EXTRACTING BIOLOGICAL KNOWLEDGE FROM DNA SEQUENCES

F. M. DE LA VEGA[1], D. THIEFFRY[2,3], and J. COLLADO-VIDES[3]

[1]*Departamento de Genética y Biología Molecular, CINVESTAV-IPN.*
A.P. 14-740, México D.F. 07000, México

[2]*Département de Biologie Moléculaire, Université Libre de Bruxelles.*
Rue des Chevaux, 67, B-1640 Rhode-Saint-Genèse, Belgium

[3]*Centro de Investigación sobre Fijación de Nitrógeno. Universidad Nacional Autónoma de México. Cuernavaca A.P. 565-A, Morelos 62100, México*

We are experiencing the time when sequencing genome projects of various organisms are already being completed. No doubt the understanding or deciphering of these large amounts of DNA sequences represents, at least conceptually, one of the major challenges for those working in computational biology.

This track focuses on the different theoretical and computational approaches aimed to extract biological information from DNA sequences. The approaches utilize among others, methods from statistics, information theory, artificial intelligence, linguistics, as well as combinations of these. The variety of approaches exemplified by the different contributions to this track illustrates how computational biology is quickly progressing.

The purpose is to identify functional units in sequences, looking for characteristic patterns and regularities found in a set of examples. Some of the identified units also maintain complex relationships among themselves, allowing to explain regulatory phenomena, or evolutionary relationships. As some complete genomic sequences become available, questions involving properties of the whole cell can be addressed, as well as comparisons of complete genomes across species. A new type of sequence analysis focusing on higher order sequence patterns is emerging.

It is important to note that ultimately what all these methods do is to make use of biological knowledge that has been in one way or the other obtained experimentally, and to transform this knowledge into a predictive tool. Therefore the importance of organized databases and biological knowledge as a resource to decipher genomes and sequences.

In this sense, it is interesting to think about the conclusions that anthropologists have drawn from their lessons in deciphering ancient human languages. Some of "the fundamental pillars" summarized by Michael Coe on which successful deciphering enterprises have rested are the following: i) the database must be large enough; ii) there should be a bilingual inscription, such as the rosette stone; and iii) the language must be known [1]. Etruscan is an example where condition three is not fulfilled, therefore, it can be read but not understood.

In the case of molecular biology, the first condition is being fulfilled by the genome sequencing projects; the second one by the organized biological knowledge in databases where sequences and functions can be matched; the third one is quite an astonishing condition. We leave it as food for thought to the reader.

The organizers of this track acknowledge all anonymous referees who helped in selecting and improving the papers submitted.

References

[1] Coe M. *Breaking the Maya Code* (Thames and Hudson, New York, N.Y., 1992).