A New Method for Analysis of DNA and Protein Sequences by 2D-Pattern Formation with Coloration

Tetsuhiko Yoshida amadeus@bbs.bekkoame.or.jp
Nobuaki Obata obata@math.nagoya-u.ac.jp
Kenji Oosawa kenji@math.nagoya-u.ac.jp
Graduate School of Polymathematics, Nagoya University
Chikusa, Nagoya 464-01, Japan

One of the most important aspects of analyses for the data bases of DNA and protein sequences is searching for sequences with specific properties. We have been developing a new method for this purpose. The method involves 2D alignment of a sequence and coloration of the aligned sequence. The most impressive character of this method is putting in parallel of 2D alignments of a sequence with different numbers of columns. By this arrangement, some signals from weak patterns can be easily discriminated by visualization of color patterns. We used this method for analyses of DNA and protein sequences in data bases. For DNA sequences, we found that there were some differences in patterns between coding and non-coding regions of cDNA sequences, for example, the cDNA sequence of human G protein beta subunit (Figure 1; cswb86). Also, we found that there were some tandem repeat sequences with different repeat length in *Escherichia coli* (bpbpbrcgrmgdkgrms97) and yeast genomes (bkzvcfhokbsds95). For protein sequences, there were specific repeating patterns for alpha-helix and collagen-helix structures from myosin heavy chain (ngs91) and procollagen (tkssbjp88), respectively. Also, we found that there were some tandem repeat sequences, for example, UDP-N-acetylglucosamine acyltransferase (rr95). Flexibility of coloring methods allowed us to color sequences based not only on one letter, e.g., A, T, G, C for DNA sequence, but on multiple letter combinations, e.g., a specific codon for an amino acid. Therefore, we can treat genomic DNA data bases both for DNA and protein structures. Furthermore, we can use various indices for coloration, e.g., hydrophobicity index of amino acids. We are now searching tandem repeat sequences with relatively long repeat periods from the *Escherichia coli* genome.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Exploratory Research (09874043 and 09878140) from The Ministry of Education, Science, Sports and Culture in Japan.

Bibliography

1. Blattner, F. R., Plunkett, G. III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y., "The complete genome sequence of Escherichia coli K-12," Science, 277: 1453-1474, 1997.

2. Bussey, H., Kaback, D. B., Zhong, W. W., Vo, D. T., Clark, M. W., Fortin, N., Hall, J., Ouellette, B. F. F., Keng, T., Barton, A. B., Su, Y., Davies, C. J. and Storms, R. K., "The nucleotide sequence of chromosome I from Saccharomyces cerevisiae," Proc. Natl. Acad. Sci. USA, 92: 3809-3813, 1995.

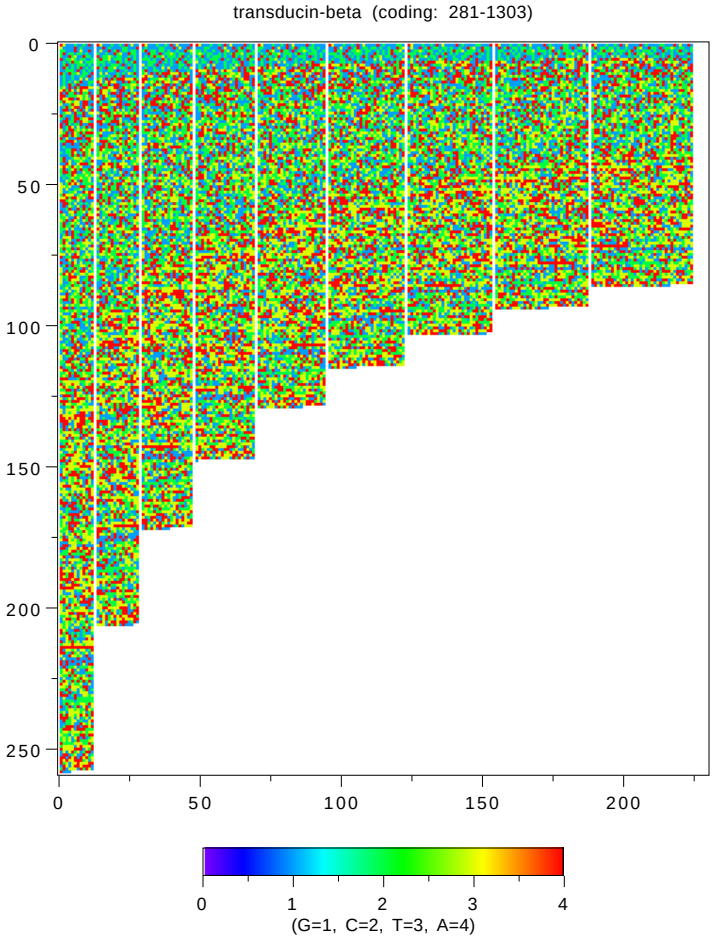

Figure 1: 2D-pattern formation with coloration of human G protein beta subunit cDNA sequence. The numbers of Columns were set to 12, 15, 18, 21, 24, 27, 30, 33, 36 from left to right of figure. Coloration was set according to nucleotide as shown in figure. The coding region is from 281 to 1303 from 5' end of cDNA.

3. Codina, J., Stengel, D., Woo, S. L. C. and Birnbaumer, L., "β-Subunits of the human liver Gs/Gi signal-transducing proteins and those of bovine retinal rod cell transducing are identical," FEBS Lett., 207: 187-192, 1986.

4. Nyitray, L., Goodwin, E. B. and Szent-Györgyi, A. G., "Complete primary structure of a scallop striated muscle myosin heavy chain. Sequence comparison with other heavy chains reveals regions that might be critical for regulation," J. Biol. Chem., 266: 18469-18476, 1991.

5. Raetz, C. R. and Roderick, S. L., "A left-handed parallel β helix in the structure of UDP-*N*-acetylglucosamine acyltransferase," Science, 270: 997-1000, 1995.

6. Tromp, G., Kuivaniemi, H., Stacey, A., Shikata, H., Baldwin, C. T., Jaenisch, R. and Prockop, D. J., "Structure of a full-length cDNA clone for the preproα1(I) chain of human type I procollagen," Biochem. J., 253: 919-922, 1988.