Prediction of Readthroughs Based on the Statistical Analysis of Nucleotides Around Stop Codons*

Sanghoon Moon, Yanga Byun, and Kyungsook Han**

School of Computer Science and Engineering, Inha University, Inchon 402-751, Korea
jiap@inhaian.net, quska@inhaian.net, khan@inha.ac.kr

Abstract. Readthrough is an unusual translational event in which a stop codon is skipped or misread as a sense codon. Translation then continues past the stop codon and results in an extended protein product. Reliable prediction of readthroughs is not easy since readthrough is in competition with standard decoding and readthroughs occur only at a tiny fraction of stop codons in the genome. We developed a program that predicts readthrough sites directly from statistical analysis of nucleotides surrounding all stop codons in genomic sequences. Experimental results of the program on 86 genome sequences showed that 80% and 100% of the actual readthrough sites were found in the top 3% and 10% prediction scores, respectively.

1 Introduction

Standard decoding of the genetic information is initiated from the start codon AUG and terminated by any of the three stop codons UAG, UAA and UGA. But the standard decoding rule can be changed occasionally by an event called 'recoding' [1]. A recoding event can occur during the elongation step (frameshift and hopping) or in the termination step (readthrough) of translation [2, 3].

In the case of readthrough, reading the stop codon is suppressed since stop codons are skipped or misread as sense codons. As a result, extended protein product is made from the readthrough process (Fig. 1). The codon usage of the local sequence surrounding stop codons is not random [4, 5], and in fact the sequence context around the stop codon is the major determinant that affects the efficiency of the translation termination [4, 5, 6, 7]. Namy *et al.* [8] showed 'poor termination contexts' that are rarely used at general termination sites. In *E. coli* and *S. cerevisiae*, the upstream sequence affects the termination efficiency [9]. The downstream sequence following the stop codon also affects the termination efficiency [4]. Bonetti *et al.* [5] demonstrated that translation termination is determined by synergistic interplay between upstream and downstream sequences.

There were previous computational and/or statistical approaches to finding readthrough sites in eukaryotes, prokaryotes and viruses [8, 10, 11, 12], but they are

* This work was supported by the Korea Science and Engineering Foundation (KOSEF) under grant R01-2003-000-10461-0.
** Corresponding author.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 276 – 283, 2006.

not fully automated. Some approaches find readthrough contexts that match the well-known readthrough context 'CAA UAG CAA' or 'CAR YYA' (Y is C or T, R is A or G), but our study shows that there are more readthrough sites not conforming to the context than those conforming to it.

We developed a fully automated program that finds readthrough sites directly from genome sequences with no prior knowledge. It considers SORF (semi open reading frame) and dORF (downstream open reading frame) in all three frames (-1, 0, +1) [10, 11]. We tested the program on 86 genome sequences from a number of organisms and successfully found readthrough sites. The rest of the paper presents the method and its experimental results in more detail.

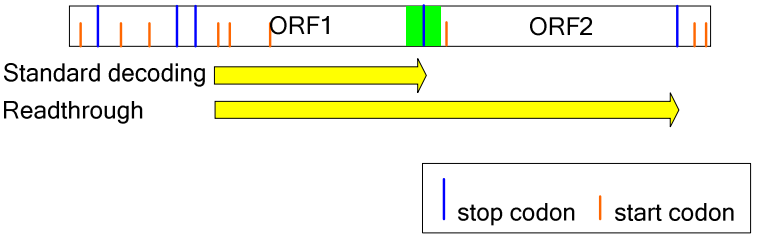

Fig. 1. In the standard decoding, translation is terminated at the stop codon. In the readthrough process, translation continues past the stop codon an extended protein product is produced.

2 Methods

2.1 Finding Readthrough Sites in Genome Sequences

Total 86 genome sequences were used as test data. 28 complete genome sequences of virus were obtained from the GenBank database [13]. 34 complete genome sequence and 24 complete coding sequences (CDSs) were obtained from the RECODE database [14].

We defined ORF1 and ORF2 in the same way as Namy *et al.* [11]. ORF1 is the area between the start codon and the first stop codon. ORF2 is the area between the first stop codon and the second stop codon. In the work by Namy *et al.* [11] the minimum length of ORF2 and the length between the first stop codon and the first start codon of ORF2 are fixed. However, different organisms have different lengths between the first stop codon and the first start codon of the second ORF. Our program allows the user to choose the length (Fig. 3D), so it is flexible to find readthrough sites of various organisms. Prediction of readthrough sites consists of five steps.

1. All genomic regions where two adjacent open reading frames (ORFs) are separated by a stop codon are identified (see Fig. 2).
2. For the second stop codon, 10 nucleotide positions (1 upstream nucleotide, stop codon and the following 6 nucleotides) are examined (colored region in Fig. 2).
3. Probabilities of A, C, G, and T are computed at each of the 10 positions, and the position specific score matrix (PSSM) is constructed from the probabilities (Fig.

3E). A score is computed for the second stop codon by equation 1 using PSSM and the probabilities of nucleotides.

4. If both ORFs do not have a start codon, the region is not considered as an open reading frame and excluded from candidates.
5. Scores are sorted in increasing order. Sites with lower scores are better ones than those with higher scores.

$$score = \sum_{k=0}^{9} \frac{p_{ki}}{p_i}, i \in \{A, C, G, T\} \quad (1)$$

where p_{ki} is the probability of observing base i at position k including the stop codon positions and p_i is the probability of observing base i at any position.

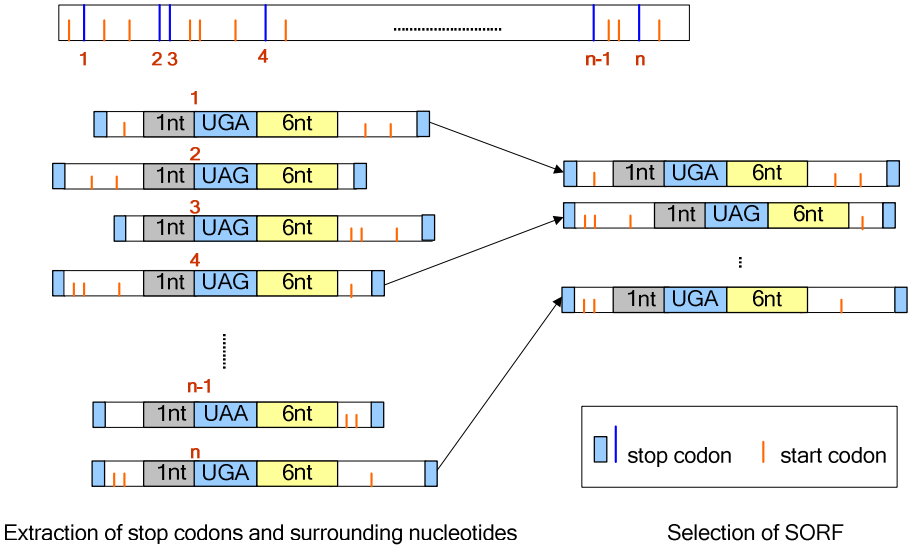

Fig. 2. Example of selecting semi open reading frames (SORF) in the genome sequence. SORFs 2, 3, and n-1 are excluded since start codons do not exist in both ORF1 and ORF2.

2.2 Implementation

The program for finding readthrough sites was implemented in Microsoft C#. As shown in Fig. 3A, stop codons of all three frames are examined by default, but the user can choose the frame that he/her wants to analyze. The user can also adjust the number of nucleotides surrounding a stop codon (Fig. 3B). Thus the user can compute the probability of the upstream and downstream nucleotides irrespective of the inclusion of stop codons (Fig. 3C). 10 in the first box of Fig. 3D indicates the minimum length of ORF1, 600 in the second box indicates the maximum length between the first stop codon and the first start codon of ORF2 and 10 in the third box

means the minimum length of ORF2. From the parameters in Fig. 3A-D, the position specific score matrix is constructed (Fig. 3E). Based on our scoring scheme the scores are sorted in increasing order (Fig. 3F). Because our program uses the data grid control, the user can move data in Fig. 3E and F easily to Microsoft Excel or notepad with the copy and paste operations. The user can also sort data by column titles. Fig. 3G shows the graphical view of the genome sequence. The user can see ORFs in all frames and analyze them easily.

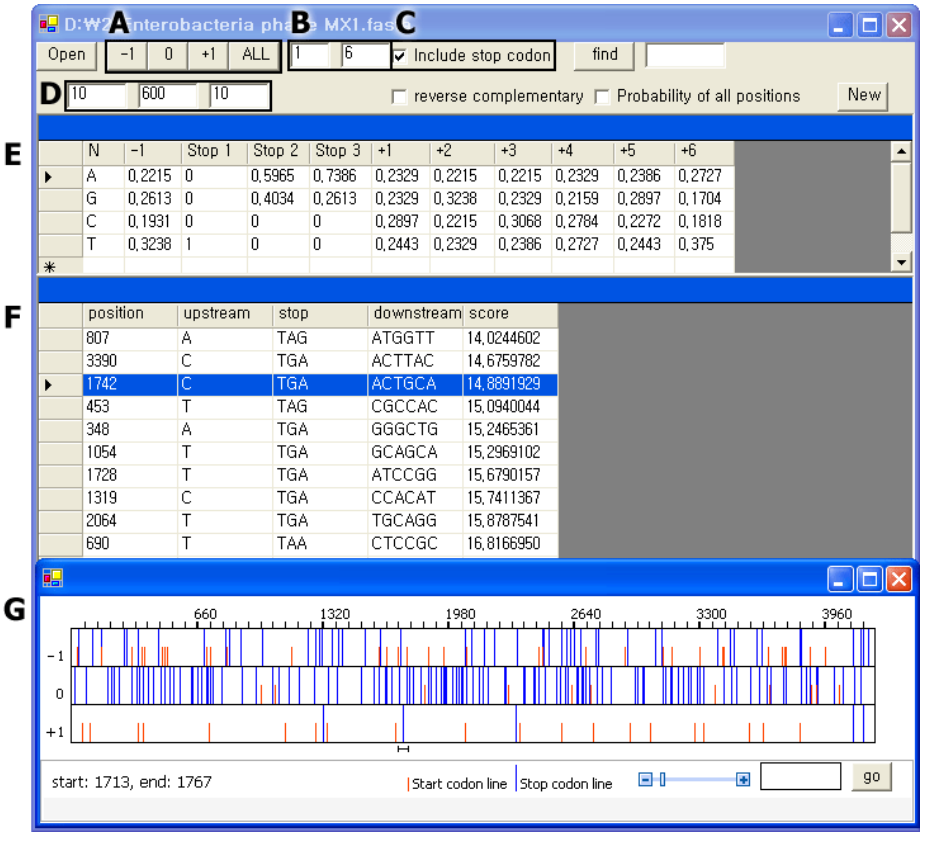

Fig. 3. Example of the user interface of our program

3 Results and Discussion

For the 86 genomic sequences we analyzed the average occurrence of each nucleotide at positions -3 to +6 from every stop codon. When we consider all stop codons in the genomic sequences, four nucleotides are observed in the positions with almost equal frequency (Fig. 4). However, the frequencies dramatically change for the stop codons in which readthroughs actually occur. Fig. 5 and Fig. 6 show sequence logos visualized by WebLogo [15] for the 86 genomic sequences and for other 26

sequences, respectively. The sequences in both data sets are known to have readthrough sites and obtained from RECODE.

Generally the role of the upstream codon for the stop codon efficiency is unclear. In Fig. 5, the nucleotides in -2 upstream codon (positions -4, -5, and -6) are distributed uniformly. In -1 upstream codon (positions -1, -2, and -3), nucleotide A is most abundant. T and A are abundant in Fig. 5. Interestingly G is very rare at position -1 (2% and 4% in Fig. 5 and 6, respectively). Most previous work on readthrough analyzed genomic sequences in terms of codons, but analysis in terms of nucleotides seems necessary in studying readthrough.

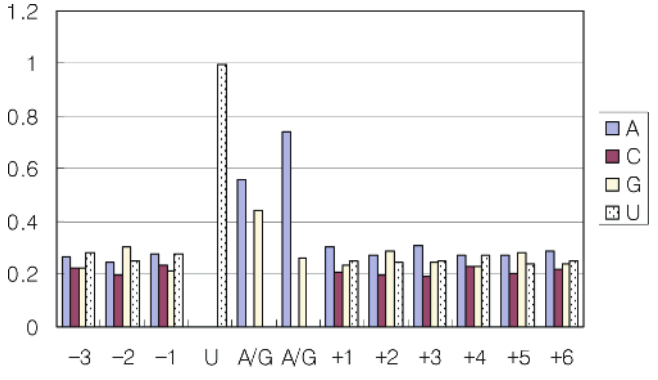

Fig. 4. The nucleotide composition around all stop codons in the 86 whole genome sequences

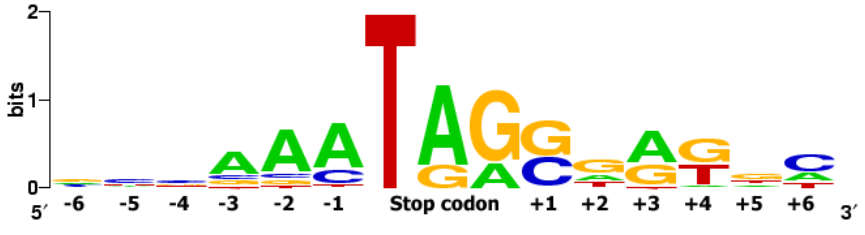

Fig. 5. The nucleotide composition in actual readthrough sites in the 86 genome sequences

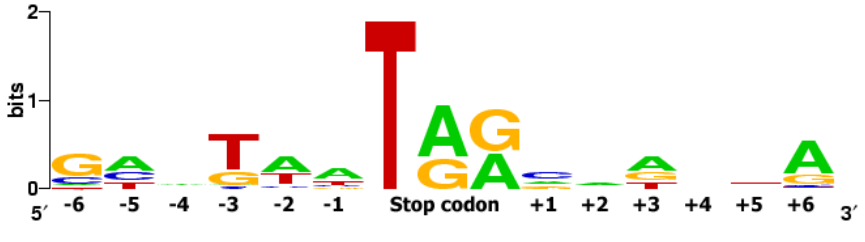

Fig. 6. The nucleotide composition in actual readthrough sites in 26 sequences from RECODE. The 26 sequences are not included in the 86 sequences used as test data.

Since G is very rare at -1 position, we examined the nucleotide at -1 position, a stop codon and the next 6 nucleotides downstream the stop codon. Tables 1-3 show the results of the analysis of stop codon readthrough. In the 86 sequences, total 23,735 (=6,416+7,343+10,154) stop codons were found. In Table 1, readthrough was not predicted for Middleburg virus (GI number: 28193965, accession number: AF339486). We obtained the complete CDS sequence of the virus from GenBank hyperlinked by RECODE, but the sequence was different from that in the RECODE database. We suppose the sequence (AF339486) does not have readthrough sites.

Table 1. Results of 24 complete CDSs from the RECODE database. Total number of stop codons: 6,416. *: no signal.

GI number	# of stop codons	Rank of the real RT in candidates (%)	GI number	# of stop codons	Rank of the real RT in candidates (%)
221091	190	4 (2.11%)	1841517	219	2 (0.91%)
332610	294	6 (2.04%)	2344756	573	1 (0.17%)
335192	153	5 (3.27%)	2582370	138	3 (2.17%)
393006	452	18 (3.98%)	3928747	191	3 (1.57%)
398066	376	10 (2.66%)	5714670	334	13 (3.89%)
409255	384	8 (2.08%)	6580858	373	11 (2.95%)
436017	286	2 (0.7%)	6580874	203	8 (3.94%)
533388	71	2 (2.82%)	7634686	394	5 (1.27%)
533391	63	3 (4.76%)	7634690	191	11 (5.76%)
786142	612	1 (0.16%)	8886896	305	4 (1.31%)
1016784	156	2 (1.28%)	10644290	147	4 (2.72%)
1236294	311	4 (1.29%)	28193965	191	*

Table 2. Results of 28 complete genomes from GenBank. Total number of stop codons: 7,343.

GI number	# of stop codons	Rank of the real RT in candidates (%)	GI number	# of stop codons	Rank of the real RT in candidates (%)
9625551	226	3 (1.33%)	20806010	366	2 (0.55%)
9625564	370	11 (2.97%)	20889313	136	2 (1.47%)
9629160	192	3 (1.56%)	20889365	337	3 (0.89%)
9629183	140	1 (0.71%)	22212887	382	5 (1.31%)
9629189	153	4 (2.61%)	25140187	311	4 (1.29%)
9635246	205	13 (6.34%)	50080143	128	6 (4.69%)
11072110	160	5 (3.13%)	30018246	214	3 (1.4%)
12018227	366	3 (0.82%)	30018252	201	7 (3.48%)
13357204	323	7 (2.17%)	33620701	839	15 (1.79%)
18254496	321	6 (1.87%)	38707974	221	5 (2.26%)
19881389	206	1 (0.49%)	39163648	174	2 (1.15%)
19919921	211	6 (2.84%)	50261346	154	1 (0.65%)
19919909	290	6 (2.07%)	51980895	228	1 (0.44%)
20153395	326	3 (0.92%)	66478128	163	3 (1.84%)

Table 3. Results of 34 complete genomes from RECODE. Total number of stop codons:10,154.

GI number	# stop codons	Rank of the real RT in candidates (%)	GI number	# stop codons	Rank of the real RT in candidates (%)
62128	323	6 (1.86%)	2801519	614	1 (0.16%)
218567	318	5 (1.57%)	3136264	308	9 (2.92%)
323338	195	6 (3.08%)	3396053	498	4 (0.8%)
331993	294	4 (1.36%)	3420022	334	9 (2.69%)
332140	222	4 (1.8%)	5442471	435	23 (5.29%)
333921	447	15 (3.36%)	5931707	368	10 (2.72%)
334100	423	11 (2.6%)	6018638	399	15 (3.76%)
335172	217	7 (3.23%)	6018642	163	4 (2.45%)
335243	374	4 (1.07%)	6143718	194	3 (1.55%)
408929	168	3 (1.79%)	6531653	178	2 (1.12%)
1335765	575	3 (0.52%)	7262472	144	2 (1.39%)
1685118	153	4 (2.61%)	7288147	438	9 (2.05%)
1752711	220	5 (2.27%)	7417288	151	1 (0.66%)
1902985	381	3 (0.79%)	10801177	246	20 (8.13%)
2213430	155	8 (5.16%)	30027702	151	3 (1.99%)
2231198	168	3 (1.79%)	30027703	218	1 (0.46%)
2801468	309	3 (0.97%)	30267510	195	4 (2.05%)

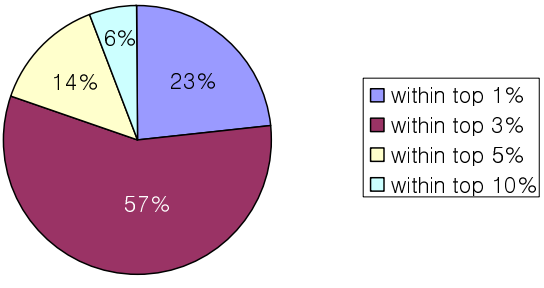

Fig. 7. The proportion of actual readthroughs in candidate readthroughs around all stop codons in the genome

It should be noted that only 0.36% (85 sites) of the total 23,735 stop codons are actual readthrough sites, and the actual sites were found with high prediction scores. 23% and 80% (=23%+57%) of the actual readthroughs were included in the top 1% prediction scores and 3% scores, respectively. All the actual readthroughs were found within the top 10% scores (see Fig. 7). Interestingly, only 23 of the 85 actual readthroughs conform to the well-known context 'CAA UAG CAA' or 'CAR YYA', suggesting that the well-known context is not sufficient for finding readthrough sites.

4 Conclusion

Finding readthrough genes is important because the readthrough process is associated with protein production and genetic control such as autoregulation. But prediction of

readthrough is very difficult because readthrough process is in competition with the recognition as a sense codon and termination as a non-sense codon.

Based on codon preference and readthrough context, we have developed a program that predicts candidate readthrough sites. Our program focused on -1 upstream nucleotide, stop codon and 6 downstream nucleotides to find readthrough sites. The program does not require any prior knowledge or manual work. It finds candidate readthrough sites from the statistical analysis of genomic sequences only. The program was tested on 86 genome sequences and successfully predicted all known actual readthrough sites. If the user provides the information of the approximate ORF length, the prediction can be done more efficiently. We believe this is the first fully automated program capable of predicting readthrough sites.

References

1. Gesteland, R.F., Weiss, R.B., Atkins, J.F.: Recoding: re-programmed genetic decoding. Science 257 (1992) 1640-1641.
2. Gestealand, R.F., Atkins, J.F.: Recoding: dynamic reprogramming of translation. Annu. Rev. Biochem. 65 (1996) 741-768.
3. Namy, O., Rousset, J., Napthine, S., Brierley, I.: Reprogrammed genetic decoding in cellular gene expression. Mol. Cell 13 (2004) 157-169
4. Poole, E.S., Brown, C.H., Tate, W.P.: The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. EMBO Journal 14 (1995) 151-158
5. Bonetti, B., Fu, L., Moon, J., Bedwell, D.M.: The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. Journal of Molecular Biology 251 (1995) 334-345
6. Namy, O., Hatin, I., Rousset, J.: Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO reports 2 (2001) 787-793.
7. Harrell, L., Melcher, U., Atkins, J.F.: Predominance of six different hexanucleotide recoding signals 3' of read-through stop codons. Nucleic Acids Res. 30 (2002) 2011-2017.
8. Namy, O., Duchateau-Ngyen, G., Rousset, J.: Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. Molecular Microbiology 43 (2002) 641-652
9. Mottagui-tabar, S., Tuite, M.F., Isaksson, L.A.: The influence of 5' codon context on translation termination in Saccharomyces cerevisiae. Eur. J. Biochem. 257 (1998) 249-254
10. Williams, I., Richardson, J., Starkey, A., Stansfield, I.: Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. Nucleic Acids Res. 32 (2004) 6605-6616
11. Namy, O., Duchateau-Nguyen, G., Hatin, I., Denmat, S.H., Termier, M., Rousset, J.: Identification of stop codon readthrough genes in Saccharomyces cerevisiae. Nucleic Acids Research 31 (2003) 2289-2296
12. Sato, M., Umeki, H., Saito, R., Kanai, A., Tomita, M.: Computational analysis of stop codon readthrough in D.melanogaster. Bioinformatics 19 (2003) 1371-1380
13. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: GenBank. Nucleic Acids Res. 30 (2002) 17-20
14. Baranov, P., Gurvich, O.L., Hammer, A.W., Gesteland, R.F., Atkins, J.F.: RECODE. Nucleic Acids Res. 31 (2003) 87-89
15. Crooks, G.E., Hon, G., Chandonia, J., Brenner, S.E.: WebLogo: A sequencer logo generator. Genome Research 14 (2004) 1188-1190