Conceptual Modeling Meets the Human Genome

Oscar Pastor

Centro de Investigación en Métodos de Producción de Software –PROS-.
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
opastor@dsic.upv.es

Abstract. Looking backwards, it makes sense to discuss the value that Conceptual Modeling has provided to the Information Systems Design and Development area. Thinking about the present, the most advanced Software Engineering approaches oriented to producing quality software propose using extensively conceptual model-based approaches. Conceptual Modeling is widely used in the Information Systems domain. Nevertheless, in terms of Conceptual Model Evolution, we should wonder which new application domains will become more challenging for Conceptual Modeling in the very near future. In an attempt to answer that question, one path to follow is associated to the Bioinformatics domain and specifically, to confront the problem of precise understanding of the Human Genome. The problems related to this topic have become first-order issues in which, curiously, the role of Conceptual Modeling has not yet been fully exploited. The comprehension of the Human Genome is an extremely attractive topic for future research taking into account the continuous and increasing interest that is being generated. Therefore, it is worth analyzing how Conceptual Modeling principles, methods and techniques could help to face the problem and how Conceptual Modeling could aid to provide more efficient solutions. The basic goal of this talk will be the introduction and the discussion of these ideas. If we look at the Human Genome as the representation of some Conceptual Model –which is not yet known-, interesting analogies with the modern Model-Driven Software Development principles appear. As a precise interpretation of the Human Genome would be much easier if the underlying model were known, Conceptual Modeling can provide new ways of facing that problem in order to obtain new and better strategies and solutions.

Keywords: Conceptual Modeling, Bioinformatics, Human Genome.

1 Introduction

If we look at the past, it makes sense to discuss the strong value that Conceptual Modeling has provided to Information Systems Design and Development. If we look at the present, we see how the most advanced Software Engineering approaches, which are oriented to producing software with the required quality, extensively use models under the acronyms of Model Driven Development (MDD), Model-Based Code Generation (MBCD), Model-Driven Architectures (MDA), etc. Nowadays, Conceptual Modeling is widely used in the Information Systems domain. If we look

Q. Li et al. (Eds.): ER 2008, LNCS 5231, pp. 1–11, 2008.

at the future, we could wonder what kind of new application domains could become more challenging for Conceptual Modeling.

Specifically, the Bioinformatics domain in general (and the understanding of the Human Genome in particular) are currently considered to be first-order issues, where the role of Conceptual Modeling has not yet been fully exploited. Considering the continuous and increasing interest of this domain, analyzing how Conceptual Modeling principles, methods and techniques could help to improve the current ways of facing the problem and how Conceptual Modeling could help to provide more efficient solutions, is an extremely attractive topic.

The introduction and the discussion of these ideas are the basic goals of this keynote speech. Describing the Human Genome as a representation of some mainly still unknown Conceptual Model, analogies with the main Model-Driven Software Development conventional principles will be analyzed in order to achieve a precise interpretation of the Human Genome.

Current and future scenarios based on these ideas will be introduced. If Conceptual Modeling is being an effective approach for providing a sound linkage between concepts and their associated software representation –facilitating the understanding of where programs (seen as conceptual model representations) come from-, why not conclude that Conceptual Modeling can be equally effective in understanding the Human Genome (seen as a representation of a Conceptual Model) by extracting the concepts that are behind it? Since the interpretation of the Human Genome is a big challenge for the scientific community, the use of Conceptual Modeling-based notions and methods to undertake this problem will open exciting scenarios finding more efficient scientific strategies with their corresponding set of original solutions, tools and subsequent practical applications.

2 Will Genome Conceptual Modelling Really Make Things Better?

Why use Conceptual Modelling for understanding genomic information? How can the application of Conceptual Modeling techniques help to improve the current strategies that are used to confrontfor facing that challenging problem? These are the main questions to be answered in this talk. Firstly, virtually everything that is known about genomes and genome expression has been discovered by scientific research: it is said to be quite normal to learn “facts” about genomes without knowing very much about “why” they happen as discovered. This lack of conceptual understanding is an interesting first problem.

Secondly, understanding requires precise definitions. Too often, vague descriptions appear associated to basic bioinformatics concepts. This imprecise definition of concepts is a second problem. As already commented in [1], the provision of clear and intuitive models is fundamental to be able to have an effective description and management of genomic data. Also, we read in [2] that the idealistic goals of systems and synthetic biology will not be feasible without the engaged contribution of computer scientists. The role of Computer Scientists should not be limited to the implementation level. Indeed, I would like to emphasize that the role of Conceptual Modeling is still more important. We could say that the conventional view of a computer scientist

in the Bioinformatics domain is that of an engineer providing more processing power and more refined algorithms intended to process larger and larger amounts of information, normally trying to discover or infer specific patterns in the genome. Instead, a computer scientist should be perceived as a conceptual modeller of reality, in this case, as a modeller of how life works.

That fundamental conceptual perspective is too often just not present. If biological cells are seen as an alternative to current hardware, it is logical to conclude that a software analogy should be engineered to direct cells to produce useful artifacts or substances. However models should play a fundamental role in that scenario, as they are considered to play a basic role in what we could call "conventional" software production methods. Additionally, through the use of models, it would become possible to characterize conceptual patterns seen as modelling primitives related to those human aspects that are represented in that "biological" low-level software that the human genome constitutes. In that case, we could reuse the Model-Driven Development (MDD) "metaphor" of moving from the Problem Space (a conceptual model of the human genome now) to the Solution Space (the human beings that constitute its running implementation).

Conceptual Modeling could provide the necessary basis to have a consistent and integrated representation of genomic data. While the biological community is building and dealing with increasingly complex models and simulations of cells and lots of biological entities, it is our belief that conceptual models can improve communication among the involved actors. Through the use of conceptual models, it will be possible to fix the relevant concepts, abstracting those biological system components that are really required to describe and understand how the human genome works. By having a proper conceptual model, the relevant biological information will be preserved and understood by all the different researchers involved in the challenge of interpreting the human genome. Models are useful to provide different views of the relevant information that are properly adapted to different user needs. Another important issue is that such a conceptual model should include both system structure and system behaviour. On the one hand, a data model must be provided to characterize the static system architecture, specifying the classes and their relationships that make up the structural genome. On the other hand, a process model has to fix the behaviour that is attached to this structure. In most of the text books about biological systems, functional biological information often appears to be drawn with somewhat inconsistent or at least highly complicated nodes and arrows. As a consequence the readers are often confused because a vague view about how a biological system works is provided to them, especially when they do not have enough knowledge about the biological systems. Putting the information in a Process Diagram –which is complementary to the former Entity-Relationship Diagram- will package the precise information required to understand the basic structure and the subsequent behaviour of the involved biological complex system that is analyzed.

In that context, understanding how a genome specifies the biochemical capability of a living cell, and subsequently, the rules that determine our perceived behaviour, is the major research challenge of modern Bioinformatics. Conceptual Modeling can provide extremely attractive and efficient answers to this challenge. In the next section, we are going to explain in further detail that process of conceptual modelling analysis, starting from a basic entity-relationship modelling intended to characterize

the structural view of the genome. This view should be complemented with the subsequent process modelling perspective, which is not dealt with in this paper although it will be briefly analyzed in the talk. This behavioural model includes the processes of transcription –in which individual genes are copied into RNA molecules- and translation -where the proteins that make up the proteome (the cell's repertoire of proteins) are synthesized by translation of the individual RNA molecules present in the transcriptome (RNA copies of the active protein-coding genes). According to this genome taxonomy, transcriptome (the result of the transcription process) and proteome (the result of the translation process) ([8]), we are going to focus our next modelling efforts on the Genome, seen as a store of biological information that is in its own unable to release that information to the cell, because the utilization of its biological information requires the coordinated activity of enzymes and other proteins which participate in a complex series of biochemical reactions referred to as genome expression.

3 A Conceptual Schema for the Human Genome

If, for instance, we want to elaborate an Entity-Relationship Diagram to represent the basic genomic concepts, their precise definition will be a need. At least, we are forced to determine in detail what we mean by any given concept. A common agreement in a shared definition for such a fundamental concept as the "gene" concept becomes an extremely interesting issue. We could start saying that the most modern definition basically accepts that a gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products ([3],[4]). This definition manifests how integral the concept of biological function is in defining genes: they are not characterized by their precise structure, which probably exists, but which is mostly still unknown. What precisely characterizes a gene is what it does from a functional perspective, specifically which protein or proteins it can exactly code.

However, as stated above, things are not so simple. When a Conceptual Modeler enters the game of understanding what exactly a gene is, and when one tries to characterize the "gene" entity, it is surprising to see how many different definitions exist, and how difficult it becomes to fix the subset of observable properties associated to the gene notion, as position (start and end of the DNA chain), sequence of nucleotides, etc. that should allow a gene in a DNA chain to be identified uniquely. These properties can vary in different individual of the same specie. In other modelling domains, if this situation occurs, one external property is added to the object to identify it. For instance, a tree in a forest or sheep in a flock can be labelled. But it is not so easy to label a gene within a DNA chain. In the genomic data repositories, genes are named, but that name is even not unique.

In the literature, there are different gene definitions that come from different prominent authors and works. Currently, there are clear discrepancies between what we could call a previous protein-centric view of the gene, and one that is revealed by the extensive transcriptional activity of the genome. For instance, in [3] and [4], emphasis is placed on genomic sequences at the DNA level and what they do in terms of protein production, while in [5] the point is that at the DNA level, the gene cannot yet

be directly identified and the formation of the mRNA and its expression must be analyzed in time and space to characterize the gene function at translation time. Taking into account these works, we could even question whether the gene concept exists as a precise concept! Obviously, when genetists make an experiment and they look for a given gene in an DNA sequence, they find it and they manipulate it with certainty. Furthermore, when they talk about a gene, they know what they are talking about. The immediate conclusion is that observable properties that allow it to be recognized do exist. But which properties we are talking about exactly is not so clear when we enter in further detail. Conceptual Modeling can provide a lot of knowledge in that context.

As an example of these ideas, we now present an Entity-Relationship Diagram to describe a gene. The intention is twofold: i) on the one side, to show how conceptual modelling forces us to understand and to define with precision what we are talking about, and ii) to open the door to implement a concrete database corresponding to the conceptual schema, whose context would be clearly structured and ready to be used as a data repository of reference for further, concrete experimentation. Such a unified database including all the information related with genes, their characteristics and their concise behaviour would constitute a first-order result in the current context, where information is spread in a lot of different repositories, with strong problems of interoperability and often with inconsistencies and useless information.

This problem has been intensively reported in the last years. Ram in [9] discusses how difficult it is to connect all those data sources seamlessly unless all the data is transformed into a common format. Different solutions to try to overcome this problem exist, as those based on using the notion of seed [10], from which an extraction ontology can be generated in order to collect as much related information as possible from online accessible repositories. But in any case all these solutions are always partial solutions, and the underlying problem of lack of uniformly structured data across related biomedical domains is a barrier that is always present.

3.1 A Model for Chromosomes

Even if in this section we restrict ourselves to the gene concept due to obvious size constraints, let us first introduce some basic genome concepts to show the huge amount of complexity associated to the goal of modelling any genome in general, and certainly for the human genome in particular.

In biology, the genome of an organism is its whole hereditary information and is encoded in the DNA (or, for some viruses, RNA). A genome includes all the genetic material present in the cells of an organism. In eukaryote beings – those whose cells are organized into complex structures enclosed within membranes, including a nucleus- genome refers to the DNA contained in the nucleus and organized in chromosomes. A very basic hierarchy of concepts can be seen in Fig.1, where the highest level is constituted by the cell, and the lowest level is made up of chromosomes and genes.

The genome of an organism is a complete genetic sequence on one set of chromosomes. Chromosomes are organized structures of DNA and proteins that are found in cells. A gene is basically a locatable region of genomic sequence, corresponding to a unit of inheritance, although an attempt to define it precisely is the goal of this section.

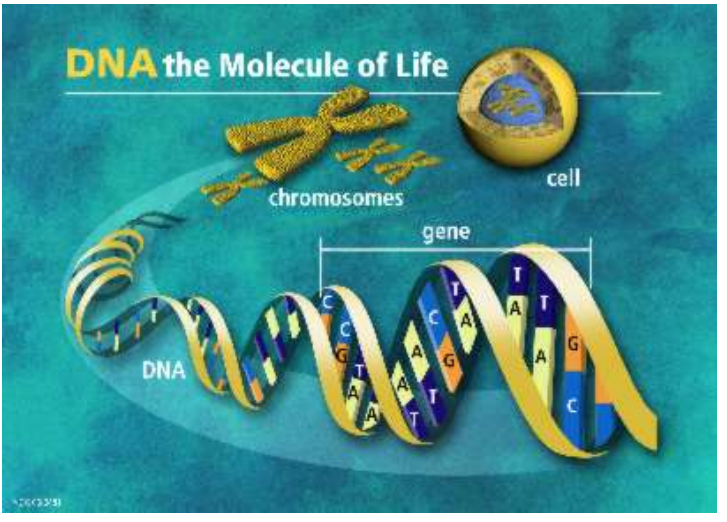

Fig. 1. From the cell to the genes, through the chromosomes

For our modelling purpose, the chromosome segment is the first relevant structure within a chromosome to be specified. By a chromosome segment we refer to i) a set of genic sequences that also includes regulator sequences, or ii) nongenic sequences including other chromosomic elements and intergenic elements. This basic structure is graphically depicted and can be analyzed in Fig. 2.

A chromosome segment could be seen as a DNA sequence constituted by different genic or non-genic sequences. Instead of looking at a chromosome segment as a union of different types of DNA sequences, we will model a chromosome segment as a conceptualization of any relevant type of DNA sequence that is present in a chromosome. From the modelling representation perspective, this means that a specialization relationship will be used instead of an aggregation or association relationship. As we will see later, a parent entity ChromosomeSegment will be specialized in a set of disjoint, descendent entities that represent the different existing, relevant types of chromosome segments. In this way, the chromosomes can be defined as an ordered sequence of chromosome segments that have a precise functionality and that can overlap.

We will refer to these types of chromosome segments as genic and non-genic segments. By a genic segment we mean a DNA segment made up of the following components:

- A Promoter, which is a DNA sequence that controls the start of the transcription process.
- A Transcribed Sequence, constituted by a set of nucleotides that contain the instructions that have to be transcribed and translated in order to synthesize a given protein for a specific gene. It has a precise start and end point defined by particular codons (combinations of three nucleotides with that special function).
- A Terminator, which is a DNA sequence that signals the end of a gene, or more precisely the end of a transcription chain.
- An Enhancer is a DNA sequence that includes the instructions that fix where and how much a gene will express itself.

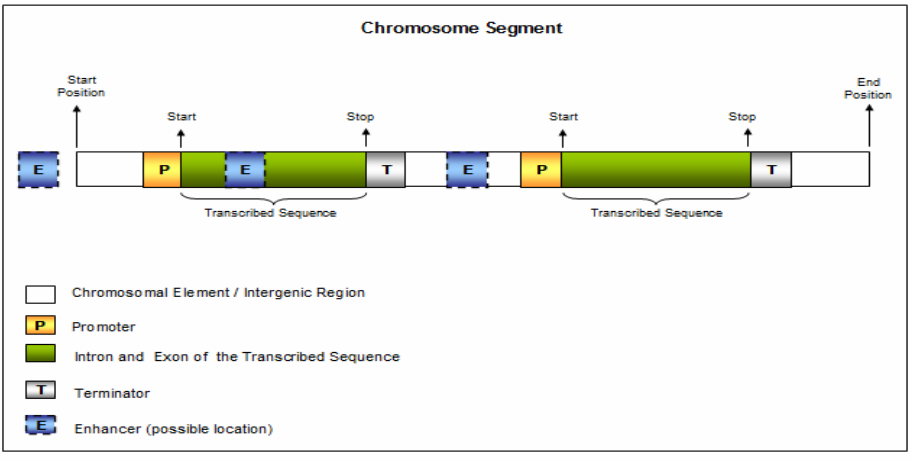

Fig. 2. A chromosome segment can denote either a genic segment (a transcribed sequence, a promoter, a terminator or an enhancer), or a non-genic region which includes non-coding DNA located in between the genic segments

Promoters and terminators are called regulators sequence. To include them in the genic segment is a controversial issue. We have decided to model them in that way, considering that they are an important part in explaining and understanding the gene function. Similarly, how to model the enhancer is an interesting issue. It can be located either within the Transcribed Sequence, or in the non-genic segment, or even it could even be outside of the considered chromosome segment.

The non-genic segments refer to sequences of nucleotides that are considered non transcribable material. Nevertheless, they are part of the chromosome segment as chromosomal elements or intergenic regions, with some particular function probably still unknown. By chromosomal elements we mean the following three different types of regions:

1. ORI, which is are a specific DNA sequence required to start a replication.
2. Centromeres, which are regions that are often found in the middle of the chromosome. They are involved in cell division and the control of gene expression.
3. Telomeres, that constitutes the ending extreme of the chromosomes. They are non-coding DNA regions, which are highly repetitive and which have the main function of assuring chromosome structural stability.

3.2 A Conceptual Schema Proposal for the Human Genome

Taking into account the previous information, we can have an idea of the degree of complexity attached to all the components together with all their interactions involved in the genome structure. In order to understand the genome fundamentals, we propose a Conceptual Schema that could adequately represent all the introduced concepts. This Conceptual Schema, which includes the classes specification, the definition of relationships and the declaration of integrity constraints, provides a basis to fix the main features that are considered relevant to characterize the basic components of the human genome. Fig. 3 presents that proposed schema.

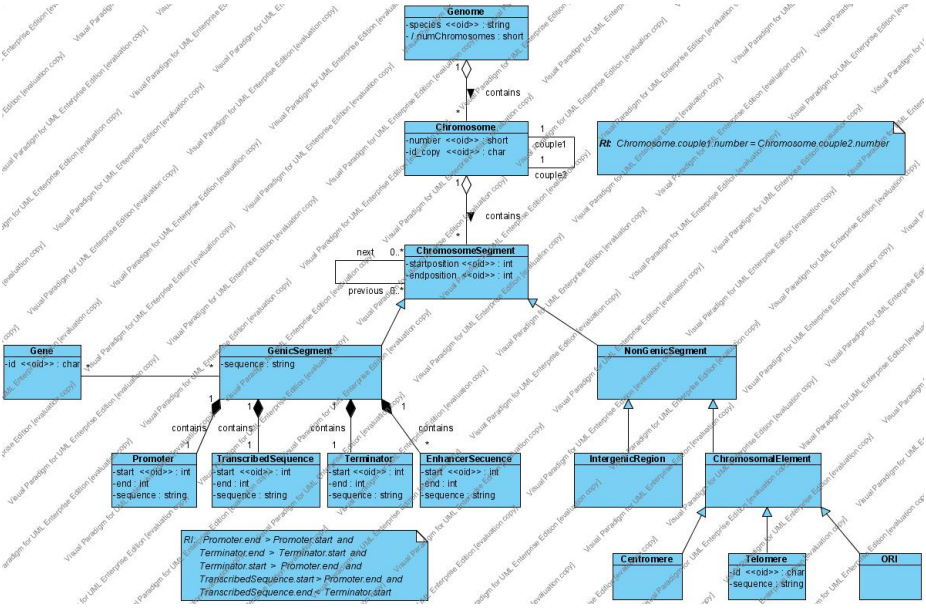

Fig. 3. An ER Conceptual Schema for representing the main components that are relevant to understand the structure of the Human Genome

In this paper, we list the attributes of entities only when these are considered to be important to the understanding of the model as a whole. Obviously, listing all the relevant attributes of all the entities would consume a prohibitive amount of space. Further details will be provided during the keynote address. As usual in the scope of Conceptual Modeling, the final selection of entities, attributes, relationships, cardinalities for relationships, etc. fixes and defines a specific structure that is the result of the modeler decisions, which has direct implications on the intended use of the model in empirical settings. A corresponding database would include the contents according to the model structure. A further characterization of functional products associated to particular genes would have, for instance, a strong impact on the management of biological, biomedical and healthcare knowledge representation. This could, for example, accelerate the development of solutions for the pharmacological and medical industries by benefitting from knowledge reuse and inferencing capabilities.

Many aspects are to be considered, which will be analyzed carefully during the keynote address. The model is full of extremely relevant details. In this short, written presentation, let's mention some of them. For instance, we can observe that:

- a genic segment is associated not only with a transcribed sequence, but also with its functionally relevant regulator components: promoter, terminator and enhancer;
- inter-chromosome segments do not exist since each chromosome segment only belongs to one chromosome;
- chromosome segments are specialized in genic and non-genic segments to represent those parts of the segment that correspond to the genes, and those that

represent intergenic regions or chromosomic elements, composed by (at least as far as it is currently known) non-coding DNA sequences;

- genes are related to genic segments in a many-to-many way, which covers the gene view as a union of genomic sequences encoding a coherent set of potentially overlapping functional products;
- integrity constraints can be specified to declare specific properties. For instance, it appears to be a natural order in the way in which the different types of sequences appear within a genic segment: first, a promoter; then a transcribed sequence; and finally, a terminator. The following constraint based on the start and end attributes of the involved parts could be used to specify that property:

RI: Promoter.end > Promoter.start and
Terminator.end > Terminator.start and
Terminator.start > Promoter.end and
TranscribedSequence.start > Promoter.end and
TranscribedSequence.end < Terminador.start

Broadly speaking, the Conceptual Schema must answer important questions that are present even today in the genomic domain. Is there a distinction between genic and intergenic segments? The Conceptual Schema provides a positive answer by distinguishing between genic and nongenic segments, depending on the coding DNA sequences associated to the rich tapestry of transcription involving alternative splicing. Splicing (including alternative splicing) and intergenic transcription are also some of the most problematic aspects.

Other modeling approaches could be considered. In the works presented in [1] and [6], chromosome segments are specialized in Transcribed and Non-Transcribed regions, instead of the presented specialization of Genic and Non-Genic segments. A Transcribed Region is then associated to a set of Primary Transcripts, while Regulatory Sequences are attached to Non-Transcribed Regions. As regulatory regions are important for gene expression, we suggest that they should be considered as an essential part of the gene, which is in itself a controversial decision. Hence, we assume that regulation is an integral concept in the gene definition, and we adopt the tradition of defining a gene in molecular terms as “the entire nucleic acid sequence that is necessary for the synthesis of a functional polypeptide” ([7])

At the same time, many challenging open problems assure the evolution of the model. It appears that some of the regulatory elements may actually themselves be transcribed. This could mean that promoters, terminators or enhancers could also be seen as transcribed sequences. In that case, integrity constraints such as the one mentioned above would be incorrect. In the extreme, we could even declare the current concept of the gene dead and try to come up with something completely new that fits all the open, not well-solved challenges. Here, we have introduced a tentative attempt to understand the current notion of a gene by means of Conceptual Models. On the one hand, the proposed Conceptual Schema clarifies the currently most accepted definitions, and on the other hand, it leaves the door open to conceptually rethinking and adapting the existing models to the new biogenomic information that is discovered day after day.

4 Conclusions

Imagine for a moment that humans are able to develop to a very sophisticated species of robots, whose specialized behavior is in many aspects a replication of particular human activities. We have seen this kind of fiction in recent, successful movies. Imagine that due to the widely commented current global climate change there is a natural disaster that makes our civilization disappear, while this other silicon-based life –created by humans- persists in time. Imagine for a moment that after centuries of evolution, individuals of that silicon-based species start to wonder about where they come from, and what the fundamentals of their life processes are –which are assumed to be based on binary sequences of 0s and 1s-.

Trying to answer those questions just by exhaustively analyzing the execution model of programs is as difficult as looking for a needle in a haystack. But isn't that just what we humans are doing when we try to understand the working mechanisms of our life by directly exploring our intricate biological-based execution model? In our case, instead of huge sequences of 0s and 1s, we face huge sequences of four nucleotides (A,C,G,T), and we try to discover hidden patterns that should allow us to understand why life processes happen as we perceive them.

How difficult could it be to discover, for instance, the notion of a foreign key –a basic and trivial data modeling concept when it is described at the conceptual modeling perspective- just looking for it in the executable, machine-oriented version of a program? Nevertheless, it is quite trivial to model a foreign key using the DDL of any Relational Data Base Engine. In some sense, within the current Bioinformatics domain, we are looking for the foreign key concept directly in the assembler version of a program. The position of this keynote address is quite clear: this is not the right way.

To understand the program encoded by any genome, we should be able to elaborate and manipulate the models that constitute the source of which a particular implementation is an individual –a human being for the human genome- Conceptual Modeling is consequently a basic strategy that could become the essential approach for guiding the research in Bioinformatics.

We have outlined how a Conceptual Schema can be built to characterize the Human Genome. If such a Conceptual Schema were widely accepted, it would make sense to create a Human Genome database whose contents would include all the essential information to determine which genes synthesize which proteins. As protein elaboration can be associated with particular human behaviors, this will open the door to linking genes with behaviors in order to create a complete catalog of human characteristics. At the same time, this level of understanding can be used to understand the effect of mutations that cause undesired effects –illnesses- and consequently, it would become much more feasible to face and correct them. By applying conceptual modeling-based techniques, we shall not only find ourselves equipped with precise definitions for understanding gene expressions in terms of Molecular Biology, but we shall also be able to devise and apply model-based transformations that could analyze gene storage and expression in terms of information systems processing. This is a real challenge to be overcome by the Conceptual Modeling community in the near future.

References

1. Paton, N., et al.: Conceptual Modeling of Genomic Information. Bioinformatics 16(6), 548–557 (2000)
2. Cohen, J.: The Crucial Role of CS in Systems and Synthetic Biology. Communications of the ACM 51(5) (2008)
3. ENCODE Project Consortium: Identification and Analysis of Functional Elements in 1% of the Human Genome by the Encode Pilot Project. Nature 447, 779–796 (2007), doi:10.1038/nature05874
4. Gerstein, M.B., et al.: What is a gene, post-ENCODE? History and updated definition. Genome Res. 17, 669–681 (2007)
5. Scherrer, K., Jost, J.: Gene and genon concept: coding versus regulation. Theory Biosci. 126, 65–113 (2007)
6. Paton, N., et al.: Conceptual Data Modeling for Bioinformatics. Briefings in Bioinformatics 3(2), 166–180 (2002)
7. Lodish, H., et al.: Molecular cell biology, 5th edn. Freeman and Co., New York (2000)
8. Brown, T.A.: Genome 3. Garland Science Publishing (2007)
9. Ram, S.: Toward Semantic Interoperability of Heterogeneous Biological Data Sources. In: Pastor, Ó., Falcão e Cunha, J. (eds.) CAiSE 2005. LNCS, vol. 3520, p. 32. Springer, Heidelberg (2005)
10. Tao, C., Embley, D.: Seed-Based Generation of Personalized Bio-ontologies for Information Extraction. In: ER Workshops 2007, pp. 74–84 (2007)