A Novel Application of GeneMark-RC to the Analysis of Prokaryotic Genomes and Human cDNAs: Sequence Data with Statistical Deviations Are Rich in Important Biological Information

Makoto Hirosawa [1] **Osamu Ohara** [1] **Katsumi Isono** [2]

hirosawa@kazusa.or.jp ohara@kazusa.or.jp isono@biol.kobe-u.ac.jp

[1] Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba 292 Japan
[2] Department of Biology, Faculty of Science, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan

1 Defining a self-consistent data-set for coding region assignment

Assignment of coding regions is the first step in genome sequence analysis. Although the complete genome sequences of sixteen organisms have already been published, the strategies adopted for coding region assignment are different from one organism to another, and consequently the data are not readily suited for direct comparative analysis. Evidently, a more unified method for defining coding regions must be established. For this purpose, we introduced the concept *self-consistency* in our previous study in which we performed coding region assignment within the Synechocystis genome [2].

A *self-consistent* set of ORFs can be defined as a set of ORFs, if the statistical parameters derived from it and used in conjunction with a gene-finding program result in the prediction of the same set as a set of likely coding regions. We selected GeneMark [1] as the gene-finding program in our analysis to obtain a *self consistent* set, and named the procedure GeneMark-RC. In the same framework, classification of coding regions was performed, which enabled us to distinguish between different sets of ORFs, exhibiting general, typical and atypical traits of the species. With the statistics for atypical ORFs, the coding regions of exogenous origins can also be included in the resultant set.

However, even if the procedure can be regarded as adequate for computational analyses, it might not be acceptable if the defined set thus obtained is very different from the published set of ORFs in databases that have been established as a result of complex analyses. In the case of our analysis of Synechocystis, the final set of coding regions we defined matched 98% of the annotated set. Similar levels of matches were obtained with other organisms.

2 Mining valuable information out of statistically deviated data

A characteristic feature of our procedure is that it uses more than one homogeneous set of statistic parameters to evaluate the score (up to 1.0) for a specified coding region. The refined statistics makes the score of a coding region extremely high. Deviation from it does not necessarily mean a problem in our procedure, but rather it suggests the presence of valuable information for the coding region as described below.

2.1 Unusual ORFs in the Synechocystis genome

Generally, the scores obtained for long coding regions were extremely high. There were some exceptions, however. Of the scores evaluated for 33 long annotated coding regions (more than 3000 bp) of Synechocystis, the worst five are listed in Table 1. As can be seen, the score was lowest for sll3600.

Table 1: Annotated long coding regions with low evaluation.

ID	Length	Score	From	To	Direction	Homologues
sll1360	3351	0.5418	1065660	1069010	complement	DNA polymerase III subunit (dnaX)
sll2005	3237	0.7427	1264328	1267564	complement	DNA gyrase B subunit (gyrB)
sll1951	5226	0.7661	1422426	1427651	complement	hemolysin
slr0323	3129	0.8169	2274868	2277996	direct	alpha mannosidase (ams1)
slr0744	3006	0.8809	134437	137442	direct	initiation factor IF-2 (infB)

However, the value was not evenly low throughout the coding region: instead it was almost zero in a region of longer than 1200 bp, which was found to correspond to an intein that was experimentally determined [2, 4]. Similarly, within the second lowest, sll12005 [4], an intein has been predicted.

2.2 Analysis of human cDNAs

In principle, the concept of self-consistency can be applicable to the analysis of any protein coding sequences. Thus, we recently developed a new version of GeneMark-RC for the analysis of human cDNAs. A typical cDNA derived from mature mRNA is generally expected to carry only a single coding region with high evaluation score. In other words, coding regions with low evaluation scores are likely to contain regions with unusual characteristics as observed with prokaryotic genes.

In addition, the introduction of GeneMark-RC to cDNA analysis has opened a way to detecting spurious coding interruption caused by cloning artifacts. The well-known origins of cloning artifacts, such as frame shifts and nonsense mutations in protein coding sequences, are caused by the low fidelity of reverse transcriptase. Although these cloning artifacts can be eliminated by analyzing multiple cDNA clones, it is unpractical to do so in a large-scale cDNA sequencing project. GeneMark-RC detects these artifacts by revealing the occurrence of multiple coding regions in a cDNA sequence. Locations of possible cloning artifacts can be easily identified by viewing evaluation scores along the cDNA sequence or by simple statistic calculations.

We performed GeneMark-RC analysis with the 441 long cDNAs isolated from human brain (KIAA269 to KIAA710) in our cDNA project at the Kazusa DNA Research Institute [3]. The detection sensitivity of cloning artifacts was found to be very high which was confirmed by comparing experimentally revised cDNA sequences with the predictions given by GeneMark-RC. For example, nonsense mutations were detected in 3 cDNAs with high certainty. In particular, nonsense mutations were detected twice in the cDNA sequence of KIAA0443, which was experimentally supported.

Currently, all the cDNA sequences obtained in our project are routinely subjected to GeneMark-RC examination for possible cloning artifacts. If a possibility is detected, then the suspected cDNA sequences are experimentally re-examined to avoid wrong assignment of protein coding regions.

References

[1] Borodovsky, M. and McIninch, J.D., GENMARK:Parallel gene recognition for both DNA strands, *Computer Chemistry*, 17:123-133, 1993.

[2] Hirosawa, M. and Isono, K., GeneMark-RC, a recursive procedure with self-consistency evaluation for the detection and classification of ORFs; its application to the analysis of prokaryotic genomes, *Genome Informatics 1997*, Universal Academy Press, 197–206, 1997.

[3] Ishikawa, K. *et al.*, Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro, *DNA Res.*, 5:169–176, 1998.

[4] The New England Biolabs, Intein Database. http://www.neb.com/neb/frame_NEB.html, 1997.