Detection of Short Protein Coding Regions within the Cyanobacterium Genome: Application of the Hidden Markov Model

Tetsushi YADA1,* and Makoto HIROSAWA2
Japan Science and Technology Corporation (JST), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan1 and
Kazusa DNA Research Institute (KDRI), 1532-3 Yana-uchino, Kisarazu, Chiba 292, Japan2

(Received 18 July 1996; revised 25 November 1996)

Abstract

The gene-finding programs developed so far have not paid much attention to the detection of short protein coding regions (CDSs). However, the detection of short CDSs is important for the study of photosynthesis. We utilized GeneHacker, a gene-finding program based on the hidden Markov model (HMM), to detect short CDSs (from 90 to 300 bases) in a 1.0 mega contiguous sequence of cyanobacterium Synechocystis sp. strain PCC6803 which carries a complete set of genes for oxygenic photosynthesis. GeneHacker differs from other gene-finding programs based on the HMM in that it utilizes di-codon statistics as well. GeneHacker successfully detected seven out of the eight short CDSs annotated in this sequence and was clearly superior to GeneMark in this range of length. GeneHacker detected 94 potentially new CDSs, 9 of which have counterparts in the genetic databases. Four of the nine CDSs were less than 150 bases and were photosynthesis-related genes. The results show the effectiveness of GeneHacker in detecting very short CDSs corresponding to genes.

Key words: Cyanobacterium; gene finding; hidden Markov model; short protein coding region; oxygenic photosynthesis genes

1. Introduction

1.1. Significance of short CDS detection

The advances in large-scale sequencing has accelerated the production of long contiguous nucleotide sequences. Today, more than 50 contiguous nucleotide sequences longer than 100 kb in length are available through the World Wide Web (WWW).1 This accumulation of the determined contiguous nucleotide sequence necessitates the development of effective algorithms to assign protein coding regions (CDSs) in nucleotide sequences.

Algorithms and programs that utilize codon usage bias within coding regions to discriminate the coding regions from the intergenic regions have been studied by many researchers.2–9 Nearly all of the programs have been developed with the human genome in mind. For prokaryote sequences, GeneMark, developed by Borodovsky et al., is being widely used for CDS identification.3

The Kazusa DNA Research Institute has determined the whole sequence (3.6 Mb) of the unicellular cyanobacterium Synechocystis sp. strain PCC6803 in 1996. The institute released 3.6 Mb of sequence data in 1995, including a continuous 1.0 Mb sequence.10,11 The determination of the complete genome of this species, which carries a complete set of relatively short genes for oxygenic photosynthesis, has led to a greater need for determining short CDSs properly. For example, the SwissProt database (Release 32.0) contains 672 cyanobacterium proteins. Of these, 170 are less than 100 amino acids (aa), and 80% of the short proteins are related to photosynthesis.

1.2. CDS assignment in Synechocystis

The assignment of CDSs in the 1.0 Mb sequence was carried out in two phases. In the first phase, a traditional scheme which utilizes the length of CDSs and a similarity search was used to assign 818 CDSs.11 In this phase, all possible CDSs longer than 150 bases were extracted from the sequence for a similarity search. Then, CDSs whose lengths were from 150 to 300 bases were classified as CDSs if their optimal scores in the similarity search were more than 100. On the other hand, CDSs longer than 300 bases were classified as CDSs irrespective of their optimal scores in the similarity search. Then, CDSs whose lengths were from 150 to 300 bases were classified as CDSs if their optimal scores in the similarity search were more than 100. On the other hand, CDSs longer than 300 bases were classified as CDSs irrespective of the results. A total of 333 CDSs were assigned without significant similarity, solely based on their length. In the second phase, the assignment of CDSs was re-examined using GeneMark. GeneMark supported 752 out of the 818 previously assigned CDSs and suggested 50 new CDSs.12

The 91.9% (752/818) CDS detection rate by GeneMark is good. However, if we focus on CDSs between 150...
GeneHacker is a gene-finding program developed by the first author (TY) in collaboration with the second

2. Materials and Methods

2.1. Sequence data

The contiguous 1 Mb nucleotide sequence from Synechocystis sp. strain PCC6803, which was registered in GenBank/EMBL/DDBJ in eight portions (accession numbers D63999–D64006), was analyzed using GeneHacker. Statistics necessary to analyze one of the eight entries in the GenBank/EMBL/DDBJ was extracted from the remaining seven entries. The concatenated sequence (1,003,450 bp) is available from Cyanobase, a World Wide Web based database related to the cyanobacterium project at the Kazusa DNA Research Institute.

2.2. GeneHacker

GeneHacker is a gene-finding program developed by the first author (TY) in collaboration with the second author (YK) at the Institute of Physical and Chemical Research. The program was applied to 1.0 Mb sequence of Synechocystis sp. strain PCC6803, which was already registered in GenBank/EMBL/DDBJ in eight portions (accession numbers D63999–D64006). The GeneHacker was applied to the remains of the remaining seven entries of the GenBank/EMBL/DDBJ. The concatenated sequence (1,003,450 bp) is available from Cyanobase, a World Wide Web based database related to the cyanobacterium project at the Kazusa DNA Research Institute.
Table 2. CDSs shorter than 300 bases and results of prediction by GeneMark (GM) and GeneHacker (GH). The 8 CDSs in this list are annotated CDSs in D63899-D64006 (GenBank/EMBL/DDBJ) whose lengths are less than 300 bases. The results of the detection by GeneMark and our algorithm GeneHacker are described in a Yes/No sequence. Counterparts in the SwissProt study are also described.

| CDS     | 5'end  | 3'end | Strand | Length GH | GM | Gene |
|---------|--------|-------|--------|-----------|----|------|
| ssl0563 | 19539  | 19808 | C      | 270       | Yes| Yes  |
| ssr0062 | 20075  | 20320 | C      | 246       | Yes| No   |
| ssr1426 | 137850 | 138250| C      | 204       | Yes| No   |
| ssr0061 | 180963 | 181143| C      | 183       | No | 30S ribosomal protein S21(Sw:P02379) |
| ssr0020 | 217631 | 217924| C      | 294       | Yes| Yes  |
| ssr0330 | 478675 | 478902| C      | 228       | Yes| Yes  |
| ssr1398 | 563847 | 564044| C      | 198       | Yes| Yes  |
| ssr1399 | 564141 | 564356| C      | 216       | Yes| Yes  |

Figure 1. HMM network employed in GeneHacker. Indices a ~ d indicate the classified parameters from the genomic structural points of view. See text for a detailed description of the network.

In GeneHacker, a network representing the genomic structure of the cyanobacteria is constructed (Fig. 1). The network is a directed cycle whose components are the intergenic region, start codon, internal codon and stop codon. The order of the components in the cycle is the same as above. CDS consists of start, internal and stop codons. From the genomic structural points of view, the parameters can be classified into four groups, labeled a ~ d in Fig. 1. We determined these parameters for examination of given sequences by statistical analysis. The Viterbi algorithm was used to detect the CDSs within genomic sequences.

The intergenic region is represented by state \( S_1 \) in the network. The state outputs characters A, C, G and T (a) according to base composition in the region, and possesses transitions to itself and to start codons. The former and latter transition probabilities \( b \) and \( c \) represent the average length of the region and start codon frequencies, respectively. ATG and GTG codon (states \( S_2 \) and \( S_3 \)) are assigned as the start codons. Internal codons are represented by \( S_4 \sim S_{61} \). These 61 states correspond to the codons from AAA to TTT exclusive stop codons. The state of internal codon possesses transitions for all internal codons including itself and the stop codons. The
transition probabilities \( (d) \) represent di-codon frequencies within CDSs. TAA, TAG and TGA codons are assigned as stop codons \( (S_{65} \sim S_{67}) \).

A detailed description of GeneHacker can be found in the first author's paper.\(^{16}\) However, the present version of GeneHacker differs from the version described in the previous paper, in that it has a module which enables the detection of overlapped CDSs, which could not be detected previously.

2.3. Search for a sequence similarity of the newly found CDSs

A search for similarity in the translated amino acid sequences of the CDSs that have newly been predicted by GeneHacker as likely CDSs in the *Synechocystis* data was carried out by running the FASTA program\(^ {19}\) against the sequences in the SwissProt\(^ {20}\) and PIR\(^ {21}\) databases and by running TFASTA\(^ {19}\) against the sequences in GenBank\(^ {22}/\)EMBL. A cut-off value for an optimal score of 80 was used for these analyses.

3. Results and Discussion

3.1. Overview of CDS detection by GeneHacker

The present version of GeneHacker (ver. 0.9), which contains the module detecting overlapping CDSs, detected 854 CDSs. It detected 92.9% \((760/818)\) of the annotated CDSs,\(^ {11}\) which was slightly higher than the prediction rate by GeneMark (91.9%, 752/818).\(^ {12}\) The newly introduced module enabled GeneHacker to detect five more annotated CDSs, namely sll032, slr0770, srl0514, srl0799 and sll0780. GeneHacker detected 94 potential new CDSs.

3.2. Evaluation of the GeneHacker detection of short CDSs

Eight CDSs, whose lengths were less than 300 bases, were registered in the annotation in GenBank/EMBL/DDBJ entries (D63999–D64006).

GeneHacker successfully detected all but one CDS, 30S ribosomal protein S21 (Table 2). In the previous study, GeneMark had failed in detection of two more CDSs, PscC2, 50S ribosomal protein L20.\(^ {12}\) Although this statistical sample was small, GeneHacker detected 87.5% \((7/8)\) of CDSs successfully and its prediction rate was higher than that of GeneMark, 62.5% \((5/8)\).

3.3. Examination of newly found short CDSs

Table 3 is a list of newly detected CDSs by GeneHacker. The new CDSs were tentatively named with a ‘gh’ prefix followed by a three-digit number. The CDSs which had been detected by GeneMark, their names assigned in the second author’s study\(^ {12}\) are indicated in the table. Table 4 shows the distribution of lengths of the CDSs detected by both algorithms.

As CDSs of very short length \((<150\text{ bases})\), GeneHacker detected 36 CDSs. Since, in the first phase assignment of the CDSs, CDSs with less than 150 bases were not analyzed by a similarity search, it is valuable to have an in-depth examination of results of a homology search of these CDSs.

As CDSs of moderately short length \((150\text{ bases} \leq \text{ length} < 300\text{ bases})\), GeneHacker detected 57 CDSs. Among them, 36 CDSs had been detected by GeneMark as well. When we detect a certain biological structure, and there are multiple methods of detecting the structure, it is regarded that the one detected by multiple methods is more probable than the one detected by a single method. In the case of gene-finding, Harris et al. proved the effectiveness of combinatorial detection.\(^ {23}\)

The number of CDSs detected both by GeneHacker and GeneMark was 759, 94.9% \((720/759)\) of which were annotated CDSs. It means that the combinatorial detection also can be effective here and that the remaining 39 CDSs have a high probability of being true ones. It is valuable to closely examine the 36 moderately short CDSs by applying a similarity search against the latest database. The results of the similarity search are summarized in Table 5.

3.3.1 Very short CDSs \((\text{length} < 50\text{ bases})\)

Among the 36 CDSs, the counterparts for four very short CDSs were found in the databases. All the CDSs had similar counterparts in *Cyanella Cyanophora paradoxa* (GED:U30821). The CDSs are the photosynthesis-related genes psh1, psh1T, psh1X and petX. The query match score of CDSs to database entries were 100%, 64.3%, 65.9% and 34.0%, respectively. Independently, Kaneko found these four genes by searching annotated CDSs in GED:U30821 (the whole sequence of *Cyanella Cyanophora paradoxa*) against this 1.0 Mb sequence of *Synechocystis* (personal communication).

Only these four genes, classified as very short CDSs, were found in Kaneko's trial. The fact that the four CDSs were among the 36 very short CDSs detected by GeneHacker proved that GeneHacker has the ability to detect CDSs corresponding to genes.

3.3.2 Moderately short CDSs \((150\text{ bases} \leq \text{ length} < 300\text{ bases})\)

Among the 57 moderately short CDSs, five counterparts were found in the databases (gh313, gh517, gh603, gh203 and gh710 in Table 5). Although all five were detected by GeneMark,\(^ {12}\) only 2 of them (gh203 and gh710) were detected by Hirosawa et al.\(^ {12}\) using a similarity search.
Table 3. List of newly found CDSs by GeneHacker. CDSs newly identified by GeneHacker are listed. CDSs detected by GeneMark (GM) are marked with their names ('gm' plus a three-digit number) described in the second author's previous paper.

| Entry | CDS name | 5'end | 3'end | Length Strand | Name in GM |
|-------|----------|-------|-------|---------------|------------|
| D63999 | gb101 | 12805 | 12927 | 123 | D | gm101 |
| gb102 | 20493 | 20582 | 90 | C | gm201 |
| gb103 | 23515 | 23700 | 186 | D | gm201 |
| gb104 | 62888 | 63100 | 213 | C | gm305 |
| gb105 | 82881 | 82997 | 117 | C | gm305 |
| gb106 | 98477 | 98589 | 114 | D | gm305 |
| gb107 | 120008 | 121030 | 96 | C | gm305 |
| D64000 | gb201 | 139977 | 140153 | 177 | D | gm304 |
| gb202 | 147326 | 147420 | 96 | D | gm201 |
| gb203 | 150007 | 150207 | 201 | C | gm201 |
| gb204 | 164277 | 164438 | 162 | D | gm201 |
| gb205 | 193942 | 194049 | 108 | C | gm201 |
| gb206 | 218071 | 218178 | 108 | D | gm201 |
| gb207 | 218156 | 218263 | 108 | C | gm201 |
| gb208 | 242564 | 242755 | 192 | D | gm204 |
| gb209 | 249269 | 249490 | 222 | C | gm206 |
| gb210 | 266597 | 266806 | 210 | D | gm206 |

Entry | CDS name | 5'end | 3'end | Length Strand | Name in GM |
|-------|----------|-------|-------|---------------|------------|
| D64001 | gb301 | 276466 | 276759 | 292 | C | gm302 |
| gb302 | 277675 | 277964 | 300 | C | gm303 |
| gb303 | 285816 | 286079 | 264 | D | gm303 |
| gb304 | 286236 | 286529 | 294 | C | gm304 |
| gb305 | 296916 | 297131 | 216 | D | gm304 |
| gb306 | 309833 | 309921 | 189 | D | gm305 |
| gb307 | 310032 | 312298 | 267 | C | gm305 |
| gb308 | 329361 | 329519 | 159 | C | gm306 |
| gb309 | 337311 | 337411 | 105 | D | gm306 |
| gb310 | 342736 | 343146 | 141 | D | gm306 |
| gb311 | 346222 | 346341 | 120 | C | gm306 |
| gb312 | 357667 | 357766 | 102 | C | gm306 |
| gb313 | 363352 | 364551 | 200 | D | gm306 |

| Entry | CDS name | 5'end | 3'end | Length Strand | Name in GM |
|-------|----------|-------|-------|---------------|------------|
| D64002 | gb401 | 379042 | 379263 | 222 | D | gm401 |
| gb402 | 382947 | 383090 | 144 | D | gm401 |
| gb403 | 391417 | 391624 | 188 | D | gm401 |
| gb404 | 391391 | 391675 | 285 | C | gm401 |
| gb405 | 408948 | 409166 | 219 | C | gm401 |
| gb406 | 427914 | 428045 | 132 | D | gm401 |
| gb407 | 428016 | 428114 | 106 | D | gm401 |
| gb408 | 428866 | 428944 | 270 | C | gm402 |
| gb409 | 429290 | 429577 | 288 | D | gm403 |
| gb410 | 444160 | 444378 | 219 | C | gm404 |
| gb411 | 451533 | 451660 | 129 | C | gm404 |
| gb412 | 454985 | 455110 | 126 | C | gm405 |
| gb413 | 468163 | 468329 | 177 | C | gm405 |
| gb414 | 471453 | 471656 | 204 | C | gm406 |
| gb415 | 471659 | 471874 | 216 | C | gm406 |
| gb416 | 471453 | 471432 | 150 | D | gm407 |
| gb417 | 481087 | 481890 | 213 | D | gm407 |

This time, the results of the similarity search of the CDSs detected by both algorithms were closely examined with a lower threshold of optimal score than before. This intensive examination led to the detection of counterparts of three more CDSs, gb313, gb517 and gb603, in the databases. Gh603 corresponds to ferredoxin, a photosynthesis-related gene (Sw:P07838). Gh515 corresponds to ycf34, annotated CDSs in the whole sequence of Cyanelle Cyanophora paradoxa (GED:U30821).

3.4. Advantage of GeneHacker on short CDS prediction

In most gene-finding algorithms such as GeneMark, a fixed sliding window was used to evaluate the likelihood that a subsequence within the window is a part of the CDS. In short CDS detection, a window size smaller than the CDS length was preferable because the likelihood is normalized according to size. However, the employment of a small window caused false-positive detections because of insufficient statistics.

On the other hand, GeneHacker uses a Viterbi algo-

Table 4. Distribution of the length of newly predicted CDSs by GeneHacker (GH). Distribution of the newly found CDSs in different ranges of length. Information on the previously found CDSs using GeneMark (GM) are described as well.

| Length | GM | GM and GH | GH |
|--------|----|-----------|----|
| < 100  | 3 | 1 | 1 |
| 100 > l > 150 | 43 | 36 | 57 |
| l > 150 | 4 | 2 | 36 |
| Total | 50 | 39 | 94 |
Table 5. Summary of similarity search. Counterparts of two very short CDSs and those of five moderately short CDSs were found by similarity searches.

| CDS   | 3'end | 3'end | Length | Strand | GM   | Gene                      |
|-------|-------|-------|--------|--------|------|---------------------------|
| gh105 | 82881 | 82997 | 117    | C      | psbl(Sw:P17747)            |
|       |       |       |        |        | sp. (PCC 6301)             |
|       |       |       |        |        | (GED:U30821)               |
|       |       |       |        |        | Cyanophora paradoxa        |
| gh202 | 147325| 147420| 96     | D      | psbT(Sw:P20176)            |
|       |       |       |        |        | Cyanelle                   |
| gh311 | 346222| 346341| 120    | C      | psbX(GED:U30821)           |
|       |       |       |        |        | Cyanelle                   |
| gh714 | 850932| 851042| 111    | D      | petX(GED:U30821)           |
|       |       |       |        |        | Cyanelle                   |
| gh313 | 364352| 364591| 240    | C      | peml-like protein          |
|       |       |       |        |        | Escherichia coli           |
|       |       |       |        |        | ycf34(GED:U30821)          |
|       |       |       |        |        | Cyanelle                   |
| gh517 | 590248| 590496| 249    | D      | ycf33(GED:U30821)          |
|       |       |       |        |        | Cyanelle                   |
| gh603 | 689899| 690105| 207    | D      | blt101(GEDZ25537)          |
|       |       |       |        |        | Cyanelle                   |
| gh710 | 817161| 817325| 165    | D      |                     |
|       |       |       |        |        |                     |
|       |       |       |        |        |                     |
|       |       |       |        |        |                     |

rithm, a kind of Dynamic Programming (DP), in its evaluation of possible CDSs, with no sliding window. In DP, the evaluation was performed for every base. It is essentially the same as selecting proper window sizes and window positions according to CDSs, which are fixed in the case of GeneMark. Therefore, it is expected that GeneHacker possesses adequate prediction ability irrespective of CDS length.

4. Summary

We used GeneHacker to detect CDSs in a 1.0-Mb sequence of *Synechocystis* sp. strain PCC6803. GeneHacker detected 36 very short CDSs. We found that four CDSs among them had counterparts in the databases and corresponded to photosynthesis-related genes. These results show that GeneHacker is effective in detecting very short CDSs, which were discovered among photosynthesis-related ones at relatively higher rate. The prediction rate of GeneHacker was 1.0% higher than that of GeneMark and the CDSs corresponding to the 1.0% are presumed to be the short CDSs detected by GeneHacker but not by GeneMark.

Its ability to restrict the number of very short CDSs to around 50 CDSs with a high possibility of being genes is very useful because it is possible to extract 2,500 very short possible CDSs (length of 90 bp to 150 bp), mentioned in Section 1.2, in the 1.0-Mb nucleotide sequence (Table 1) and because analysis of results of their similarity search is not easy. Presently, the percentage of very short genes in the databases is low except in well-studied genes, such as photosynthesis-related genes. However, the percentage will be increased as studies on other kinds of genes progress. Therefore, it can be assumed that more very short CDSs that can be detected by GeneHacker will have similar counterparts in the databases.

We closely examined the results of similarity searches for moderately short CDSs detected by both GeneMark and GeneHacker. This was done by lowering the cutoff value for similarity searches. As a result, we found three more CDSs corresponding to genes than before. It showed that CDSs detected by multiple gene-finding algorithms are more likely true genes. When multiple gene-finding algorithms are available, we suggest, it will be better to use these algorithms combinatorially.

Acknowledgments: We acknowledge Takakazu Kaneko of Kazusa DNA Research Institute for valuable discussion. This work was carried out as a part of the ALIS (Advanced Lifescience Information System) project for genome analysis by the Japan Science and Technology Corporation (JST), and was supported in part by a grant from the Kazusa DNA Research Institute (KDRI).

References

1. Robinson, K., 1996, The 100 Kb Club, http://golgi.harvard.edu/100kb/.
2 Fickett, J. W. and Tung, C.-S. 1992, Assessment of Pro-
tein Coding Measures, *Nucleic Acids Res.*, 20, 6441–6450.

3 Borodovsky, M. and McIninch, J. D. 1993, GENMARK: Parallel gene recognition for both DNA strands, *Comput. Chem.*, 17, 123–133.

4 Uberbacher, E. C. and Mural, R. J. 1991, Locating protein-coding region on human DNA sequences by a multiple sensor-neural network approach, *Proc. Natl. Acad. Sci. U.S.A.*, 88, 11261–11265.

5 Guigo, R., Kundsen, S., Drake, N., and Smith, T. 1992, Prediction of gene structure. *J. Mol. Biol.*, 226, 141–157.

6 Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. 1994, Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucleic Acids Res.*, 22, 5156–5163.

7 Dong, S. and Searls, D. B. 1994, Gene Structure Prediction by Linguistic Methods, *Genomics*, 23, 540–551.

8 Krogh, A., Mian, I. S., and Haussler, D. A. 1994, Hidden Markov Model that Finds Gene in E. coli DNA, *Nucleic Acids Res.*, 22, 4768–4778.

9 Snyder, E. E. and Stormo, G. D. 1995, Identification of Protein Coding Regions in Genomic DNA, *J. Mol. Biol.*, 248, 1–18.

10 Kotani, H., Kaneko, T., Matsubayashi, T., Sato, S., Sugiiro, M., and Tabata, S. 1994, A Physical Map of the Genome of a Unicellular Cyanobacterium *Synechocystis* sp. strain PCC 6803, *DNA Res.*, 1, 303–307.

11 Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiiro, M., and Tabata, S. 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.

12 Hirosawa, M., Kaneko, T., Tabata, S., McIninch, J. D., Hayes, W. S., Borodovsky, M., and Isono, K. 1995, Computer Survey for Likely Genes in the One Megabase Contiguous Genomic Sequence Data of *Synechocystis* sp. Strain PCC6803, *DNA Res.*, 2, 239–246.

13 Borodovsky, M., Rudd, K. E., and Koonin, E. V. 1994, Intrinsic and extrinsic approaches for detecting genes in a bacterial genome, *Nucleic Acids Res.*, 22, 4756–4767.

14 Baum, L. E., Petrie, T., Soules, G., and Weiss, M. 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.*, 41, 164–171.

15 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.

16 Yada, T. and Hirosawa, M. 1996, Gene Recognition in Cyanobacterium Genomic Sequence Data Using the Hidden Markov Model, *Proc. of the 4-th Int. Conf. on Intelligent Systems for Molecular Biology*, 252–260.

17 Hirosawa, M., Kaneko, T., and Tabata, S. 1995, CyanoBase: Visual representation of information on the genome of Cyanobacterium *Synechocystis* sp. strain PCC6803 through WWW, *Proc. of Genome Informatics Workshop VI*, 102–103.

18 Hirosawa, M. 1996, CyanoBase, http://www.kazusa.or.jp/cyano/cyano.html.

19 Pearson, W. R. 1990, Rapid and Sensitive Sequence Comparison with FASTP and FASTA, *Methods Enzymol.*, 183, 63–98.

20 Bairoch, A. and Apweiler, R. 1996, The SWISS-PROT protein sequence data bank and its new supplement TREMBL, *Nucleic Acids Res.*, 24, 21–25.

21 George, D. G., Barker, W. C., Mewes, H.-W., Pfeiffer, F., and Tsugita, A. 1996, The PIR-International Protein Sequence Database, *Nucleic Acids Res.*, 24, 17–20.

22 Benson, D. A., Boguski, M., Lipman, D. J., and Ostell, J. 1996, GenBank, *Nucleic Acids Res.*, 24, 1–5.

23 Harris, N. L., Reese, M. G., Helt, G. A. and Eeckman, F. H. 1996, GeneNomi, http://www-hgc.lbl.gov/homes/nomi/.
