously this is not so, but the question has been provocatively raised by two researchers in *Isn't bioinformatics just about building databases?* (see GIBAS, C., JAMBECK, P., *Developing bioinformatics computer skills* O'Reilly 2001, p. 8.

[12] This definition will sound to general for the experts, but will suffice the purpose of this article.

[13] The GenBank (created in 1982 in the U.S.A.), the *EMBL Data Library* (Europe 1980) and the Japanese DDBJ are known as "primary databanks " and have historical and documentation functions.

[14] Swiss-Prot is a derivative database that deals with proteins whose added value is exactly opposite to that of the primary databases. Primary databases, as mentioned in footnote 14, focus on gathering data in an (almost) raw format. On the contrary, derivative databases concentrate on the quality of the annotations, that is, circumstantial information regarding the sequence.

[15] TRAMONTANO, A. *op. cit.* p. 2.

made it indispensable for researches to resort to *Database Management Systems* (DBMS) that do not necessarily work on open and compatible formats.

One aspect of bioinformatics research is that it takes good advantage of a converging approach that was possible to implement only thanks to the Internet and to the expansion of distributed computing projects also outside the academic community. This is the case of the *Folding@Home*[16] (FaH) project by Prof. Vijay S. Pande, a professor at Standford University (USA). Downloading certain software developed by this group of American researchers that can be run on independent computers, any participant can become a member of the computer network in which each one carries out a small part of the protein sequencing work. In this way, a multitude of single computers is transformed into the equivalent of the most powerful of supercomputers.

Biology and information technologies

The classification of the methods used by researchers summed up and partially described before suggests that some of these methods (such as sequence analysis) are actually merely "computing transpositions" - improved and refined, though - of techniques that had already been developed before the use of computers. Other methods, such as the use of distributed computing or the writing of mathematical simulation patterns to understand the complex balances that determine the functioning of the biological systems, are the autochthonous child of the convergence of Life, Mathematics and Information Sciences. This merger of sciences has collateral effects (some of which may be predicted) that derive from the substantial approach that characterizes applied informatics. In a few words, the results are a "system" that, all in all, "doesn't work too bad". [17]

The quality of software, data and information

The fact that users consider themselves satisfied – or, to put it in legal terms, if the obligation undertaken by the software manufacturer is deemed fully performed by these users– when a certain computer program or an operative system runs "more or less" satisfactorily is due to a perverse addiction to (intentionally created) disinformation and certainly not to the objective and intrinsic features of the development techniques employed in the computer program. [18] In other words, that free-of-error software does not exist does not necessarily mean that it is legal to market malfunctioning products before which users are helpless because of the restrictions imposed by the instrumental use of the protection granted by copyrights. [19] If these arguments are certainly relevant

---

[16] The home page of the project is http://folding.stanford.edu and the software is available at http://folding.stanford.edu/download.html.

Whoever has dealt with ICT knows that there is no software without errors, and that in case of problems or malfunctioning "one usually sits and awaits that the patch or the subsequent version comes out" because "that's the way things are" and whoever uses a certain software does so at their own risk. Actually, the question of the structural (mal)functioning of software is articulate and complex which, to be well understood, demands detailed study of the development and marketing models of such software, of the commercial policies that guide the choices of multinational companies in the sector, of the permeability of the sector to the law, of the relationship between the developers' community and the corporate *top management* sector. For a (discouraging) reconnaissance of these aspects see COOPER, A., *The inmates are running the asylum*, Sams Publishing; 2 Ed 2004 - Ed. it. *Il disagio tecnologico* Apogeo 1999.

[17] COOPER, A., *op. cit.* p. 27. *The prodigies of silicon are so overwhelming that we willingly ignore its collateral costs. If after a shipwreck you are pushed to a desert island, you won't certainly mind if the ship that comes to rescue you has its hulk in bad conditions and is full of rats. The difference between having a software solution for our problem and having none is so huge that we are prepared to accept any inconveniences or difficulties that the solution brings along.*

18 KANER, C., PELS, D., *Bad software* Wiley 1998, MINASI, M., *The software conspiracy* Mcgraw-Hill 1999

[19] Operative systems, programming environments, compiler programs.

in a general context, they become crucial where bioinformatics research is concerned, since the reliability and accountability of the results are – literally – of vital importance. Scientists, in fact, must necessarily rely on applications written by people who are foreign to the specific sector of molecular biology[20] and who – the reference is to companies and single programmers - carry on their backs the cultural background that is responsible for the production of bad software. [21] On the other hand, experts of many and different disciplines have gone into the twists and turns of programming tools for bioinformatics[22] without having an adequate knowledge of programming languages and software design, with the consequence that programs that were developed *ac hoc* by researchers run the risk of containing defects unknown to their authors. This state of the affairs translates into a substantial technical and legal instability of the whole "bioinformatics system", which may be not evident to many, but still happens and can potentially reveal itself without warning and with unforeseeable effects.[23] Indeed, this is absolutely not a theoretical question if we accept that key element in research activities is the "quality" of the information gathered. "Are there errors in the databanks?" – asks herself Anna Tramontano - "Yes: in every databank there are errors of two types: in the data themselves and in the annotations."[24] The errors that afflict the data may be caused either by those who produced the data (and therefore, among other causes, by software errors) or by those who annotated them when inputting the data in the databases, whereas the errors in annotation "may be due to the method used to annotate the sequence, to the spreading of preexisting mistakes, to the fact that when filing the data, some characteristics of the molecule were unknown or known wrongly".[25] In the light of these comments, the notion of "qual-

---

[20] The approach on programming that bears more responsibility than others for software errors is the "code recycling" approach, that is, reusing portions of the source code written by third parties. Besides the foreseeable implications from the point of view of the authorship of the derivate work, when this method is practiced without adequate control, it becomes extremely dangerous because it works as if contaminating a theoretically healthy organism with infected material. Bioinformatics indeed is not immune to this plague, as pointed out by GIBAS, C., JAMBECK, P., *op. cit*., pag. 20: "*an efficient programmer is a lazy programmer. He does not make useless efforts in writing a program if others have already provided one that perfectly fulfills the requirements. If you are looking for something that is merely routine .... you may be sure that someone else has already written the software you need and that.... you will probably even find the source code to look at*.

[21] GIBAS, C., JAMBECK, P., *Developing bioinformatics computer skills* O'Reilly 2001, p. 3 *Researchers approach bioinformatics from as dissimilar fields as mathematic, information sciences and linguistics*.

[22] Programming is an extremely complex and delicate activity. More even so is the programming of biology applications, where besides specific programming knowledge one needs to master operative systems and environments, programming languages (essentially C++ and Perl), mathematics and statistics. Yet, having a sound preparation in these fields does not seem to be, in some people's opinion, a priority or an essential requisite among the skills of a bioinformatics expert. The introduction to an essay on Bioinformatics reads: "*Programming is to computing science as the work of a bricklayer to architecture. Both are creative activities, but the former are crafts and the latter are arts. Many bioinformatics students ask me whether it is important to learn to write complex computer programs. In my opinion (which is not shared by everyone in this field): it is not important, unless you want to specialize in this area .*". A.LESK, *Introduction to bioinformatics* McGraw-Hill 2002 - Ed. it. *Introduzione alla bioinformatica* McGraw-Hill 2004 p. 13

[23] TRAMONTANO A., *op.cit*. p. 15.

[24] TRAMONTANO A, Id.

[25] Among which is the possibility of changing a program, duplicating and redistributing it, avoiding in many cases the possibility of "undue appropriation" of software developed in this way by unscrupulous software companies. In addition, thanks to their efficiency, sturdiness and versatility, "unix-based" systems such as Linux are renowned as being the preferred software for scientific research.

ity" of bioinformatics data may be better explained as the "*sacra trimurti*" of "integrity", "reliability" and "non-repudiability" that pervades every bend of ICT world.

The ownership of software and file formats

The problem of the "quality" of the data contained in the biological databases we have discussed so far leads directly to the question of the "ownership" of the software tools used in research and of the data produced by such research activity. The bioinformatics community heavily employs operative systems, web applications, scripting and programming languages and programs regulated by open source licenses; whereas the "raw" results produced by the laboratories end up in public databases that are freely accessible to researchers from all the world. The reasons for such a choice seem obvious. On the one hand, open source software guarantees great flexibility of use without risks of legal disputes (also instrumental) arising from the infringement of copyrights.[26] In addition, the free circulation of the research results is unanimously considered an essential value. This could seem too good to be true, and in fact, we ask ourselves whether that of bioinformatics' is actually such a perfect world, untouched by the interests (in the negative sense) of industry and by other interests.

Some healthy pragmatism is found in the *Frequently Asked Questions* included in the cited *Folding @ Home* project. After having provided and explanation to the first question "what does *protein folding* mean¨", the document goes straight into *in medias res* asking rhetorically "Who owns the results?" and then "Why isn't the source code (of distributed computing *software*, A/N) available?" The answer to the first question textually is: "Unlike other distributed computing projects, Folding@home is run by an academic institution …. which is a nonprofit institution dedicated to science research and education. We will not sell the data or make any money off of it. Moreover, we will make the data available for others to use…. The analysis of the simulations will be submitted to scientific journals for publication, and these journal articles will be posted on the web page after publication. Next, after publication of these scientific articles which analyze the data, the raw data of the folding runs will be available for everyone, including other researchers, here on this web site." To the question as to the non-availability of the source code, the answer is: "Most of the critical parts of FAH are publicly available. The *Tinker* and *Gromacs* source codes can be downloaded and run. Unlike many computer projects, the paramount concern is not functionality, but the scientific integrity, and posting the source code in a way that would allow people to reverse engineer the code to produce bogus scientific results would make the whole project pointless."

The legal status of information

The position of the Standford researchers is so interesting because with a few phrases they are accomplishing a truly Coperican revolution in the traditional way of looking at the legal nature of biological (in this case) data. Contrarily to others, they declare "**we** will not sell the data that **you** provide us with." And then: rather than in the functionality of the programs, we are **interested** in the **scientific integrity** that the **data** and this is why we do not post a part of the source codes of our software. We are not so concerned about the paternity of the information as much as about its (scientific, but also economic) **value**. Thus, the sense of asking "who owns the results*?*" is much deeper than it appears *prima facie* because it involves both the information as such and the tools used for producing this information. Which necessarily leads to the problem of the abstract possibility of figuring out some legal principle that encompasses information as a "legally protected interest".

---

[26] Larry Thompson *Genes, Politics and Money: Biologist Most Ambitious Project Will Cost a Fortune, but Its Value Could Be Out beyond Measure* Washington Post Z12 (24/2/1987)

Who owns the information?

The first attempt to "protect" – or, better, to take possession – of genetic information was made by the American National Health Institute (NIH), which applied for the patenting of 2.750 cDNA sequences without knowing how they worked. The US Patent Office (USPO) denied the application on the grounds that the sequences had been "discovered" and not "invented", that they lacked the novelty requirement because they came from public files, and that they lacked the utility requirement because it was not known what they were useful for. However unsuccessful, the attempt of the NIH showed a change in the logics of seeking patenting protection by using patent "preventively": you don't patent what you invented; you patent something that could have an economic value even if it is not an "invention" in the technical and legal sense, and without even knowing whether you actually have "the goose that lays the golden egg". Trying to unroot the problem from the core, Europe issued European Directive 98/44CE (Protection of biotechnological inventions), ratified in Italy with DL 3/06, in 4 c. I) a), which expressly forbids the patentability of a "mere discovery of one of the elements of the body, including the sequence of a gene, or the partial sequence thereof" and of "DNA's simple sequence, a partial sequence of a gene, used to protect a protein or a partial protein". "Unless" – the decree states further on - "the applicant furnishes the indication and the description of some function that is useful for evaluating the requirement of the industrial application and that the corresponding function is specifically claimed; each sequence is considered autonomous for patenting purposes in the case of overlapping sequences only in those sections that are not essential to the invention." As may be easily understood, this exception follows the same motivations of the USPO's rejection of NIH's application and is a way of securing, surreptitiously, the patentability of sequences that have a value, leaving those without value to the public domain.

Information and copyrights

The impossibility of securing patents did not stop the attempts to establish some sort of "ownership" on the genetic information, and alternative ways have been sought. As far back as 1987, Walter Gilbert, one of the pioneers in bioinformatics research, declared to the Washington Post: "I don't believe in the patentability of the genome. What we are actually interested in is securing copyrights on the sequences. This means that if someone wishes to read the code, they will have to pay us to get access. Our goal is to make the information available to everyone. Provided they pay a price." [27] This position does not seem viable, at least in Europe, because the legal treatment given by the law to personal data - and genetic data being considered "sensitive" data - is the first shield to this attempt of "undue appropriation"[28]. Moreover, genetic sequences as such do not have a "creative" character and therefore may not be protected by copyrights.

However, the laws that govern the treatment of personal data do not have jurisdiction over the sequences because in many cases – especially in cases where research does not depend on the association between genetic samples and their "source" – the sequences are substantially anonymous and as such, may be dealt with freely. In addition, about the claim of "lack of creativity", may be the sequences do lack creativity but certainly the protection granted to databanks by Community[29]

---

[27] Reference to art. 646 of the Italian Civil Code is not casual: in the case of genetic data sequences, the research laboratory does not "produce" the raw material with which it works but receives it from external sources. We could therefore say that, in certain instances, there is undue *interversio possessionis.*

[28] Dir 96/9/CE (In the Italian Official Gazette – 2nd Special Series n.34 - 6 May 1996).

[29] L. 633/41 art. 102, second and third paragraphs.

and Italian[30] laws grant the maker of a databank the same protection as for "abusive drawing". This actually gets around the problem of the legal nature of information, and in this case specifically, of genetic information. In other words, in terms of protection of information, rather than the "nothing" obtainable from patents, it could seem advisable to follow the road of the "something" offered by copyrights: a dangerous and deceitful road, though.

The prospective imposition of copyrights on genetic sequences is, as will be explained in brief, extremely alarming, and this assertion is backed by that fact that the data unavailable through the Internet is hardly useful. More even so, we might say that usefulness of these data decreases inversely proportionately to their growth into enormous quantities. Nevertheless, when an enormous quantity of data can be produced, managed and manipulated by computers, things change, and what was a weakness - numerousness – becomes, thanks to computers, an actual strength. As we said before, genetic information is memorized in formats that are different from the *flat-files*, and may therefore be read by any application. Consequently, treating genetic information requires the use of specific software, capable of reading those formats. As it follows, those who have the copyrights of the formats in which the genetic information was memorized at first practically become the exclusive owners of the information. In any case, the problem does not only have to do with the file "formats" but also extends to the tools used for manipulating those files. "We've come to a point - James Boyle writes – where genetic information is seen as information in the first place. We are concerned with the informative message – the sequence of A, G, C, S and T – and not with the biological medium. The genome project is just a huge cryptography exercise. The same as the archeologists that worked on the Rosetta stone, we too have broken the code and now can use DNA as a "language" to "be spoken", and no longer as "an object" to "be examined".[31] The development of the biotechnologies and the completion of the genetic map, says Boyle, will make it possible to intervene on the genetic apparatus of an individual using word processor-like tools: his biological fate will be simply written down on a first draft, which we will spell check, slightly change or even fully rewrite. Then, whoever will own the rights on software - that could be called "genetic spell-checker" or " tri-dimensional protein modeler" - will find himself in the condition, in fact and by law, of selecting (for example by establishing license prices) who may and who may not afford conducting genetic research; ultimately, who may qualify for an "author" of life. [32] Unacceptable as it may be, this is a necessary conclusion if we accept extending copyrights – a right granted to individuals - to genetic sequences.

Information: from copyrights to tangible property rights

To come out of the trap of such unacceptable syllogism, we need to change the main assumption, that is, that genetic information is protected by copyrights, and to fit this information in a different legal framework, a trendier one that allows balancing the interests at stake (research, industry, citizens). One possible solution of this kind is implicit in a semantic mistake spread out by the continuous "anglicizing" of the Italian legal institutions, and in particular the well-established assumption that, in practice, "copyrights" and "intellectual property" are equivalent. Although these terms are frequently used as synonyms, they are not at all the same thing. Indeed, the former conceptually refers to an individual's interest and the latter[33] concerns a property right ("ownership",

[30] BOYLE, J, *Shamans, software & spleens* 1997 Harvard University Press, p. 4.

[31] Including the set of rights that derive from authorship, and including the right to grant a license for disassembling information, which is completely free at present.

[32] Actually, the notion of "intellectual property" is not found in positive law and therefore, with strict rigor, its equivalence with copyrights would not be formally legitimate.

[33] To become an actual "parasite" struggling to spread out in every sector. A good example of this is tendency to force software and databanks into the sphere of protection of statute l.633/41 or, more recently the *querelle* on the qualificability of sports events or Tv formats as works of talent.

in effect). There should be a net distinction between what falls in the field of copyrights (the creative act) and what does not (an immaterial asset having economic value) to avoid "invading the field" and undesirable effects. That this problem undeniably exists is revealed by the fact that with the passing of time, copyrights have progressively lost their original connotation (i.e. protecting the creative sphere of individuals)[34] [35]. However, we are far from being able to ascertain that the solution is *tout court* to invoke a copyrights-type of protection for all immaterial property.[36] In fact, the laws in force nowadays already offer good pretexts for establishing differences between *information* as such (and the "data" that make up this information) and the works of human talent, to lead information (with a daring *fictio,* though) to the general category of *res*, finally to allow a constitutional reading of its role in the scientific, political and economic dialectics.

Italian law and information

Proof that information is considered an "autonomous entity" is mostly found, historically, in the Italian Penal Code, which punishes the infringement of classified information, trade secrets and industrial secrets.[37] Actually, the criminal protection is aimed at punishing the disloyal behavior – i.e. the breach of the duty of confidentiality – rather than the object of such behavior, that is, the revealed "secret". Article 623 of the Penal Code, refers[38] to "information" as the material object of the illicit behavior, whereas art. 392 P.C. expressly states that "the infringing behavior" requirement is met also when such behavior is aimed at a computer program. In addition to computer programs, also "data" and "information" are considered material objects by art. 635. With respect to *insider trading* aspects, for the purposes of the application of the subsequent art. 184 (abuse of privileged information) art. 181 DLGV 58/04[39] defines "privileged information as "... information of precise nature which has not been made public and which directly or indirectly concerns one or more issuers of financial instruments or one or more financial instruments which,

---

[34] Which actually protects the interests of a small group of individuals. Please see http://www.alcei.it/?p=112, ALCEI's (Italian Association for the Freedom of Interactive Electronic Communication) notice of 17 February 2006, 20 June 2006, reads: "EC Directive 2004/48 was approved in great secrecy, increasing, for the umpteenth time, the powers of the audiovisual majors". And *P2P Has the San Remo Pact Gone Bust?* in *Punto Informatico* n. 2565 of 22 June 2006.

[34] Even the "supporters" of copyrights are using more and more terms like "theft", "stealing", "misappropriation" when referring to the non-authorized copying of protected works and software. In this way, they are obviously legitimating the identification of this immaterial property with "tangible" (*res*) property.

[35] Art. 362 refers to the disclosure and use of business or official secrets, Art. 621 to the disclosure of privileged documents and Art. 622 to the disclosure of professional secrets.

[36] On the disclosing of scientific or industrial secrets, the law establishes that he shall be punished who "having become aware, by reason of his status, office, profession or art, of information which is meant to remain secret about scientific discoveries or inventions or about industrial applications, discloses such information or uses such information for his own or other persons' benefit.

[37] DLGV 24 February 1998 n. 58 (Unified Text of the Tax Authority - TUF).

[38] Enacted in the U.S.A. after the Enron scandal. See CLARKSON, A. *SOX Act e controlli IT* in ICTLEX Briefs 1/06 e 2/06 Monti & Ambrosini Editori.

[39] For instance with the non-competition and the non-disclosure agreement s.

if made public, could significantly affect the prices of such financial instruments". In this same financial field, US 2002 Sarbanes-Oxley Act (SOX)[40] imposes a complex and burdensome system of controls on the transparency and correctness of the financial information that American companies must furnish. Similarly, the prevailing trend in both the Italian[41] and Community[42] civil laws is to consider information as the "means" by which a wrong is committed rather than the actual object that must be protected. It was only by the unfortunate Act 675/96 and its deserving successor, Legislative Decree DLGV 196/03,[43] that the concept of "information"[44] actually gained greater ontological independence. Admittedly, the type of information referred to by DLGV 196/03 is directly connected to an individual or may be immediately referred to an individual; so clearly, the results obtained from the sequencing activity do not fall in this category. But the principle set by art. 1 of the cited decree[45] introduces a useful interpretative element by establishing a connection between a person having rights ("any person") and the object to be protected (the "personal data" and therefore the "information"), which connection leaves completely out of consideration any reference to copyrights or patents. Moreover, legislative decree DLGV 30/05 [46] recently introduced modifications to Italian industrial property law that opened interesting new frontiers but raised some problems, too. In fact, art. 98[47] endorses know-how and confidential information

[40] UE Court of Justice – V Section , Sentence of 24 June 1986 n. 61985J0053 - AKZO CHEMIE B. V. e CHEMIE U. K. LTD. against the EUROPEAN COMMUNITY COMMISSION: By virtue of art. *214 of the Treaty, employees of institutions shall not disclose information in their possession that be protected as privileged or secret business information. Art. 20 of Regulation 17/62 that regulates this law within the framework of company law establishes in its ss. 2 that, except for the provisions of arts.19 and 21, the Commission and the competent authorities of the Member States and their officers and other agents shall not disclose any information collected in applying this regulation and that, by its nature, be protected as privileged or secret business or official information.*

[41] A long critical report on the articles dealing with this subject published in the last ten years appears in the Interlex magazine at http://www.interlex.it/675/indice.htm.

[41] Even though as the *species* of the "personal data" *genus*.

[42]Any person has the right to the protection of information concerning his or her own personal data.

[43] Legislative Decree of 10 February 2005, n. 30 Industrial Property Code according to Article 15 of Act 273 of 12 December 2002" in the Italian Official Gazette n. 52 of 4 March 2005 – Regular Supplement n. 28.

[44] The Decree reads: "1. Company information as well as technical and industrial experiences, including commercial ones that are under the legitimate control of the holder thereof constitute objects of protection, if:

a) such information is secret, in the sense that it is not generally known or known in their precise setting or the combination of its elements nor is it easily accessible to sector experts o professionals;

b) such information has an economic value by virtue of its secret nature;

c) such information is subject to measures that are deemed reasonably adequate to maintain information secret by the persons that legitimately control it.

2. Protection is also granted to data connected to tests and other secrets the making of which involves considerable effort and whose disclosure is subject to obtaining authorization for the launching to the market of chemical, pharmaceutical or agricultural products involved in the use of new chemical substances."

[45] With the meaning that Pythagoreans gave this term.

[46] With the (relevant and useful, for our purposes) exceptions of arts. 392 and 635 2° paragraph of the Italian Civil Code.

[47] Obviously, the legal protection of a new propeller project developed in the facilities and with the resources of an airplane factory may only be guaranteed in so far as the owner of the company protects the information adequately.

as autonomous "objects" to be protected by the law and recognizes that the condition for enforcing such protection lays in their secret nature (which must be guaranteed by adequate security measures, says the Decree). Hence, on the one hand decree DLGV 30/05 helps to "reify" it but on the other hand, the decree reinforces the concept that information, as such, (or at least "certain" information) is only food for the "parasites of copyright". The second paragraph of art. 98, indeed, states that "Protection is also granted to data connected to tests and other secrets whose making involves considerable effort and whose disclosure is subject to obtaining authorization for the launching to the market of chemical, pharmaceutical or agricultural products involved in the use of new chemical substances." Again, although this rule does not necessarily apply to genetic sequences, it is wide enough to comprise - in certain conditions - also the results of bioinformatics research. If it is understandable that besides protecting the "classical" industrial property (trademarks and inventions) the Legislator also wanted to protect these new "objects" that have gained a value and relevance of their own, it is also true that – keeping in mind the grounds upheld by the USPO to reject NIH's patent application on the 2.750 cDNA sequences - "data" and "information" themselves are not ontologically comparable with or referable to industrial inventions. We must then come to the conclusion that at present industrial property is made up of **two** types of property rights: inventions (which are protected as a whole, considering the investments that are necessary to accomplish them) and information (which is protected as an "asset" of mathematical knowledge[48] - or as the "brick" to build inventions).

We may then say that the unspoken assumption in the choices of both civil and in criminal law [49] , is that there is a biunique link between the "value" and the "secrecy" of the information. Hence the legal principle by which only that over which the *jus excludendi* is actively exercised – in this particular case, by preventing the access to information by third parties – is granted protection. In any case, the equation"("secrecy" V value)→legal protection" derives from the idea that the protection of the results of a certain research that is overall carried out in secrecy, and that it is the "hiding" itself that justifies such protection. [50] Conversely, in the case of genetic information (and, in general, genetic information publicly available), the "value" is exactly and oppositely made up of the transparency and free availability. For this reason it would not be possible to demand the (full) application of the laws in force for the protection of the genetic information freely available to the public. In any case, we should still bear in mind the possibility that (also) free genetic data and information, as claimed by the Standford researchers, be considered as proper and distinguishable property of its own, and deserving the protection of the applicable laws.

Conclusion

One more step is needed before recapitulating the reasoning developed so far: to check whether it is possible to fit information within the existing legal structure without enacting new legislation. To begin with, we have seen that information has an actual and autonomous value of its own and

---

[48] Protection is granted, in any case, subject to the information being: **secret** (breach of confidentiality is punished by art. 98 DLGV 30/05), **personal** (DLGV 196/03), **intended for the "public credit"** (DLGV 58/04, *SOX Act*). Although information lacking these features (and in particular information available by the public and accumulated in databases) may be granted legal protection only indirectly through copyrights, this fact does not diminish the reification of the "information" or the "datum" as immaterial property. The Italian Penal Code rules on this aspect in art. 392 and 635 second paragraph, by stating that computer programs, data and information are law-protected property and material objects of wrongs and injury.

[49] The reference is to WELLS, H.G., *The Island of Doctor Moreau* Magnum easy eye publication, 1968.

[50] The reference is to CRICHTON, M., *Jurassic Park* Knopf, 1990.

that this situation is forcing legislators to provide specific forms of legal protection for information as such. However fragmented and contradictory, the present scenario shows that information already enjoys some autonomous legal protection of its own.[51] We have also seen that genetic information neither falls within the direct scope of protection of copyrights nor is it protected by patents because of the express ban imposed by Legislative Decree 3/06, noting that the Italian Code of Industrial Property Rights distinguishes between the protection of inventions and protection of (secret) information and of data. Last, we have shown that there is non-secret, non-personal and non-public credit information that in spite of having an intrinsic (also) economic value of its own still lacks the legal instruments that provide direct protection. In other words, we are facing an asymmetric situation that may not be justified by arguments that are traditionally held to sustain the limitations of the effectiveness of the laws of reference.

A common denominator is needed to allow a flexible handling of the cases that have not been covered by specific legislation. And this common denominator must be urgently found because the stakes are too high. Legislators have already proved - and continue to prove - how incapable they are at facing the legal implications of technological and scientific research, as shown by their poor answer to the problems posed by the information technologies. The actual danger is, then, that wrong or inapplicable laws are passed also with respect of biotechnologies, or worse, laws that are instrumented for private ends, as in the case of copyrights. Considering information as tangible property rights, without the need of specific legislation could in fact furnish an efficient solution from a systematic point of view because in this way, (genetic) information as such would be incorporated into a consolidated interpretative legal frame, which would ease the role of those who are called to decide on questions that for a long time now have been considered "new".

In general, extending information to the status of "*res*"(chose) would allow solving the problem of avoiding the unpatentability under art. 1 DL 3/06 with its hidden trick besides securing an efficient balancing of private and public interests. Too worried about the reaction of the public opinion, ignorant legislators (or acting in bad faith) thought that forbidding the patentability of the sequences would be enough to avoid the proliferation of Doctor Moreau's [52] illegitimate children , of unprejudiced *Jurassic Park* imitators, with COD primitive men side by side with prehistoric reptiles, or, even more, of Brazilian-German progeny. As pointed out before, the unpatentability of information does not preclude the existence of other forms of protection that can be even more invasive than the patent and which in the end accomplish the same result: leaving the control of human life in the hands of private entities, without any possibility of exercising governmental monitoring and restrictions. If genetic information were "*res"(*choses), the enforcement of tangible property laws on information would be limited to the social function that the Constitution grants to public and private property, which would allowing the Government's direct intervention in balancing what is public and what is not every time that the interests of society be endangered by the excessive extension of the feuds and of the "*mortmain*" of the Lords of Life.

[51] The reference is to LEVIN, I. *The Boys from Brazil* Dell Publishing, 1977.

[52] From the imposition of copyrights on databanks to the patenting of the information collection and organization process, to the control of the file formats with *Digital Right Management* systems to that on software and on DBMS.