Genomics 99 (2012) 138–143

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Statistics of N-terminal alignment as a guide for refining prokaryotic gene annotation

Naoki Sato *, Naoyuki Tajima

Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153-8902, Japan

ARTICLE INFO

Article history:
Received 19 July 2011
Accepted 19 December 2011
Available online 29 December 2011

Keywords:
Cyanobacteria
Genome annotation
Genome clustering
Initiation codon
Synechocystis sp. PCC 6803

ABSTRACT

Identification of a correct N-terminus of a protein is an important step in genome annotation. However, we sometimes encounter incorrectly annotated N-termini in genomic databases. We analyzed statistics of surplus or missing N-terminal amino acid residues in tentatively translated coding sequence of cyanobacterial database entries, and found that, on average, about 8–9% of the aligned proteins have a putative incorrect N-terminus, although the percentage was dependent on the database entry. In an attempt to find more plausible N-termini for these proteins, we were able to estimate a better-aligning N-terminus in 90% of the cases. TTG was found as a putative initiation codon in most cases of recessed N-termini. This statistical approach, applicable to any group of prokaryotes, will help identify a plausible translation initiation site for each protein-coding gene in newly sequenced genomes, and also is a method of refining the N-terminus of proteins in already published genomes.

1. Introduction

Determination of a correct initiation site of a gene or an N-terminus of a protein is an important step in genome annotation. In biochemical analysis of a protein or in genetic analysis of a gene, identification of a correct initiation site is crucial. This problem is especially important in the current trends in rapid and massive sequencing using 'new generation sequencers'. The N-terminus of a protein may be determined if it is purified in a large amount, but in most cases, the N-terminus is estimated by computational methods based on DNA sequence. In a typical flow of data processing in annotation of a bacterial genome, a coding sequence (CDS) is identified by finding an initiation codon and a termination codon. Selection of a termination codon is not a problem if a CDS of reasonable length can be found. However, initiation codon is theoretically difficult to select, because initiation codon may not be ATG (in DNA sequence), but GTG and TTG are also known to serve as an initiation codon [1]. Large-scale experimental determination of N-termini is now performed as part of proteogenomics [2,3]. But, production of valid data is still limited to abundant proteins.

There is no confident purely bioinformatic method of identification of the initiation site. The presence of a Shine–Dalgarno sequence (or ribosomal binding site) is not universal, and Shine–Dalgarno sequence is not the only element that defines the site of translation initiation [1,4]. However, some computational methods involving machine learning can give reasonably good results [5], and was applied to develop software such as MetaGeneAnnotator [6]. Even with such good tools, we cannot be finally convinced if the selected initiation codon is correct. For this purpose, comparison with orthologs in closely related organisms will be helpful.

In any genomic databases, incorrectly annotated initiation codons are inevitably present. We sometimes encounter a situation, in which a protein sequence in the database does not align well with closely related orthologs. In many cases of protruding N-terminus (we use this expression to indicate a situation, in which there are surplus amino acid residues in the N-terminus), we can find a plausible initiation codon, just by looking at the alignment. If the protein lacks a long stretch of amino acid sequence in its N-terminus with respect to other orthologs, we will have to search a possible initiation codon in the genomic sequence, but this is time-consuming and error-prone.

In the present study, we analyzed the alignments of all proteins in the currently available cyanobacterial genomes to detect ill-aligned N-terminal sequences. Cyanobacteria were selected because they are a group of highly related organisms that retain many orthologs, and more than 40 genomes have been sequenced. Based on the statistics of N-termini, we devised a simple procedure of refining the N-termini. Here, we use the word 'refinement' in the sense of a procedure to make the N-termini of homologs as consistent as possible within the databases. In the present article, 'correction' or 'corrected' are also used in this sense. To find really 'correct' N-termini, we need extensive experiments. Though very simple, our method will be helpful in both revising genome annotation in the existing database and annotating a new genome.

2. Methods

2.1. Preparation of protein alignments

We used 41 cyanobacterial genomes that had been used for constructing CyanoClust4 database [7], which were originally obtained

* Corresponding author. Fax: +81 3 5454 6998.
E-mail address: naokisat@bio.c.u-tokyo.ac.jp (N. Sato).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.12.004

from GenBank in April 2010. The details of genome database entries used in the present study are presented in Supplemental Table S2. Coding sequences that were estimated in draft assemblies of two new cyanobacterial genomes (nicknamed 'Lim' and 'Phks', respectively) were also included. To construct protein clusters, all-against-all BLASTP (version 2.2.22) analysis [8] was performed with –FF –cF –m8 options, namely, no filtering, no compositional adjustment, and table output. All 205,942 proteins in the 43 (=41+2) cyanobacterial genomes, 10 other bacterial genomes and 60 plastid genomes were clustered by the Gclust software using the parallel version 3.5.6 [9]. Gclust uses a heuristic called 'Entropy-optimized organism count method' involving several different measures of sequence similarity in selecting homologs, and constructs clusters of 'true and near orthologs' including one or several proteins per genome. Sequences in clusters were aligned by ClustalW/ClustalX version 1.8.3 [10] or Muscle version 3.6 [11].

2.2. Statistics of alignments

The statistics of an alignment containing N_0 sequences was analyzed in the following way. Let x_i be the position of the N-terminus of the i-th sequence within the alignment ($x_i=1$ if the first residue in the alignment is the N-terminus). Overall average av_0 and standard deviation std_0 of x_i were first calculated:

$$av_0 = < x_i > \quad (1)$$

$$std_0 = \sqrt{< x_i^2 > - < x_i >^2}. \quad (2)$$

For each sequence i, sequences having an N-terminus within $x_i \pm std_0$ were selected (including sequence i itself), and the average av_i and standard deviation std_i were calculated for the selected sequences (termed group S_i).

$$av_i = < x_j > (j \text{ is a member of } S_i) \quad (3)$$

$$std_i = \sqrt{< x_j^2 > - < x_j >^2} \quad (4)$$

Then, the number of sequences N_i having an N-terminus within $av_i \pm std_i$ was counted. If $N_i/N_0 < threshold$, the sequence i is marked as an ill-aligned sequence or an outlier. The index N_i/N_0 is referred to as 'majority index' hereafter for simplicity. For all analysis reported in the present paper, $threshold = 0.2$, but this can be adjusted by examining the data distribution (such as the one shown in Fig. 1). The actual calculation was performed by the *stat* option of the *getclu* command in the SISEQ package (version 1.59.41) [12], which also outputs statistics of the C-termini and the entire lengths that were calculated in similar ways.

2.3. N-terminal refinement

For refinement of N-termini, a DNA sequence file containing coding sequences as well as sequences up to 501 bases upstream of each initiation codon was used. The length of 5′ extra sequence (currently 501) was selected to match the real needs after several trials. A rationale for this is given by the following argument: the differences in N-termini in an alignment are usually within 100 amino acid residues, because, otherwise, the protein may be considered to have an additional domain and is not included in the same cluster. This is based on the methodology of Gclust, in which domain architecture is estimated using the homology information [9].

The DNA sequence file was prepared by the *cdsnuc* command of the SISEQ package using GenBank flat files as inputs. A protein sequence that was judged to be an outlier was processed in the following way. If the N-terminus is protruding with respect to other orthologs, namely $x_i < av_0$, then a putative initiation codon was searched near the position corresponding to the average N-terminus av_0. In the present study, we used GTG, TTG and CTG as alternative initiation codons. ATA was not considered explicitly in the current version of implementation, because this is a very rare codon.

For this purpose, the translation table was modified to include lower case characters 'm' for ATG, 'v' for GTG and 'l' for TTG/CTG. Other amino acids were represented by characters in upper case. In the translated sequence, lower case characters were searched from x_i to av_0. A putative initiation codon, which was nearest to av_0, was selected. If no initiation codon was found, the search went 4 residues further, but this was in only rare cases.

If the N-terminus of the outlier sequence i was shorter than those of other orthologs, then an in-frame hypothetical translation of the upstream sequence was used in the search for a possible in-frame initiation codon. The search was limited by the length of the upstream sequences provided beforehand, but very large deletion longer than 167 residues was exceptional, because the clustering of the sequences used the overall coverage of sequence similarity called 'overlap score' [9].

Finally, revised alignment was printed, and the renewed nucleic acid sequence file was prepared, containing modified 5′ termini that were made to contain exactly 501 bases with respect to the revised initiation site. All these procedures were implemented as *correct* option of the *getclu* command. Automatic refinement of many alignments can be performed by the *mass_cor* option. Finally, the results of automatic refinement were evaluated manually, one by one, using the Artemis software [13].

2.4. Implementation into SISEQ

The procedures of estimating statistics of alignment (N- and C-termini and entire length) and refining N-termini were implemented in SISEQ package version 1.59.41 or later (http://nsato4.c.u-tokyo.ac.jp/old/Siseq.html). Test data package is also available from http://nsato4.c.u-tokyo.ac.jp/old/Software/NtermRefinement.html.

3. Results

3.1. Quality analysis of N-termini

Clustering of all 205,942 protein sequences resulted in 15,405 clusters including two or more sequences, excluding 29,225 singletons. The largest cluster included 989 proteins related to two-component transcriptional regulators. The singletons included very large proteins, multi-domain proteins, fragmented proteins, and real orphan proteins. Although many of them showed some similarity to some other proteins, they were difficult to analyze in the present study. We focused on the clusters of proteins that are aligned over their entire length. To get reliable measures of statistics, we used clusters having seven or more proteins. Very large clusters containing more than 280 members were also excluded. Hence, we analyzed the statistics of the N-termini of 5165 alignments of all cyanobacterial proteins.

Figs. 1A–C show histograms of 'majority index' (see Section 2.2) for the N-terminus, C-terminus and the entire length. This measure represents proportion of sequences having an N-terminus (or C-terminus, or entire length) of similar length for each protein. If the majority index for a sequence i ($=N_i/N_0$) is high, the sequence i belongs to the majority with respect to N-termini. If the proportion of such sequences having similar length (with respect to N- or C-terminus or entire length) is small, the protein being analyzed may be an outlier. The histogram shows that there were a significant number of outlier sequences with respect to N-terminus, typically having a majority index less than 0.2. We set the threshold as low as possible to avoid unnecessary changes in the N-terminus. In this regard, the