DNA Research 4, 179–184 (1997)

Prediction of Translation Initiation Sites on the Genome of *Synechocystis* sp. Strain PCC6803 by Hidden Markov Model

Makoto Hirosawa,*[,1] Takashi Sazuka,[1] and Tetsushi Yada[2]

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba 292 Japan[1] *and*
Japan Science and Technology Corporation (JST), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan[2]

(Received 14 February 1997; revised April 9, 1997)

Abstract

We developed a computer program, GeneHackerTL, which predicts the most probable translation initiation site for a given nucleotide sequence. The program requires that information be extracted from the nucleotide sequence data surrounding the translation initiation sites according to the framework of the Hidden Markov Model. Since the translation initiation sites of 72 highly abundant proteins have already been assigned on the genome of *Synechocystis* sp. strain PCC6803 by amino-terminal analysis, we extracted necessary information for GeneHackerTL from the nucleotide sequence data. The prediction rate of the GeneHackerTL for these proteins was estimated to be 86.1%. We then used GeneHackerTL for prediction of the translation initiation sites of 24 other proteins, of which the initiation sites were not assigned experimentally, because of the lack of a potential initiation codon at the amino-terminal position. For 20 out of the 24 proteins, the initiation sites were predicted in the upstream of their amino-terminal positions. According to this assignment, the processed regions represent a typical feature of signal peptides. We could also predict multiple translation initiation sites for a particular gene for which at least two initiation sites were experimentally detected. This program would be effective for the prediction of translation initiation sites of other proteins, not only in this species but also in other prokaryotes as well.

Key words: translation initiation site, Cyanobacterium, Hidden Markov Model, GeneHackerTL

1. Introduction

The complete genomic sequences of four species, *Haemophilus influenzae* Rd.,[1] *Mycoplasma genitalium*,[2] *Synechocystis* sp. strain PCC6803[3] and *Methanococcus jannaschii*,[4] have already been determined. The nucleotide sequences and potential coding regions of each species, assigned on the basis of certain assumptions, are available from GenBank/EMBL/DDBJ. However, for understanding of the complicated physiological processes in living organisms such as the control of gene expression and post-translational modification of proteins, the precise assignment of protein coding regions is indispensable. This precise assignment makes it possible to comparatively study the four species to elucidate common and unique features among them.

The precise amino-terminal sequences of a total of 96 proteins in the *Cyanobacterium*, *Synechocystis* sp. strain PCC6803, have been previously determined by micro-sequencing of two dimensionally resolved proteins of the whole cell extract.[5,6] By comparison with the genomic DNA sequences, the actual translation initiation sites of 72 proteins were assigned, but the amino-terminal of the remaining 24 proteins could not be correlated to the translation initiation sites because of the lack of potential initiation codons at the amino-terminal positions. These results strongly suggest that the latter class of proteins have been processed post-translationally. Thus, the use of a computer program may be useful for the prediction of potential translation initiation sites on the basis of information extracted from the experimentally determined proteins.

We previously reported a program, GeneHacker, which predicts the coding regions by applying the Hidden Markov Model (HMM).[7,8] We have now developed another program, named GeneHackerTL, which predicts the translation initiation sites. GeneHackerTL uses information which is extracted from a set of nucleotide sequences surrounding the translation initiation sites according to the framework of HMM.[9] HMM is a higher paradigm of the Markov Model and can express complex sequence patterns, such as the promoter sequences, better than *the regular expression*, which is used to express motifs in Prosite.[10] HMM is conceptualized as a network composed of nodes, to which the nucleotide frequency is assigned in our study, and directed paths between them, to which the transition probability is as-

Communicated by Toshihisa Takagi

* To whom correspondence should be addressed. Tel. +81-438-52-3919, Fax. +81-438-52-3918, E-mail: hirosawa@kazusa.or.jp

signed. Here we will denote information expressed in this framework of HMM as HMM. Tracking of nodes by selecting a path with the highest transition probability leads to a sequence pattern represented by a particular HMM. A simple presentation of HMM by taking promoter sequences can be seen in Fig. 1 of our previous paper.[11]

We applied GeneHackerTL for analysis of the cyanobacterial genes, whose products are assumed to be processed post-translationally, and it successfully predicted the translational initiation sites.

2. Materials and Methods

2.1. GeneHackerTL

GeneHackerTL is a newly developed computer program which predicts the most probable translation initiation site of a given nucleotide sequence. Necessary information for GeneHackerTL is as follows:

In-1:	The base composition in non-coding regions
In-2:	An HMM representing nucleotide sequences surrounding the translation initiation sites
In-3a:	The base compositions of the first, second and third positions of the amino acid-triplets in-frame in the coding regions
In-3b:	The average length of the coding regions

The first and second HMM are constructed from the information on In-1 and In-2, respectively, and the third HMM from the information on In-3a and In-3b. The three $HMMs$ are concatenated to produce an integrated HMM. GeneHackerTL identifies the sequence structure most similar to the integrated HMM in a given sequence with the Viterbi algorithm,[9] which eventually identifies the most probable translation initiation site. The score of the identified translation initiation site is calculated with the Viterbi algorithm as the similarity of the sequence surrounding the translation initiation site against the second HMM. The logarithm likelihood of the calculated translation initiation site is normalized to obtain its score by setting the threshold of the translation initiation site to 1.0.

2.2. Sequence data

The training sets, which were used for the creation of HMM surrounding the translation initiation sites (the second HMM in the previous subsection), are the 38-bp regions from nucleotide positions −25 to 13 of the 72 proteins experimentally determined (the aminoterminal position was regarded as nucleotide position +1). The nucleotide sequences from positions −1000 to the termination sites of the 72 proteins were also used in the evaluation process of the prediction rate (See 2.4).

For prediction of the translation initiation sites of the unassigned 24 proteins, the sequences from positions −1000 to the termination sites were cut out.

2.3. Prediction by GeneHackerTL

The most probable translation initiation sites of the 24 proteins, except for those corresponding to sequence nos. 29 and 74, were predicted by GeneHackerTL with the information below. The base composition in non-coding regions (In-1) was deduced from the entries D63999–D64006 in GenBank/EMBL/DDBJ. As HMM representing the sequence surrounding the translation initiation site (In-2), the HMM derived in the previous study was used [11]. The base compositions of three respective frames (In-3a) and the average length of the coding regions (In-3b) were derived from the coding regions of the 72 proteins in the dataset.

2.4. Evaluation of prediction rate by GeneHackerTL

The prediction rate of GeneHackerTL was estimated by prediction of the translation initiation sites for the coding regions of each of the 72 proteins using information extracted from the remaining 71 proteins. The base compositions in the coding regions (In-3a), the average length of the coding regions (In-3b) and parameters in HMM representing sequences surrounding translation initiation sites (In-2), except that for initiation codons, were recalculated based on the 71 remaining translation initiation sites and their coding regions.

3. Results and Discussion

3.1. The Prediction rate of GeneHackerTL

The most probable translation initiation sites of the 72 proteins were predicted based on the information of the remaining 71 sequences. The results shown in Table 1, in which the predicted positions of translation initiation sites relative to their experimentally determined positions are listed. GeneHackerTL predicted the translation initiation sites in all but one sequence, corresponding to no. 37. Of the 71 translation initiation sites predicted by GeneHackerTL, 62 sites coincided with sites experimentally determined[5] giving a prediction rate of 86.1% (62/72), and a false-positive error rate of 12.6% (12/71).

Among the nine sites differently predicted (sequences nos. 26, 42, 49, 57, 60, 81, 86, 94, and 95), the predicted site for sequence no. 49 was far upstream from the experimentally determined position. But as described in the next section, translational initiation at multiple sites was suggested for this coding sequence.

The rate of correct prediction was apparently different with the species of initiation codons. The sites of 58 out of 65 proteins initiating with ATG were correctly predicted by GeneHackerTL (89.2%), but only 4 out of 7 (57.1%) were predicted for proteins initiating with rare

Table 1. Validation of the Prediction by GeneHackerTL[a)]

SeqNo	Initiation site (predicted)		Score	SeqID	Initiation site (determined)	
	Codon	Position			Codon	Position
2	ATG	0	1.157	sll0436	ATG	2540656
3	ATG	0	1.158	slr2076	ATG	915721
8	ATG	0	1.179	sll0018	ATG	2471670
9	*GTG*	*0*	*1.140*	*sll2580*	GTG	725367
10	ATG	0	1.103	slr0131	ATG	2181013
11	TTG	0	1.075	slr2051	TTG	596498
12	ATG	0	1.101	sll0617	ATG	2646071
13	ATG	0	1.136	slr1719	ATG	1471414
15	ATG	0	1.140	slr0006	ATG	2475195
17	ATG	0	1.178	sll1261	ATG	1722006
19	ATG	0	1.142	sll1479	ATG	3395676
22	ATG	0	1.059	slr1259	ATG	1404085
24	ATG	0	1.157	*sll0807*	*ATG*	*1713256*
25	ATG	0	1.087	slr0520	ATG	2932510
26	ATG	580	1.172	slr0065	ATG	2593329
28	ATG	0	1.089	slr1198	ATG	879408
31	ATG	0	1.128	sll1621	ATG	1325404
32	GTG	0	1.097	slr1622	GTG	2034294
33	ATG	0	1.174	slr0476	ATG	2612907
34	ATG	0	1.089	sll1444	ATG	1890474
35	ATG	0	1.165	slr1516	ATG	1607353
37[b)]	—	—	—	*sll0145*	*GTG*	*2192681*
38	ATG	0	1.117	sll1825	ATG	823973
39	ATG	0	1.190	slr0676	ATG	414871
40	ATG	0	1.175	sll1815	ATG	832265
41	*ATG*	*0*	1.137	sll1577	ATG	727466
42	ATG	55	1.026	sll1316	ATG	1166728
43	ATG	0	1.101	sll1578	ATG	726837
44	GTG	0	1.075	slr0001	GTG	2464527
46	ATG	0	1.140	slr2057	ATG	1430119
47	ATG	0	1.187	sll0422	ATG	2093508
48	ATG	0	1.088	slr1251	ATG	1392972
49	ATG	−168	1.203	slr1761	ATG	538658
50	ATG	0	1.125	slr1992	ATG	1436699
51	*ATG*	*0*	*1.098*	*slr0172*	*ATG*	*2304094*
53	ATG	0	1.297	slr0243	ATG	114061
54	ATG	0	1.203	slr0737	ATG	126639
55	ATG	0	1.126	slr1034	ATG	563342
57	GTG	358	1.093	slr0145	ATG	2177900
58	ATG	0	1.173	sll1852	ATG	2234753
60	ATG	157	1.069	sll1675	ATG	247569
61	ATG	0	1.162	slr1986	ATG	1431000

Table 1. (Continued.)

SeqNo	Initiation site (predicted)		Score	SeqID	Initiation site (determined)	
	Codon	Position			Codon	Position
62	ATG	0	1.162	sll0754	ATG	2417207
63	ATG	0	1.184	slr1732	ATG	1311828
68	ATG	0	1.146	slr1894	ATG	843105
69	ATG	0	1.026	slr0455	ATG	3507060
70	ATG	0	1.182	sll0359	ATG	2146386
72	ATG	0	1.094	sll0822	ATG	2862394
75	ATG	0	1.259	ssl0020	ATG	2485183
77	ATG	0	1.150	slr0823	ATG	2961130
78	*ATG*	*0*	*1.141*	*sll1130*	*ATG*	*1008753*
79	ATG	0	1.132	slr1330	ATG	1668531
80	ATG	0	1.279	sll1398	ATG	50231
81	TTG	−60	1.131	ssl0707	TTG	2152629
82	ATG	0	1.175	sll1746	ATG	923230
83	ATG	0	1.087	ssl3093	ATG	724345
84	ATG	0	1.146	slr2075	ATG	915313
85	ATG	0	1.193	slr0012	ATG	2480477
86	TTG	88	1.024	ssr1600	ATT	2021108
88	ATG	0	1.190	ssr2837	ATG	*1982049*
89	ATG	0	1.142	ssr1480	ATG	1135337
90	ATG	0	1.118	sll0230	ATG	145469
91	ATG	0	1.092	slr1839	ATG	956275
93	*ATG*	*0*	1.168	sll1029	ATG	219077
94	ATG	70	1.153	ssl2781	ATG	1590382
95	ATG	109	1.032	slr1139	ATG	794362
96	ATG	0	1.028	ssl0563	ATG	2287579
97	ATG	0	1.127	ssl1972	ATG	390634
98	ATG	0	1.125	ssr0330	ATG	*2745934*
100	ATG	0	1.194	ssr1698	ATG	637606
101	ATG	0	1.173	sll1769	ATG	1222927
102	ATG	0	1.107	ssl0352	ATG	2735598

[a)] The most probable translation initiation sites of the 72 experimentally determined proteins were predicted by GeneHackerTL by calculating the score for each with the parameters extracted from the 71 remaining proteins. The columns from left to right represent: Protein sequence number (SeqNo.); initiation codons, initiation sites relative to those experimentally determined and their scores by GeneHackerTL; corresponding ORFs (SeqID) assigned by Kaneko et al.[3]; initiation codons and positions experimentally determined.[5]

[b)] No translation initiation site was predicted.

initiation codons, GTG, TTG or ATT. The lower rate for rare initiation codons may be because the number of training sets was too small to extract the sequence characteristics.

The false-positive error can be reduced if a score threshold higher than 1.0 is taken. When 1.1 was used as the threshold, 52 translation initiation sites had scores above the threshold and the number of falsely predicted sequences was reduced to three (sequence nos. 26, 49 and 81), giving a the false-positive error of 5.8% (3/52).

3.2. Prediction of multiple translation initiation sites

Although 9 out of 71 sites predicted by GeneHackerTL were different from those assigned experimentally, some of the proteins may have initiated at multiple sites, such as the proteins corresponding to sequence nos. 29 and 49, because the amino-terminal sequences of these proteins was assigned to the same open reading frame (ORF), named slr1761, and the amino-terminal position of sequence no. 49 starts 66 nucleotides downstream of sequence no. 29. Therefore, it is likely that these proteins are initiated at different initiation sites of the same ORF.

Thus, we tried to predict multiple translation initiation sites using Gene-HackerTL. As described above, Gene-HackerTL predicts only the most probable translation initiation site in a given coding sequence. However, it is possible to predict another translation initiation site by introducing a mutation in the predicted initiation codon followed by analysis of the mutated coding sequence. Thus, for the protein corresponding to sequence no. 49, the third position of the initiation codon was mutated from G to C. The previously predicted initiation site be-

Table 2. Psooibility of multiple translation initiation sites of sequence no. 4 9[a)]

Position	Score
−189	1 062
−168	1.199
−123	1.038
−114	1 065
−42	1.047
1	1.127

The translation initiation sites of sequence no. 49 were investigated by GeneHackerTL by introducing a mutation (G to C) into the third position of initiation codon of the predicted translation initiation site.

comes undetectable because of this mutation, which leads to prediction of another translation initiation site. More translation initiation sites can be predicted by repetition of the above procedure. The result of the calculation is shown in Table 2.

Among the six predicted initiation sites, two sites have scores higher than 1.1; the highest (1.199) was at position −168 and the next (1.127) was at position +1. These results suggest that sequence no. 49 corresponds to the protein initiated at position +1 without processing, and that the protein with sequence no. 29 is initiated at position −168. After processing, a signal peptide of 34 amino acids long is generated (Table 4).

3.3. Prediction of the initiation sites of 24 unassigned proteins

Among the proteins for which the amino-terminal sequences were micro-sequenced, the initiation sites of 24 proteins have not been assigned[5] because of the lack of a potential initiation codon at the amino-terminal position. The initiation sites of these proteins have tentatively been assigned to either ATG or GTG in the upstream region,[3] and these proteins have been grouped into categories A and B, mainly with respect to the length of putative processed peptides. We applied GeneHackerTL for prediction of the initiation sites of these proteins, and the predicted sites are indicated in Table 3 with their scores.

3.3.1. The sites for proteins in category A

Translation initiation sites were predicted by GeneHackerTL in all but sequence no. 45, and the 15 predicted sites were upstream of the experimentally determined sites. According to this assignment, all the peptides between the predicted to experimentally determined sites are abundant in hydrophobic amino acids (See Table 4), and 11 out of 15 possible signal peptides contain one to three basic amino acids (Lys or Arg) near their amino-termini (4.5 amino acid positions from the termini on average). The average length of the 11 possible signal peptides was 27.6 amino acids, and there was an average of 1.9 basic amino acids. Since signal peptides are characterized by abundant hydrophobic amino acids throughout and by basic amino acids near their amino-termini,[16,17] at least the 11 putative processed peptides carry this characteristics. The remaining four processed peptides (nos. 7, 64, 66 and 73) were hydrophobic but did not contain basic amino acids at their amino-termini. The failure to prediction sequence no. 45 may be because the characteristics of its initiation site sequence are different from those of the training sets.

Table 3. Prediction of translation initiation sites[a)]

SeqNo	Initiation site [predicted] Codon	Position	Score	SeqID	N-terminal position [determined]
#Category A					
6	ATG	−78	1.196	slr0447	2098999
7	ATG	−45	1.085	sll0319	2438734
14	ATG	−69	1.110	slr1409	1182734
16	ATG	−84	1.106	sll0427	2087045
20	ATG	−81	1 068	slt0314	2447225
23	ATG	−90	1.105	sll1785	1188795
45[b)]	—	—	—	sll1483	3392109
56	GTG	−75	1.062	sll0258	2132023
59	ATG	−84	1 104	slr1273	1865117
64	ATG	−48	1.103	sll1620	1326578
66	ATG	−39	1.250	sll0172	2310931
67	ATG	−66	1.066	slr2101	1569654
71	GTG	−90	1.095	slr1406	1181154
73	ATG	−24	1.102	sll0630	3322200
87	ATG	−108	1.114	sll1194	297780
92	ATG	−84	1 048	sll0199	2526123
#Category B					
4	ATG	−93	1.118	slr1506	1590724
18	ATG	−72	1.075	sll1782	1233611
21	ATG	25	1 015	slr1668	3476792
29	ATG	−102	1.203	slr1781	538592
52[b)]	—	—	—	sll0100	2981815
65	ATG	−183	1 033	sll0274	2102378
74	ATG	−516	1 080	sll0422	2092092
76[b)]	—	—	—	sll0756	2415820

[a)] The most probable translation initiation sites of the 24 proteins were predicted by GeneHackerTL. The columns from left to right represent: Protein sequence number (SeqNo.); initiation codons, initiation sites relative to those experimentally determined and their scores by GeneHackerTL; corresponding ORFs (SeqID) assigned by Kaneko et al.[3]; positions experimentally determined.[5]
[b)] No translation initiation site was detected by GeneHackerTL.

3.3.2. The sites for proteins in category B

The translation initiation sites were predicted for six out of eight sequences (nos. 4, 18, 29, 65, 74 and 21). The sites of the first five were located upstream of the experimentally determined site, and the lengths between the predicted initiation site and the respective amino-termini were 31, 24, 34, 61 and 172 amino acids (Table 4). The first four truncated peptides, especially the first three, were presumed to be processed as signal peptides, because of their short length and their characteristic sequence features.

Table 4. Putative processed peptides predicted by GeneHackerTL[a)]

Seq No.	Length	Putative processed peptide
#Category A		
6	26	MTNPFGRRKFLLYGSATLGASLLLKA
7	15	MAVGAILAPFTPVSA
14	23	MRIFPVFLLTFSLFLIKEEIVTA
16	28	MRFRPSIVALLSVCFGLLTFLYSGSAFA
20	28	MNLIRNRWAQIFTQSILGVLIAGGTAWA
23	30	MLLKVKLWGIGLVLTLTLGTILFLQNFSVA
45[a)]		
56	25	VKRFFLVAIASVLFFFNLMVGSANA
59	28	MFLTALRSFLLFLAVTCLSLAIAMPAWA
64	16	MGALLAVLLSGMVWFA
66	13	MICAVLFAGTAAA
67	22	MKFISSFFALATVLACQPTVFA
71	30	VKVSKRFFQFCLVLGISLGLSLALVPDALA
73	8	MAIAPANA
87	36	MKFISRLLVACSLLIGLMGFLGADLAQALTPNPILA
92	28	MSKKFLTTLAGLLLVVSSFFLSVSPAAA
#Category B		
4	31	MVTFPLNLRRWLQSVCLGALTAIAVQLPGKT
18	24	MLKQFSATFIGLLLATVGAQAAIA
21[b)]	−8	
29	34	MRDILISLTVVTFFSLVLISVAIPGKSSPSAIAA
52[a)]		
65	61	MNFLVICQKFFTFNLFPWKAIARVQREKPQSLGRWQFVVRTGILVATFILA LGSLASPSLA
74	172	MTPKLIIIGGASSLDDKGGLATVRQSLHQIVAAVYETLTAGGSAMDAVVQG CELLENEPRFNAGTGSVLQSDGQVRMSASLMDGDRQNFSGVINVSRIKNPI QMAQFLQGQTDRILSDYGADLAREMQLPIYDPATDFRIQEWMEERGEDVRK KMARLIADPTVGIEARKG
76[a)]		

[a)] No translation initiation site was detected by GeneHackerTL. [b)] The site predicted by GeneHackerTL was 8 amino acids downstream of the experimentally determined amino-terminus.

For the protein corresponding to sequence no. 74, the resulting processed peptide was much longer than those of typical signal peptides (172 amino acids), suggesting that a post translational processing other than the typical signal processing is operated. The predicted initiation site of the protein starting with sequence no. 21 was eight amino acids inside of the experimentally deduced site. As this protein is classified to the minor class of genes,[18] it suggests that an additional *HMM* element corresponding to the minor class is required for this class of genes to enhance the prediction rate of GeneHackerTL.

3.4. Comparison with other methods

In the sequencing projects of the *Escherichia coli* genome,[12,13] a translation initiation site of each protein coding region was assigned by calculating the efficiency of the ribosomal binding site using a scoring matrix.[14] This method, konwn as multiple regression analysis, is essentially based on the weight matrix. There is an other method for the assignment of translation initiation sites,[15] but this is also based on the weight matrix. From a view point of computer science, prediction of the translation initiation sites based on HMM appears to be superior to that based on the weight matrix, because the former can utilize more information which cannot be represented in the weight matrix. Conceptual comparison of the two types of methods is done here by taking the Shine-Dalgarno sequence as an example.

In our previous study[8] in which the *HMM* representing nucleotide sequences surrounding translation initiation sites was created, we identified four distinct regions upstream of the initiation site. Among them, Region U2 (See Fig. 2 in ref. 8) contains 97.2% (34/35) of the Shine-Dalgarno sequence.[5] The average length of Region U2 is 6.24 nucleotides with a standard deviation of 3.07. The average position of the center of Region U2, which ranges from −22.5 to −4.5, is −11.4 with a standard deviation of 3.7. If nucleotide sequences are represented in the weight matrix, the distinctive features of Region U2 would become vague owing to the high standard deviation of this region in position. This variance of Region U2 in position and both the prototype and variant patterns of this region can be represented in HMM, and provide a more predictive power on GeneHackerTL. Furthermore, some features detected in the previous study would completely disappear in the weight matrix because of masking by the influence of Region U2. An example is the existence of

a thymine just downstream of Region U2. The disappearance of such information in the weight matrix would reduce the predictive power of the multiple regression analysis.

The present version of GeneHackerTL is for prediction of the translation initiation sites of genes on *Synechocystis* sp. strain PCC6803. GeneHackerTL can be applicable to other prokaryotes as well if the necessary information is provided. Its application to eukaryotes should be investigated.

Acknowledgments:

We thank Dr. Takeshi Itoh of Nara Institute of Science and Technology for valuable discussion. This work was supported by funds from Kazusa DNA Research Institute and was carried out to examine the effectiveness of software developed in the Advanced Lifescience Information System project for genome analysis by the Japan Science and Technology

References

1. Fleischmann, R. D., Adams, M. D., White, O. et al. 1995, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science*, **269**, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O. et al. 1995, The minimal gene complement of Mycoplasma genitalium, *Science*, **270**, 397–403.
3. Kaneko, T., Tanaka, A., Sato, S. et al. 1995, Sequence Analysis of the Genome of the Unicellular Cyanobacterium *Synechocystis* sp. strain PCC 6803. I. Sequence Features in the 1 Mb region from Map Positions 64% to 92% of the Genome, *DNA Res.*, **2**, 153–166.
4. Bult, C., White, O., Olsen, G. J. et al. 1996, Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschi*, *Science*, **273**, 1058–1073.
5. Sazuka, T. and Ohara, O. 1996, Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC 6803 by amino-terminal protein sequencing, *DNA Res.*, **3**, 225–232.
6. Sazuka, T. and Ohara, O. 1997, Towards a proteome project of Cyanobacterium *Synechocystis* sp. strain PCC 6803: Linking 130 protein spots with their respective genes, *Electrophresis*, in press.
7. Yada, T. and Hirosawa, M. 1996, Recognition incyanobacterium genomic sequence data using the Hidden Markov Model, *in the Proceedings of International Conference on Intelligent System for Molecular Biology* (*ISMB*-96), 252–260.
8. Yada, T. and Hirosawa, M. 1996, Detection of short protein coding regions within the cyanobacterium genome: Application of the Hidden Markov Model, *DNA Res.*, **3**, 355–361.
9. Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. 1983, An Introduction to the Application of the Theory of Probabilistic Function of a Markov Process to Automatic Speech Recognition, *Bell Syst. Tech. J.*, **62**, 1035–1074.
10. Bairoch, A. 1991, Prosite: A dictionary of protein site and pattern: User manual Release 7.00.
11. Yada, T., Suzuka, T., and Hirosawa, M. 1997, Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome by using hidden Markov model, *DNA Res.*, **4**, 1–7.
12. Aiba, H., Baba, T., Hayashi, K. et al. 1996, A 570-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 28.0–40.1 min region on the linkage map, *DNA Res.*, **3**, 363–377.
13. Itoh, T. Aiba, H. Baba, T. et al. 1996, A 460-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 40.1–50.0 min region on the linkage map, *DNA Res.*, **3**, 379–392.
14. Barrick, D., Villanueba, K., Childs, J. et al. 1994, Quantitative analysis of ribosome binding sites in *E. coli*, *Nucleic Acids Res.*, **22**, 1287–1295.
15. Stormo, G. D., Schneider, T. D., and Gold, L. 1986, Quantitative analysis of the relationship between nucleotide sequence and functional activity, *Nucleic Acids Res.*, **14**, 6661–6679.
16. vonHeijne, G. 1984, Analysis of the distribution of charged residues in the N-terminal region of signal sequences: impliations for protein export in prokaryotic and eukaryotic cells, *EMBO*, **3**, 2315–2318.
17. vonHeijne, G. 1986, A new method for predicting signal sequence clearge sites, *Nucleic Acid Res.*, **14**, 4683–4690.
18. Hirosawa, M., Isono, K., William, S. H., and Borodovsky, M. 1997, Gene identification and classification in the synechocystis genomic sequence by recursive GeneMark analysis, *DNA Sequence*, in press.
19. Wilkins, M. R., Pasquali, C., Appel, R. D. et al. 1996, From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis, *BIO/TECHNOLOGY*, **14**, 61–65.