ibc Interdisciplinary Bio Central

Open Access

IBC 2013;5:2, 1-3 • DOI: 10.4051/ibc.2013.5.1.0002

ESSAY

Development of Information Biology (I)

Yoshio Tateno[1,*]

[1]School of New Biology, Daegu Gyeongbuk Institute of Science and Technology, Daegu, Korea

Subject areas; Bioinformatics/Computational biology/Molecular modeling

Author contribution; Y.T. wrote this article.

Correspondence and requests for materials should be addressed to Y.T. (yt.tateno@gmail.com)

Editor; Hong Gil Nam, Daegu Gyeongbuk Institute of Science and Technology, Korea

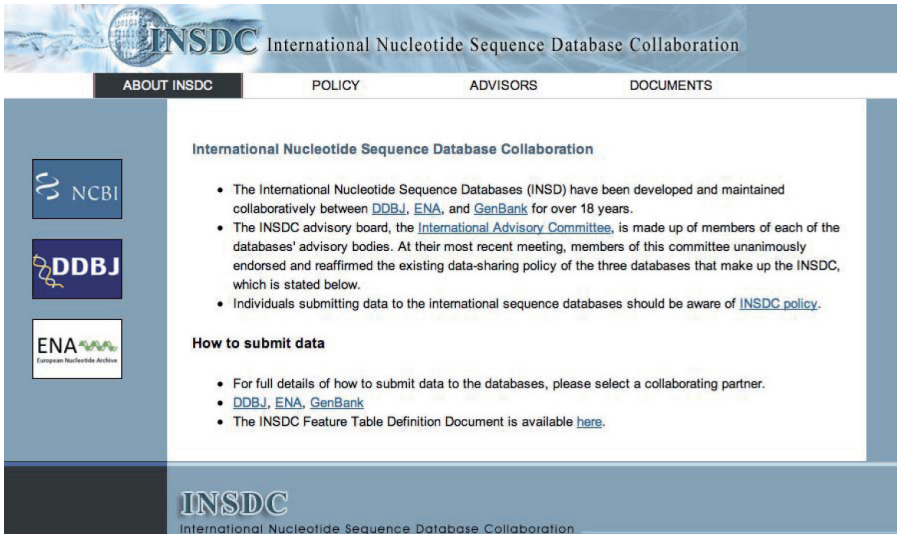
Received March 24, 2013;
Accepted March 30, 2013;
Published March 31, 2013

Citation; Tateno, Y. Development of Information Biology (I). IBC 2013, 5:2, 1-3.
doi: 10.4051/ibc.2013.5.1.0002

Competing interest; All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

Birth and development of information biology are introduced with its definition and scientific basis. The discipline lives on the two types of nutrition, one is a huge amount of biological data on genomes, gene expressions, proteomes, protein 3D structures, protein networks, and so forth. The other is the method of using them on a computer. The scientific basis of the two is evolution. To collect genome and gene expression data form laboratories in the world, annotate and dissimilate back to researchers worldwide, they built the EMBL database in Europe in 1982, GenBank in USA in 1984 and DNA Data Bank of Japan in 1987. On the other hand, the methods of using and analyzing those data have accordingly been developed. The two aspects advance the discipline further and further.



Key Words: information biology; bioinformatics; DNA database; data sharing; homology search; computer tools

First, I would like to clear myself about the name and definition of information biology (or bioinformatics). When I talked to D. Lipman, Director of the National Center for Biotechnology Information in NIH, USA, about the name more than 10 years ago, he mentioned that "bioinformatics" placed emphasis more on "informatics", while "information biology" placed it more on "biology". That was exactly what I thought, and since then I have used "information biology" to refer to the discipline, as I am a biologist. On the other hand, the definition may be in three holds; to study by using large amounts of biological data on genome, gene expression, proteome, protein network, protein 3D structure and others by using computer tools, to collect large amounts of such data and distribute them in the same format with annotation, and to develop computer tools for using and analyzing the data. Here, "large amounts of" is essential to define the discipline, because every researcher regularly deals with a relatively small amount of data in his/her study in biology. No one might deny strongly the definition, but what is the scientific basis of information biology? You will see the answer later.

Then, the question is, when information biology was born as a scientific discipline. Perhaps, no one knows exactly about that, though we know the parents. They are biology and informatics both of which mate to give birth to it. But, when? In my bold opinion, it was born in 1970, when the method of protein sequence alignment was published[1]. By that time protein sequence data were collected and compiled by M. Dayhoff and her colleagues[2]. Thereafter, it has been possible to compare a protein sequence to another on computer. Then, who did invent the method of protein sequencing? In 1953 F. Sanger published a couple of papers in which he disclosed the 51 amino acids of insulin that were completely sequenced by his own method[3,4]. For that endeavor he won his first Nobel Prize in Chemistry in 1958. Since the publication, his method was employed for sequencing many proteins for many species such as immunoglobins, hemoglobins and cytochrome c for man, mammals, birds, yeast and others.

It is noted that the comparison makes sense, solely because the two proteins in question were diverged from their common ancestral protein in evolution. With this recognition, we published a paper in 1997[5], in which we stated that the scientific basis of information biology was evolution, molecular evolution, in particular. It may be evident that every gene (and thus protein) is a product of evolution. In that sense, every biology researcher should pay attention to evolution no matter what subject he/she works on.

In 1977 two papers were published by a British and a US research groups[6,7]. The British group was led by F. Sanger, and the US group by W. Gilbert. The papers reported on the DNA sequencing methods they independently invented. For those achievements Sanger won his second Nobel Prize for Chemistry together with W. Gilbert and P. Berg in 1980. What a scholar Sanger is! When a bioinformatics research center was built in Hinxton nearby Cambrige, UK, they asked Sanger to be the director but he declined that, and reluctantly agreed that the name to be Sanger Center (now Sanger Institute). A friend of mine, G. Cameron, then Director of European Bioinformatics Institute, who knows Sanger well once told me about him saying "I don't know any other persons who think themselves so modestly." He also said "You won't find him at home, because he goes fishing every day."

The sequencing methods were welcomed worldwide, but disappointed me. In 1976 we published a paper in which we showed our method to estimate the evolutionary distance between two amino acid sequences in the number of nucleotide sites. On the contrary, in 1969 T. Jukes and C. Cantor already published a paper in which they derived a method of estimating the evolutionary distance between two DNA sequences[8]. It is used even today as the Jukes – Cantor method. You need an appropriate time to do something, but the time often is out of your control.

Since the publication of the two sequencing methods, a number of laboratories have begun nucleotide sequencing, and reported their results in Nucleic Acids Research (NAR) and other journals. Actually, NAR was launched for the purpose of data-sharing of nucleotide sequence data produced in laboratories worldwide. However, researchers kept on sequencing and submitting more and more data, and as a result, NAR was no longer able to cope with them. (It is noted that NAR now reports on the selected biological databases in the first issue of the year.)

Then, in 1980 some researchers had a meeting in Strasbourg, and discussed that problem. The conclusion of the meeting was to establish a database to collect, edit and distribute the nucleotide sequence data produced under the two major regulations. One was that every researcher who sequenced regions of a genome and whished to submit a related manuscript to a journal had to submit the sequence data beforehand to the database in the form provided by it, obtain the accession number of the sequence from it, and list it in the manuscript. The accession number makes other researchers easy to assimilate the sequence data when released from the databank. The other was that the submitters should agree that the data would immediately be released worldwide from the database, when the manuscript was published. The database called the EMBL nucleotide database began functioning in Heidelberg, Germany in 1982 (now called European Nucleotide Archive and located in Hinxton) followed GenBank in USA in 1984, and then DNA Data Bank of Japan (DDBJ) in 1987. The three databases have since collaborated for the daily data exchanges among them by using the common format and annotation, and immediate release of the

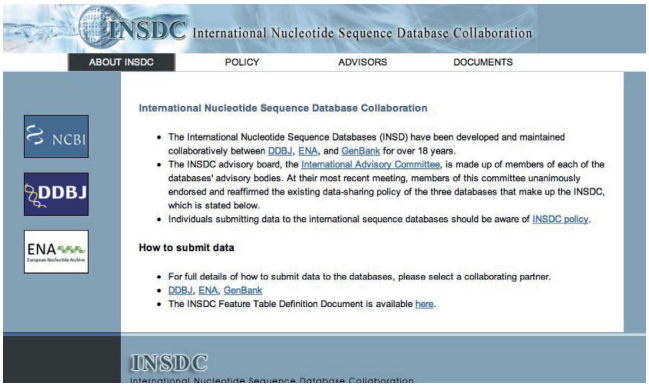


Figure 1. The homepage of INSDC that explains its members and roles.

data corresponding to publication of the manuscript. The three databases are collectedly called now the International Nucleotide Sequence Database Collaboration (INSDC) (Figure 1). When the human genome sequence was published finally after the laborious (so laborious and routine that S. Brenner once wrote that the job should be carryout our in prisons.) and time/money consuming processes[9], INSDC immediately released it worldwide. Perhaps, no other organizations could make the immediate release possible.

Meanwhile Lipman and his collaborators published a paper in 1980 in which they introduced a new method of searching similar sequences to a given sequence at INSDC, called BLAST[10]. Then, in 1994 Gibson and his colleagues reported a method of aligning multiple nucleotide (and amino acid) sequences called ClustalW[11].

The environment for the development of information biology was thus set up. Now, every researcher interested in using nucleotide or amino acid sequences in his/her study can get access to INSDC with no restrictions, as long as his/her PC is connected to the Internet.

REFERENCES

1. 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. *J Mol Biol* 48, 443-453.
2. Dayhoff, M. O. (1969). Atlas of Protein Sequence and Structure Washington, D.C.: Silver Spring : National Biomedical Research Foundation.
3. Sanger, F., and Thompson, E. O. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* 53, 353-366.
4. Sanger, F., and Thompson, E. O. (1953). The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J* 53, 366-374.
5. Tateno, Y., and Gojobori, T. (1997). DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res* 25, 14-17.
6. Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 74, 560-564.
7. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74, 5463-5467.
8. Jukes, T. H., C. C. R. (1969) Evolution of protein molecules. In: M.H. N., editor. Mammalian protein metabolism. New York: Academic Press. pp. 21–132.
9. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
11. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.