Visual presentation of complete genomic DNA sequences, and its application to identification of gene-coding regions

By Shin-ichi MAKINO,*) Naoki AMANO,** and Masashi SUZUKI***)

(Communicated by Setsuro EBASHI, M.J.A., Dec. 13, 1999)

Abstract: A new method is described for printing nucleotide sequences by assembling square micro-cells, representing nucleotides, into blocks, and by differentiating micro-cells following a color code, a gray scale, etc., corresponding to the four types of nucleotides. By printing micro-cells, whose sides are close to 250 μm, using a laser printer, the complete sequence of a genome of the size close to 1 M bases can be printed in the space of full size of A4, without losing any original information. By enlarging the print using a magnifying glass or a photo-copy machine, the original sequence can be reconstituted, visually or mechanically. Overall and regional characteristics of genomes (e.g. distribution of gene-coding regions) can be analyzed using a plot of this type. The scheme for classifying micro-cells can be modified, in order to present different types of information such as distribution of purines and pyrimidines or that of different types of dinucleotide combinations through the genome.

Key words: Gene identification; genetic information; genome analysis; genome comparison; nucleotide sequence.

Introduction. According to the editorial office of Proceedings of the Japan Academy, in a full page of this journal, at maximum, 5,500 of the narrowest characters, such as 1 and I, are printed in the form of text. By reducing the sizes of characters further, and by minimizing the gaps between lines, it is possible to print a larger number of characters in the same space. However, by ordinary printing methods, it is unlikely that the number would largely exceed 55,000.

The numbers of nucleotides composing complete genomic sequences range from the order of 10^6 (M) to that of 10^9 (G). The eubacterium Mycoplasma pneumoniae has a genome in the size close to 1 M bases (816,394 bases). In order to print this sequence completely by an ordinary method, this journal will need 15~150 pages. A human genome is composed of 3G bases, and for printing the sequence completely, 55,000~550,000 pages are needed.

Because of the large space necessary, of the scientific papers so far published reporting the determination of complete genomic sequences, none has been able to provide the most fundamental and original information, namely printed representation of the determined nucleotide sequences. Genomic sequences can be handled using a computer. Still, it is not easy to summarize the overall characteristics of a genome, by following line after line, and by scrolling page after page. It is even more difficult to visually compare two genomic sequences. In this paper, a new method for printing a genomic sequence is discussed, aiming to improve the above situation.

Micro-cellular presentation of nucleotide sequences. A symbol that has a detailed substructure will be difficult to reduce. For the best economy of the space, the simplest symbols, such as circles and squares, need to be used. Upon discriminating symbols by their shapes, their substructures are compared, and thus only fraction of the symbols. In contrast, the whole part of symbols is used, upon discriminating symbols of the same shape by their expression—e.g. by coloring, shadowing, etc. The differences needed to be fully expressed by using the maximum space available for the easiest recognition. And thus, in order to avoid creating gaps between symbols, choice of squares seem to be

---

*) AIST-NIBHT CREST Centre of Structural Biology, 1-1 Higashi, Tsukuba 305-0046, Japan.
***) Doctoral Program in Medical Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba 305-0006, Japan.
****) Graduate School of Human and Environmental Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan.

Correspondence to: M. Suzuki at AIST-NIBHT.
Fig. 1. A micro-cellular presentation, in color, of the complete genomic DNA sequence of *Mycoplasma pneumoniae* of 0.82 M bases. Different colors are used for differentiating the square micro-cells of the size of 295 μm, depending on the types of the corresponding bases; red for A, black for G, green for T, and white for C. The triangle indicates the termination of the sequence. A laser printer (Epson, LP-8000C) was used for printing the original. The size of the micro-cells is 295 μm, and the density is 86 squares per inch (spi). In each line 558 micro-cells are assembled.
In order to print close to \(10^6\) micro-squares into the space of A4 paper, the separation between neighboring micro-squares, and thus, the size of micro-squares need to be close to 250 \(\mu\)m. This can be satisfied by using a high quality laser printer (e.g. Epson, LP-8000C, having a resolution of at least 600 dpi). By printing squares of 254 \(\mu\)m, with the density of 100 squares per inch (spi), in four colors (red corresponding to A, black corresponding to G, green corresponding to T, and white corresponding to C), the genomic sequence of Mycoplasma pneumoniae, of 0.82 M bases was printed completely in a single A4 size sheet. Instead of reproducing the A4 version, in this paper, another version printed using micro-squares of 295 \(\mu\)m is reproduced, since the size available for each page (220 mm \(\times\) 165 mm) is smaller than that of A4, and thus, in any case, at least two pages are needed.

The resolution of the micro-squares was confirmed visually by using a magnifying glass. In the original print the micro-squares were well separated, so that the original information was not lost (an enlargement of part of the original print made using a photo-copy machine being shown in Fig. 2). Our expectation is that the quality of Fig. 1 reproduced in this journal is high enough. The reader is asked to examine Fig. 1 using a magnifying glass (note that the size of micro-squares in other figures, Figs. 3 and 4, is 254 \(\mu\)m and thus is smaller than that in Fig. 1).

Visual identification of gene-coding regions. The micro-cellular presentation of the genomic sequence of Mycoplasma pneumoniae (Fig. 1) shows characteristics very unlike those expected to a similar presentation of a random combination of bases (Fig. 3b). Very notable are, although not totally distinct, what looks like stripes in red and black, or in green and white that run horizontally; the reader is asked to rotate Fig. 1 by 90° in order to recognize these characteristics. Regions in genomes, that code genes, tend to have larger numbers of purines (A in red and G in black) than those of pyrimidines (T in green and C in white). Thus, regions coding genes along the strand whose nucleotide sequence is shown in Fig. 1, seem to be the origin of the reddish/blackish stripes, while regions coding genes along the complementary strand are the origin of the greenish/whitish stripes.

While producing Fig. 1, 558 micro-squares were assembled in each line. This number was changed to 150 for printing part of the genomic sequence of Pyrococcus sp. OT3 (Fig. 3a).\(^3\) The average number of bases in each coding region in this genome is close to 1,000.\(^3\) If a similar average number of bases is expected to gene-coding regions of M. pneumoniae, in Fig. 1 not all the coding regions are expected to span from the left end to the right end, but in Fig. 3a, many gene-coding regions are expected to span fully and multiple times. In Fig. 3a, broader and clearer stripes can be recognized.

Color printing is not an absolute necessity for printing nucleotide sequences, or for identifying gene-coding regions. The same type of presentation can be produced in black and white, by using tones along a scale for differentiating the micro-squares (Fig. 4). Unlike a presentation in color, the four types of bases are now lined along a single scale, by a particular type of bases being expressed closer to a particular type. Thus, the choice of expression of the four types along a gray scale is important. In Fig. 4a, where purines are printed darker than pyrimidines, stripes corresponding gene-coding regions, are recognizable, while in Fig. 4b, where C and A are darker than G and T, stripes are not seen.

A micro-cellular presentation can be produced by grouping the four types, A and G to purines (R), T and C to pyrimidines (Y), G and C to S, or A and T to W, etc. Alternatively, it can be used for presenting other types of information—e.g., the combination of each base to the following one in terms of its Y/R type, AA being RR, TA being YR, etc.

Number of symbol types used for presentation. Conventionally, four types of symbols are used for printing nucleotide sequences, corresponding to the four types of nucleotides composing DNA. If 16 types of symbols are used, instead, by correlating each symbol not to a single nucleotide, but to a dinucleotide, the total number of symbols used for presenting a sequence will be decreased to half. However, visual identification of 16 colors in a presentation similar to Fig. 1 appears to be difficult.

In general, by using 4\(^n\) types of symbols, the length of a genetic message decreases to \(1/n\). Thus, to shorten
100 pages of a nucleotide sequence, that is printed using four types of symbols, to a single page by this method, symbols of $4^{100}$ types, namely on the order of $10^{60}$ types, are necessary. An examination of systems for writing languages hints that creation of so many types of symbols is unrealistic.

For writing English language, 26 types of phonograms are used. Even if capital cases are counted distinctly from lower cases, characters with modification—e.g. underlined, are used, and some additional symbols such as numbers are included, the number of the types of symbols does not increase beyond the order of $10^5$. Approximately 50 characters are used in each of the two Japanese *kana* systems. Thus, to create $10^{60}$ types of symbols, phonograms found in over $10^{58}$ different languages need to be combined.

A larger number of symbols are used in a system of ideograms—e.g. the Chinese character system. Still, it is unlikely that the number of Chinese characters will exceed the order of $10^4$. In addition, many ideograms are composed of multiple numbers of parts, *i.e.* phonetics. An ideogram composed of 4 phonetics, needs the space the same as that is needed for 4 phonograms. In terms of the economy of the space, ideograms are not much better than phonograms.

In the known biological system for replicating genetic messages, the symbols in each pair used *in vivo* for coding the information have the complementary relationship, so that, in general, the number of types of symbols were to be kept to $2^N$, where $N$ is a natural number. If, in a hypothetical situation, the number could be increased, then the length needed for each genetic message would decrease, thereby producing advantages in terms of its maintenance, replication, and repair. However, increase of the number $N$ would create additional complications in the metabolic system; for example, for the synthesis of the larger number of symbol types. Thus, the two requirements need to be balanced. If the difficulty in maintaining a genetic message is proportional to its length, and if the difficulty in maintaining symbol types is proportional to the number of the types, the overall difficulty will change proportional to $N/\ln N$. This function has the minimum where $N$ is $e$, the base of natural logarithm (2.71828 ...), which is not so different from the $N$ number of 2 found universally in the known biological systems.

**Acknowledgements.** This work was supported by the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Corporation (JST).
MS thanks Prof. Dickerson and Prof. Oshima for their inspiring him graphical presentation of nucleotide sequences. In December 1995, Prof. Dickerson, in his room at UCLA, showed MS a board made in order to explain amino acid replacement in hemoglobin. On this board colored wooden blocks, each of which represented one of the 20 types of amino acid residues, were assembled to compose the sequence of the protein. By replacing a block at a particular position to a block of another type, an amino acid replacement was introduced. In August 1979, at Biophysics Summer School held in Shimabara, Prof. Oshima informed his research of visual interpretation of the nucleotide sequence of φX174 to MS. MS believes that, while seeking for a method of printing a genomic sequence in a limited space, he was still under the influence of his impression obtained from these experiences.

References


Fig. 4. Micro-cellar presentation, in black and white, of the sequence the same as in Fig. 3a. In (a) and (b) the four types are positioned in different orders along a gray scale.