J. Biomedical Science and Engineering, 2010, 3, 868-883 **JBiSE**
doi:10.4236/jbise.2010.39117 Published Online September 2010 (http://www.SciRP.org/journal/jbise/).

Identification of the interactive region by the homology of the sequence spectrum

Masatoshi Nakahara[1], Masaharu Takeda[2]*

[1]Department of Computer and Information Sciences, Sojo University, Ikeda, Kumamoto, Japan;
[2]Department of Materials and Biological Engineering,Tsuruoka National College of Technology, Tsuruoka, Yamagata, Japan.
Email: mtakeda@tsuruoka-nct.ac.jp

Received 4 June 2010; revised 9 July 2010; accepted 12 July 2010

ABSTRACT

The base sequence in genome was governed by some fundamental principles such as reverse-complement symmetry, multiple fractality and so on, and the analytical method of the genome structure, the "Sequence Spectrum Method (SSM)", based on the structural features of genomic DNA faithfully visualized these principles. This paper reported that the sequence spectrum in SSM closely reflected the biological phenomena of protein and DNA, and SSM could identify the interactive region of protein-protein and DNA-protein uniformly. In order to investigate the effectiveness of SSM we analyzed the several protein-protein and DNA-protein interaction published primarily in the genome of *Saccharomyces cerevisiae*. The method proposed here was based on the homology of sequence spectrum, and it advantageously and surprisingly used only base sequence of genome and did not require any other information, even information about the amino-acid sequence of protein. Eventually it was concluded that the fundamental principles in genome governed not only the static base sequence but also the dynamic function of protein and DNA.

Keywords: Spectrum of Genome Base Sequence; Homoology of Sequence Spectrum; Interactive Region; Reversese-Complement Symmetry; Multiple Fractality; Analytical Method Of Genome

1. INTRODUCTION

As described in the previously [1,2], it was very important to investigate the structure of the entire genome because the four bases should be arranged in a sophisticated fashion in the genome, and essentially the base sequences might reflect the conformations of protein, RNA and DNA. DNA sequences were deeply affected by the adjoining sequences. In other words, the non-coding sequences might play some important roles to express each gene (the coding sequences) in genome. That is, not only the coding region, but also the non-coding region might be necessary to transmit and to transform the biological information precisely, rapidly, and stably. Therefore, if we would find meaningful structure in the genome, we might also obtain important information about the functions of protein, RNA and DNA from their structure.

Previously, we showed that the four bases in genomic DNA were organized based on the generation-rules in all organisms by analyzing the appearance frequency of the bases, and we proposed three generation-rules of the base sequences in a single-strand of DNA: 1) reverse-complement symmetry of the 1 ~ 9 successive base sequences, 2) multiple fractality of each base distribution depending on the distance, and 3) bias of four bases, A, T, G and C. These rules were universally observed regardless species [1]. Further we also defined the sequence spectrum by the appearance frequency of the base sequence in genome, and we have developed the powerful method "Sequence Spectrum Method (SSM)" in order to visualize and analyze the generation-rules in entire genome explained above. As one of important results, we revealed by using SSM that there was the remarkable homology of sequence spectrum between proteins and tRNAs [2]. This fact suggested the sequence spectrum could be closely associated with the function of protein, and the homology of sequence spectrum could be related to the mutual interactive region. Identification of mutual interactive region of protein, RNA and DNA was definitely important to figure out their functions, and usually the homology of base sequence or amino-acid sequence was used for it.

To investigate the effectiveness of SSM, in this paper, we showed that SSM could identify the interactive region of the protein-protein and the protein-DNA by the homology of the sequence spectra. The advantages of the proposed method were as follows.

1) It used only base sequence of genome and did not

require any other information, even information about amino-acid sequence of protein. As SSM faithfully reflected the biological information, the conservation of the bases sequences of genomic DNA was also conserved in the translated amino acids sequence of the protein sequence [1,2].

2) It could identify the interactive region of both protein-protein and protein-DNA in completely the same manner.

3) It could be executed fully on a personal computer and did not require a special high performance computer. Moreover the identification was done in a few seconds.

2. MATERIALS AND METHODS

2.1. Sequence Spectrum Method (SSM)

SSM was carried out in the same way as the published procedures [2]. The outline of the proposed method was as follows. The base sequence of interest was sectioned by a small number of bases from the top (5'-end). The key sequences of the nine successive base sequences (d = 9) was 262,144 sequences (= 4^9, Reference [2]). The appearance frequency of the key sequence was counted in the entire genome, and was plotted at the position of the first base of the key sequence as described in the next paragraph. These procedures were carried out for the entire base sequence of interest with one base shift (p = 1). The next step was to average the appearance frequencies so that a recognizable pattern of appearance frequency was obtained for the base sequence. This pattern of the averaged appearance frequency was called the "sequence spectrum". Finally, the homology factor between two sequence spectra was calculated to determine the degree of homology. The exact procedure was explained below in a mathematical way.

Let S be an entire set of base sequences, and $B = [b_i]$ be a partial set of interest in S. A base element was denoted by b_i (i = 1..M), and M was the base sequence size of B. The base element b_i become A (adenine), T (thymine), G (guanine) or C (cytosine). The key sequence k_i and the appearance frequency f_i were defined for b_i as follows.

Key sequence k_i : base sequence comprised of sequential base elements b_i~b_{i+d-1} (d : base size of the key sequence).

Appearance frequency f_i : appearance count of k_i in S. The key sequence k_i was compared with the base sequence of the entire set S, and the appearance frequency f_i was increased by one every time the key sequence k_i matches the partial base sequence of the entire set S. This procedure was iterated for all key sequences k_i to obtain f_i (i = 1..M). In practice all f_i were counted and tabulated in advance by scanning all base sequence in S. Consequently, the appearance frequency vector $F = [f_i]$ (i = 1..M) was determined (actually, the appearance frequencies for the last (d-1) base elements of B could not be calculated; however, this was neglected because $M >> d$-1).

Next, the appearance frequency f_i was averaged as follows:

$$f_{si} = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} f_j$$

where the parameter m was average width. This averaged appearance frequency $Fs = [f_{si}]$ (i = 1..M) was called the "sequence spectrum".

The next step was to calculate the homology factor to determine the degree of homology. The homology factor determines the homologous region of a target base sequence with respect to a reference base sequence. In order to derive the homology factor, the mutual correlation function MF within the window width of homology was calculated as

$$MF_{ij}(Fsr, Fst) = \frac{1}{\|Fsr_i\|\|Fst_j\|} \sum_{k=1}^{w} (fsr_{i+k} - \overline{fsr_i}) * (fst_{j+k} - \overline{fst_j})$$

$$\|Fsr_i\| = \sqrt{\sum_{k=1}^{w} (fsr_{i+k} - \overline{fsr_i}) * (fsr_{i+k} - \overline{fsr_i})}$$

$$\|Fst_j\| = \sqrt{\sum_{k=1}^{w} (fst_{j+k} - \overline{fst_j}) * (fst_{j+k} - \overline{fst_j})}$$

$$\overline{fsr_i} = \sqrt{\frac{1}{w}\sum_{k=1}^{w} fsr_{i+k}}$$

$$\overline{fst_j} = \sqrt{\frac{1}{w}\sum_{k=1}^{w} fst_{j+k}}$$

where

Fsr— sequence spectrum of the reference base sequence

Fst— sequence spectrum of the target base sequence

w— window width of homology

The mutual correlation function MF ranges from -1 to 1, and then the homology factor HF was defined as

$$HF_{ij}(Fsr, Fst) = \frac{(MF_{ij}+1)}{2} * 100[\%]$$

The higher the homology factor, the more similar the sequence spectra were. The similar regions of the target base sequence with respect to the reference base sequence were obtained by calculating the homology factors HF_{ij} for all i (i = 0..*Mr*-*w*, *Mr*: size of reference sequence) and j (j = 0..*Mt*-*w*, *Mt*: size of target sequence).

When the base sequence was very large, elements of the sequence spectrum were skipped by the size factor p

to reduce the size as follows.

$$fs_i \rightarrow fs_{(i-1)*p+1}$$

For instance, when p = 2

$$fs_1, fs_2, fs_3 \ldots \rightarrow fs_1, fs_3, fs_5 \ldots$$

This operation reduced the size to 1/p.

The base sequences of the genomes were obtained from the databases listed below.

Saccharomyce Genome Database. (2010) (http://www. yeastgenome.org/).

NCBI genome data base. (2010) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome).

2.2. Appearance Frequencies of Bases

For nine successive bases, the appearance frequency was counted for the entire genome by matching from the start of the base sequence in a genome with one base shift (p = 1) as follows.

Ex. Nine successive bases: **AATAAAGAA**

AATAAAGAA (one base shift)

Base Sequence:

5’-ATCG**AATAAAGAA**CCGTTCGGTAAGTCG**AATAAAGAA**T-CTGGCATTT-3’

1 2

Count of **AATAAAGAA**: 2

In the case of the genome composed of the plural chromosomes such as *S. cerevisiae*, we have calculated the sum of the base frequencies of the 16 chromosomes (in numeric order) plus mtDNA [1].

2.3. The Parameters “d”-, “m”- , “p”-, and “w”-Values of SSM Analysis for the Interaction

Controllable parameters in the sequence spectrum were the base size “d” of the key sequence, the average width “m”, the skip base number (the size factor) “p” and the window width “w” of homology. The parameter “d” determined the highest resolution for extracting the structural feature of the base sequence. Therefore this parameter should be chosen to be as a large value as possible to extract the exact feature. The large “m” values were usually used to obtain the overall features of the structure, and smaller “m” values were applied to investigate the structure in detail. The value of “m” normally ranges from 1/10 to 1/100 of the base sequence size [2]. This parameter was adjusted to the base sequence size especially when the homology factor between a small reference and a large target was calculated [2]. The window width of homology, “w” determined the width of similar region to identify. In this paper the values of “d”, “m”, “p” and “w” were 9, 10, 1 and 200, respectively, to identify the interactive region of protein and DNA.

In figures of the sequence spectrum the horizontal parameter was the base size of sequence, M of each gene or genomic DNA, and the vertical parameter was the sequence spectrum. These parameters were appropriately scaled to show the similar region clearly.

2.4. Procedure of Identification of the Interactive Region by SSM

To simplify the procedure, it was assumed that the interactive region of one protein was given (shown in purple-blue), and SSM identified the interactive region of the other protein (shown in red). The procedure to identify the interactive regions of two proteins by SSM was as follows. In the following procedure one of two proteins was replaced by DNA when the protein-DNA interaction was investigated.

[Step 1] One protein with the given interactive region (shown in purple-blue) was designated as a reference protein, and the other protein with the interactive region (shown in red) which SSM identified was designated as a target protein.

[Step 2] The sequence spectra of both the reference and target proteins were calculated.

[Step 3] The similar regions between the sequence spectra of the reference and target proteins were calculated.

[Step 4] The pair of similar regions (red/purple-blue) with the highest homology factor (HF) was selected as a candidate of interactive regions.

[Step 5] The base sequence of the reference protein was converted to be the reverse complementary and the steps [2-4] were repeated because of the reverse-complement rule in genome.

[Step 6] In two candidates obtained in steps [4] and [5], the similar region of the target protein with higher HF was called first identified region, and the other was called second identified region.

3. RESULTS AND DISCUSSION

This section demonstrates that the homology of the sequence spectrum was closely associated with the mutual interaction of proteins or DNA. The identified interactive regions of the proteins were all the first identified regions in the examples below. We showed some of the interactive regions analyzed by SSM in this section.

3.1. Mutual Interaction of Protein-Protein

1) *MAS*1 and *MAS*2

Figure 1 showed the interactive region (in purple-blue) of *MAS*1 [Mas1p (*β*-MPP), Reference [3]] - *MAS*2 [Mas-2p (*α*-MPP), Reference [4]]. These proteins formed a complex to cleave the mitochondrial targeting signal of precursors. In **Figure 1(a)** the active region (in purple-blue) around the key amino acid E^{73} of *MAS*1 (Mas1p) was the reference, and the whole coding region of *MAS*2 (Mas2p) was the target (**Figure 1(b)**). Previous reports proposed a model in which the glycine-rich region of *MAS*2 (Mas2p, in red) cooperated with the active region of *MAS*1 (Ma- s1p, in purple-blue). Our results strongly supported this model because the most similar region of *MAS*2 (in red; HF = 90.5%) with the active region of *MAS*1 (in purple-blue) was completely identical to the reported glycine-rich region [5,6, in red]. Moreover, the positions of the key amino acids in both proteins (E^{73} in Mas1p and K^{296} in Mas2p) were also identical.

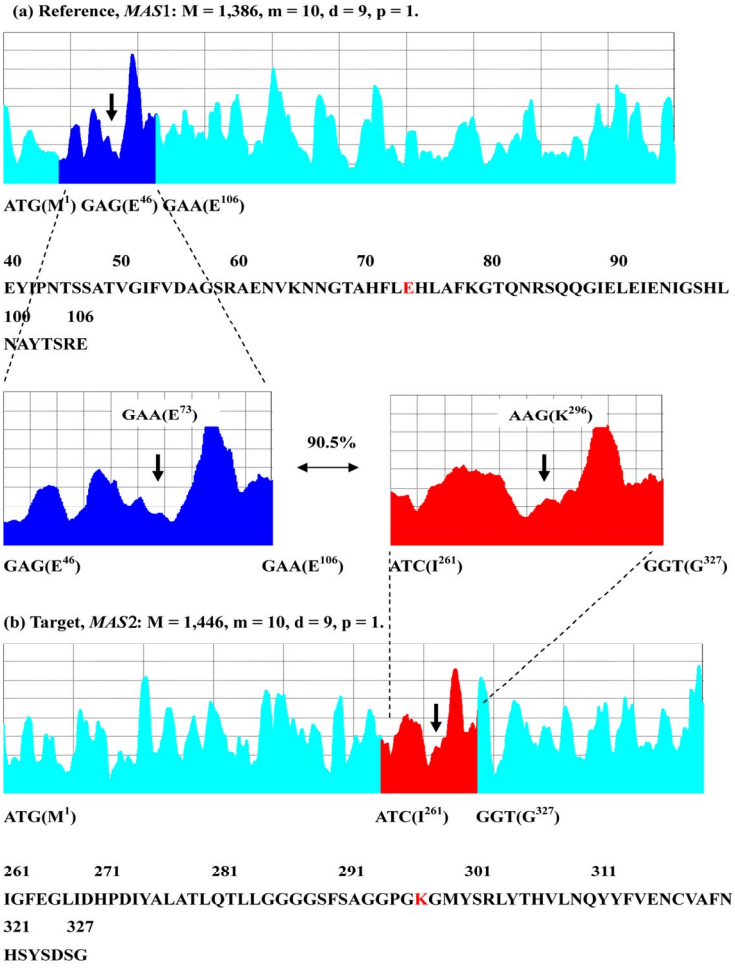

Figure 1. Sequence spectra of *MAS*1 and *MAS*2 (d = 9, m = 10, p = 1). (a) Coding region of *MAS*1 (Mas1p, M = 1,386). The active region of *MAS*1 (Mas1p, reference: M = 200, in purple-blue). This region (corresponding to E^{46} – E^{106}) carries the characteristic metal-binding motif associated with the catalytic activity (5, 6). (b) Coding region of *MAS*2 (Mas2p) containing the 5'- and 3'- non-coding region (target: M = 1,446). The region most similar to the reference is shown in red (HF = 90.5%). The most similar region is glycine-rich and closely related to the catalytic function (I^{261} – G^{327} of Mas2p). E^{73} (shown in red letter) of Mas1p presumably interacts with K^{296} (shown in red letter) of Mas2p (position of arrowhead). The scales of the axes for the sequence spectra of the similar regions were the same. The amino acid sequences of Mas1p and Mas2p neighboring the interactive regions were shown in figures, respectively.

2) *PHO*4 and *PHO*80

Figure 2 showed the sequence spectra of *PHO*4 (a, Pho4p, reference: the interactive region around the key amino acid P^{174}, in purple-blue) and *PHO*80 (b, Pho80p, target: the whole coding region). *PHO4* (Pho4p) was a transcription factor, and *PHO*80 (Pho80p) inhibited the transcriptional function of *PHO*4 (Pho4p). Ogawa & Oshima [7] and Okada & Toh-e [8] reported that there was interaction between P^{174} in Pho4p and M^{42} in Pho80p, respectively. The red region in (b) in which M^{42} (**Figure 2(b)**, arrow head) of Pho80p was located was the region most similar to the reference region of Pho4p, in which P^{174} (**Figure 2(a)**, arrow head) was located (HF = 89.1%). The interactive regions between Pho4p and Pho80p were also discussed in the Pho2p results (6) later.

3) *RPB*2 and *RPB*12

Figure 3 showed the sequence spectra of *RPB*2 and *RPB*12. The *RPB* protein family forms DNA-directed RNA polymerase II [9]. *RPB2* (Rpb2p encoding gene) and *RPB*12 (Rpb12p) were members of the family, and *RBP*12 (Rpb12p) combined with *RPB*2 (Rpb2p). Rpb12p was a very small protein with 70 amino acids whereas

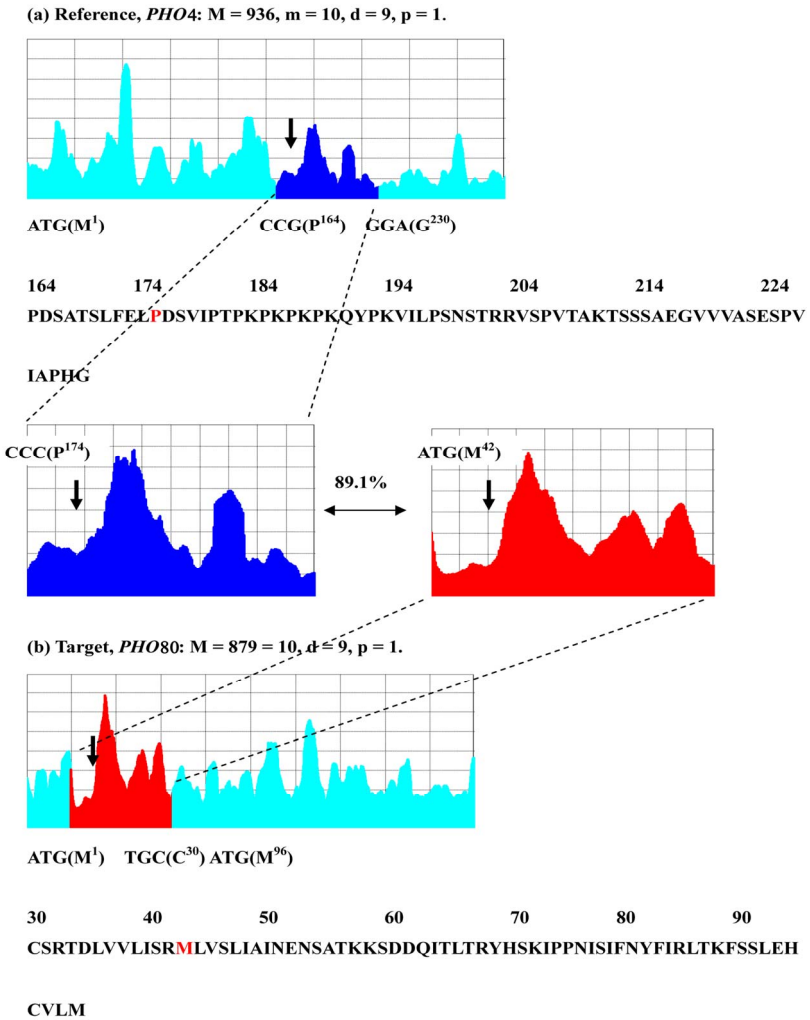

Figure 2. Sequence spectra of *PHO*4 and *PHO*80 (d = 9, m = 10, p = 1). (a) Coding region of *PHO*4 (Pho4p, M = 936, the active region was shown in purple-blue). (b) Coding region of *PHO*80 (Pho80p, target: M = 880). The region most similar to the reference is shown in red (HF = 89.1%). It has been shown that P^{174} (shown in red letter) of Pho4p interacts with M^{42} (shown in red letter) of Pho80p [7, 8]. The arrowhead in each spectrum respectively indicates the position of the amino acid P^{174} of Pho4p, and M^{42} of Pho80p. The scales of axes in (a) and (b) are the same. The amino acid sequences of Pho4p and Pho80p neighboring the interactive regions were shown in figures, respectively. The red letter indicated to report as a functional amino acid.

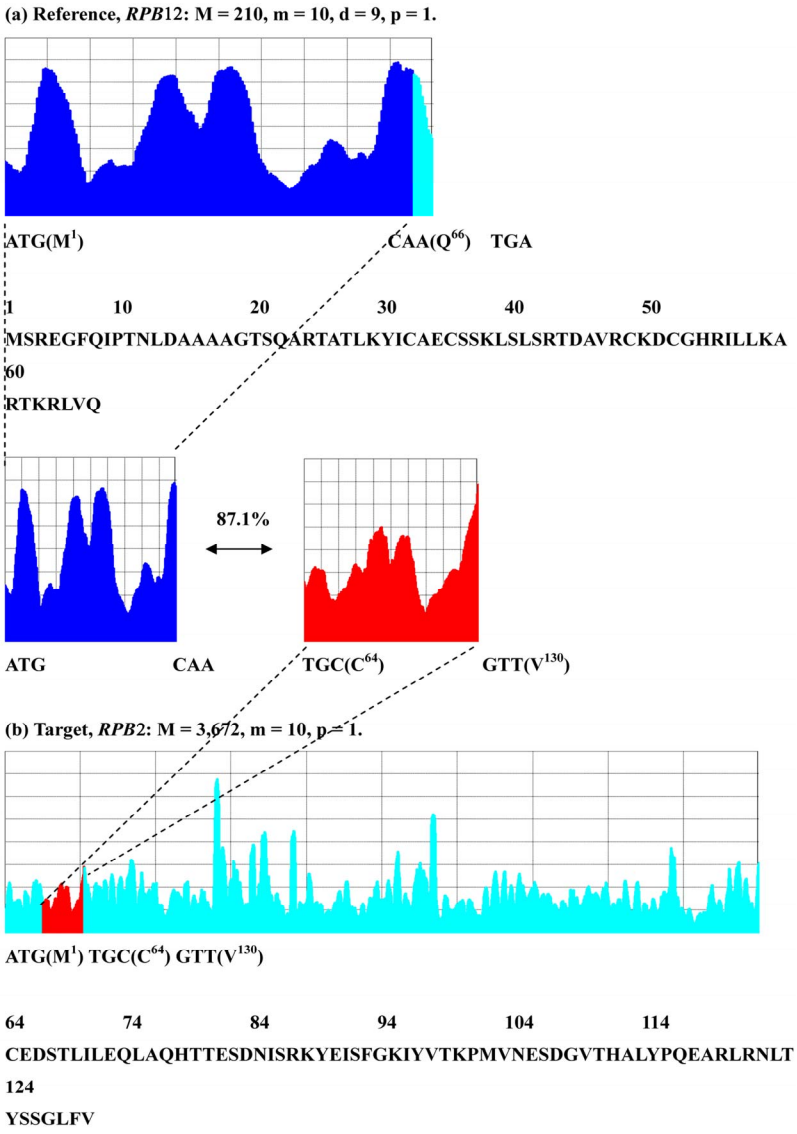

Figure 3. Sequence spectra of *RPB*12 and *RPB*2 (d = 9, m = 10，p = 1). (a) Coding region of *RPB*12 (Rpb12p, reference: M = 210). (b) Coding region of *RPB*2 gene containing the 5'- and the 3'- non-coding region (Rpb2p, target: M = 3,672). The region most similar to the reference is shown in red (HF = 87.1%). The scales of axes in (a) and (b) are the same. The amino acid sequences of Rpb12p and Rp- b2p neighboring the interactive regions were shown in figures, respectively.

Rpb2p was a large one with 1224 amino acids. Therefore in this case the whole coding region of *RPB*12 (Rpb12p) was suitable for the reference (a) and the coding region of *RPB*2 (Rpb2p) for the target (b). The result was shown in **Figures 3(a-b)**. The red region is the most similar re- gion of *RPB*2 (Rpb2p) with *RPB*12 (Rpb12p, HF = 87.1%). The literature [9] revealed that the interaction between *RPB*2 (in red) and *RPB*12 (in purple-blue) occurred at two regions of *RBP*2, and **Figure 3** showed one of these two interaction regions. This result was unlikely to be a coincidence because the target size was about 18 times larger than the reference size. In addition, interestingly the other interacting region was very close to the second identified region in the coding region (not shown), although it was not completely identical (a previous report [9] specified the region around the 900th amino acid of Rpb2p, but our results specified the region around the 940th amino acid).

4) *GCR*1 and *GCR*2

The interactive region of *GCR*1 [Gcr1p,10] and *GCR*2 [Gcr2p,11] was very interesting. In **Figure 4** the red region of *GCR*1 (Gcr1p, leucine zipper) was the first iden-

tified region (HF = 92.9%) with respect to the reference region (in purple-blue) of *GCR*2 (Gcr2p). The sequence spectra suggested that the leucine-zipper region of *GCR*1 (Gcr1p) might interact with the C-terminus of *GCR*2 (Gcr2p, purple-blue region), although considerable controversy still existed concerning the interaction between Gcr1p and Gcr2p [12,13]. This case is quite interesting for following reasons: a) the identified region was derived from the reverse-complement reference region of *GCR*2, that is, the reverse-complement base sequence of *GCR*2 was also useful to the analysis of the interactive region by SSM (designated it as the reverse-complement rule), and b) the portion of the reference region exceeded outside to the downstream region. This means that in this case the proposed method identified both the different objects, the protein region for *GCR*2 (Gcr2p) and the DNA region for *GCR*2 of the reference region. That is, the sequence spectrum of a given gene might reflect the information of both protein and DNA, and SSM could be applied to analyze both of them.

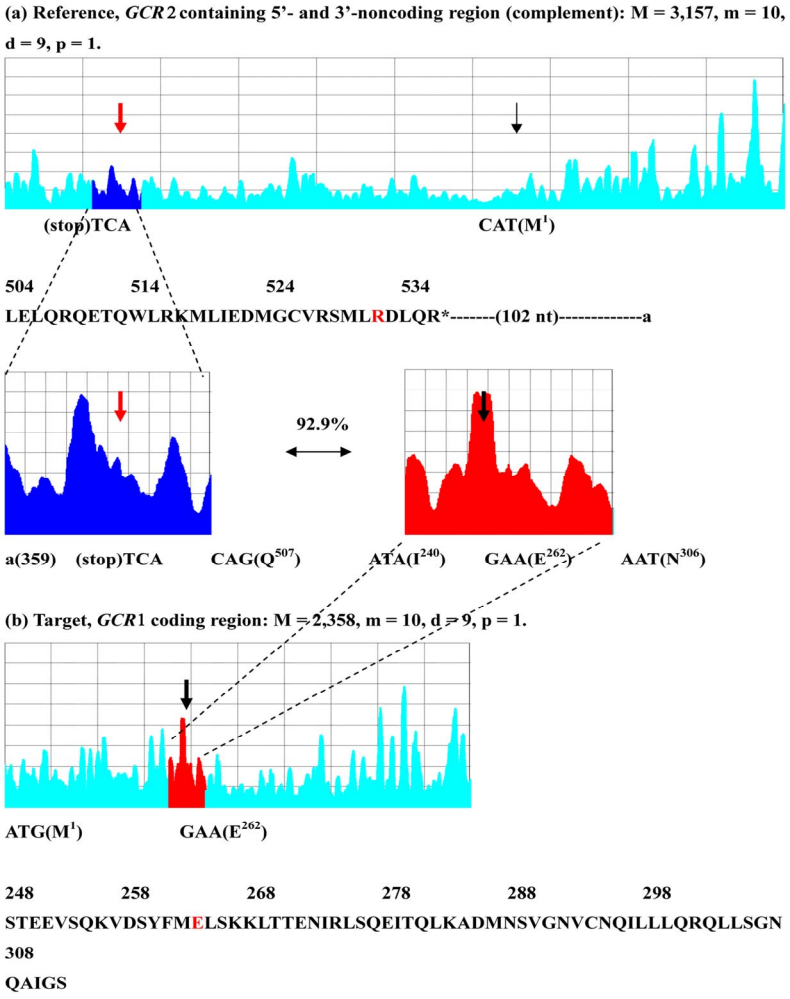

Figure 4. Sequence spectra of *GCR*1 and *GCR*2 (d = 9, m = 10, p = 1). (a) The reverse-complement sequence of whole region of *GCR*2 (Gcr2p) containing the 5'- and the 3'- non-coding region was used as the reference (M = 3,157, the active region was shown in purple-blue). (b) The functional region (K^{266} – R^{300}, leucine zipper) of *GCR*1 (Gcr1p, ref.10-13). The region most similar to the reference (HF = 92.9%). This region (leucine zipper, ref. 12, 13) of Gcr1p might interact with the reference region of Gcr2p. The scales of axes in (a) and (b) are the same. The arrowhead of black and red were the start codon (M^1) and the stop codon (TGA) of *GCR*2, respectively. The bold black arrowhead of *GCR*1 was the position of E^{262} (red letter in the amino acid sequence of Gcrp1). The amino acid sequences of Gcr2p and Gcr12p neighboring the interactive regions were shown in figures, respectively. The red letter indicated to report as a functional amino acid.

5) *SLA*1 and *SLA*2

This example proved that SSM could apply to large size proteins. The size of proteins Sla1p (coded by *SLA*1) and Sla2p (coded by *SLA*2) were 1244 and 968 amino acids respectively, and **Figure 5** showed the interactive regions of these proteins. In **Figure 5** the red region of *SLA*1 (Sla1p) was the first identified region (HF = 94.3%) with respect to the reference region (in purple-blue) of *SLA*2 (Sla2p) which was converted to be reverse complementary. The literature [14] showed that this result was valid.

The three examples 6) ~ 8) below were results of predicting the interactive regions by SSM. In these examples one of the interactive regions was known and the other was unknown, and SSM predicted the unknown interactive region.

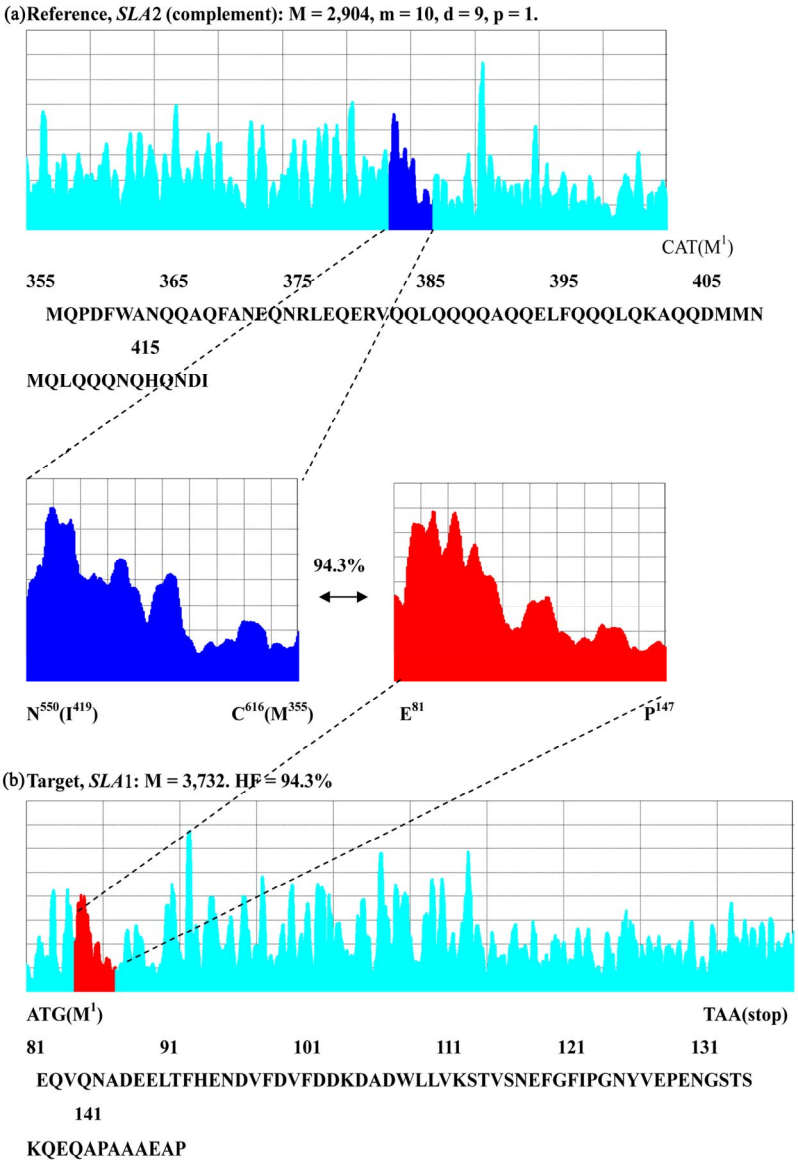

Figure 5. Sequence spectra of *SLA*2 and *SLA*1 (d = 9, m = 10, p = 1). The reverse-complement of the base sequence gave more homologous than the normal base sequence could be shown in the interaction *SLA*2 (Sla2p)/*SLA1* (Sla1p). (a) The reverse-complement sequence of coding region of *SLA*2 (Sla2p) was used as the reference (M = 2,904, the active region was shown in purple-blue). (b) The sequence spectrum region of *SLA*1 (M = 3,732. Sla1p, ref.14). The amino acid sequences of Sla2p and Sla1p neighboring the interactive regions were shown in figures, respectively. The region most similar to the reference (HF = 94.3%).

6) *PHO*2, *PHO*4 and *PHO*80 [15-17]

The identification of the interactive regions might be applied the characterization of the molecular mechanism of the metabolism. For instance, the example focusing on the interactive regions of *PHO*2 (Pho2p) - *PHO*80 (Pho80p) - *PHO*4 (Pho4p) was very suggestive. *PHO*2 was a gene coding a transcription factor, Pho2p regulating several genes like *PHO*5 with co-regulated with other transcription factor, Pho4p [15-17]. It was well known that Pho2p had a cooperative interaction with Pho4p, and the literature [15] reported that the amino acids around S^{230} of Pho2p played an important role concerning the interaction with Pho4p. In this connection SSM predicted the target interactive region of Pho4p with the reference region around S^{230} of Pho2p. The predicted region of Pho4p was located very close to or overlapped partially with the interactive region with Pho80p, and the positions of the key amino acids, S^{230} of Pho2p and P^{174} of Pho4p were identical (**Figure 6**).

As described in the above section (2) *PHO*4 and *PHO*80, P^{174} of Pho4p and M^{42} of Pho80p were functioned in the interaction of theses proteins (**Figure 2**). Namely the positions of the three key amino acids P^{174} of Pho4p, M^{42} of Pho80p, and S^{230} of Pho2p were identical

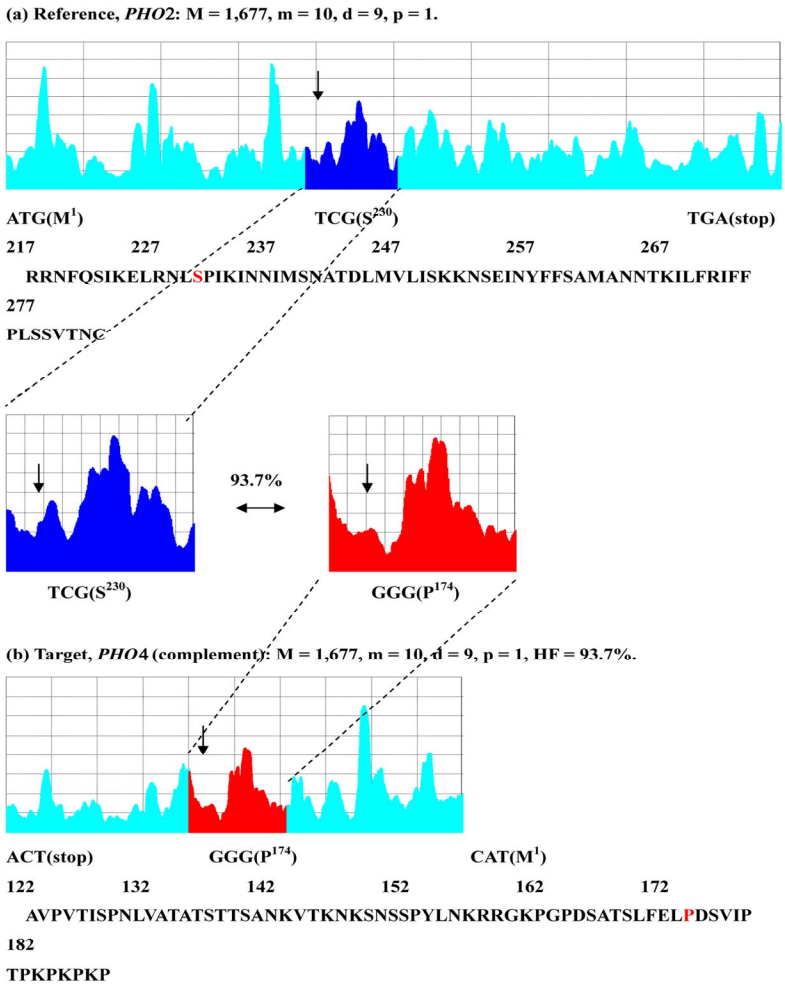

Figure 6. Sequence spectra of *PHO*2 and *PHO*4 genes (d = 9, m = 10， p = 1). (a) Coding region of *PHO*2 (Pho2p, reference: M = 1677). The region most similar to the reference is shown in purple-blue. (b) The reverse-complement sequence of coding region of *PHO*4 (Pho4p, M = 936). The active region was shown in red (HF = 93.7%). It has been shown that P^{174} (shown in red letter) of Pho4p interacts with S^{230} (shown in red letter) of Pho2p [15-17]. The arrowhead in each spectrum respectively indicates the position of the amino acid S^{230} of Pho2p, and P^{174} of Pho4p. The scales of axes in (a) and (b) are the same. The amino acid sequences of Pho2p and Pho4p neighboring the interactive regions were shown in figures, respectively. The red letter indicated to report as a functional amino acid.

in the identified interactive regions by SSM. This fact suggested that Pho80p might be interfered in the cooperation between Pho4p and Pho2p, and this result was very reasonable [15-17] although more experimental confirmations would be necessary.

7) *PHO*2 and *SWI*5 [18]

*SWI*5 was a gene encoding a transcription factor, Swi5p that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase such as *PHO*2 encoding a regulatory protein involved in cooperatively phosphate metabolism, Pho2p. The base number of the interactive region in *SWI*5 is known and unknown in *PHO*2 [18]. We predicted the unknown interactive region of Pho2p by the SSM (**Figure 7**).

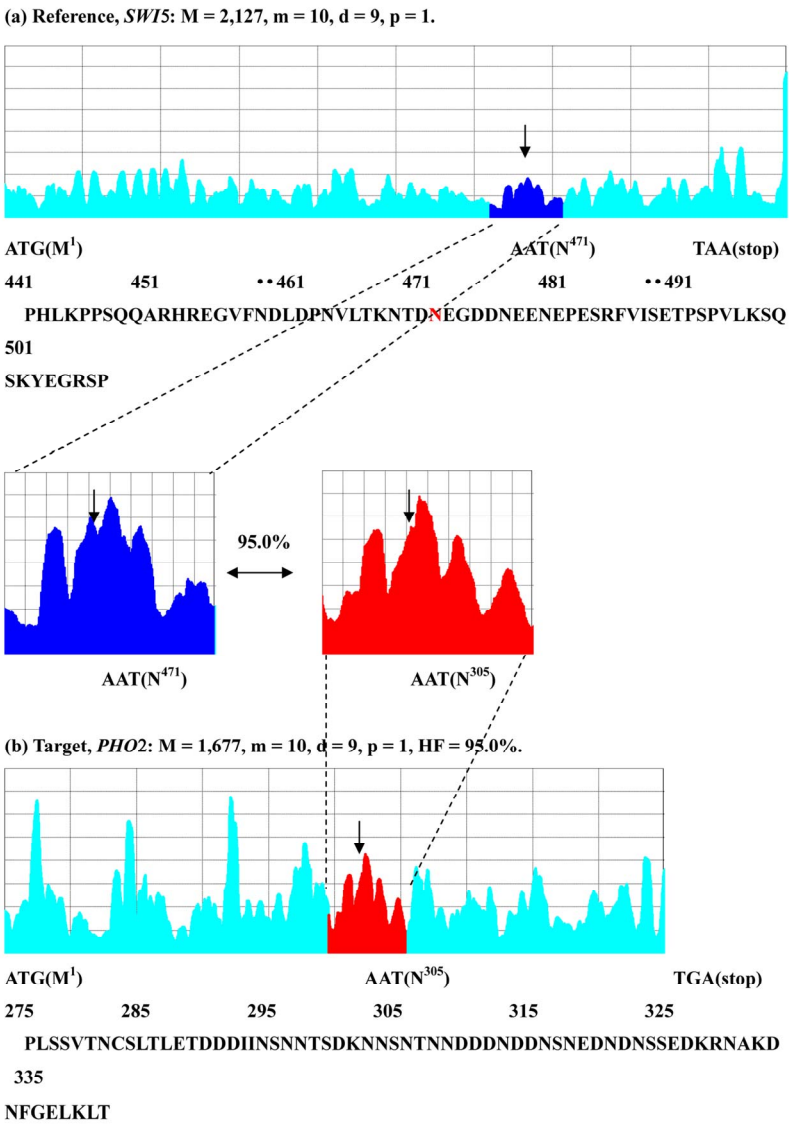

Figure 7. Sequence spectra of *SWI*5 and *PHO*2 genes (d = 9, m = 10, p = 1). (a) Coding region of *SWI*5 (Swi5p, M = 2127, the active region was shown in purple-blue). (b) Coding region of *PHO*2 (Pho2p, target: M = 1677). The region most similar to the reference is shown in red (HF = 95.0%). It has been shown that the amino acids sequences (shown in red letter) of Swi5p interacts with the amino acids sequences (shown in red letter) of Pho2p [18]. The arrowhead in each spectrum respectively indicates the position of the functional amino acid N^{471} of Swi5p, and $N3^{305}$ of Pho2p. The scales of axes in (a) and (b) are the same. The amino acid sequences of Swi5p and Pho2p neighboring the interactive regions were shown in figures, respectively. The red letter indicated to report as functional amino acids sequences.

8) *ATP*3 and *ATP*15 [19-21]

*ATP*3 and *ATP*15 were genes encoding F_1F_0-ATPase complex γ and ε subunits respectively, which participated in a rotation of the complex [19-21]. In this example the interactive regions both of *ATP*3 and *ATP*15 were unknown. However we could choose the entire coding region of *ATP*15 as the reference because the genome size of *ATP*15 was small (186 nt). Therefore, we used as w = 186 by SSM in this case. Other values, m, d, and p were the same, 10, 9, and 1, respectively as before. In addition, the reverse-complement base sequence of *ATP*15 was used because HF was higher in this analysis. We predicted the unknown interactive region of *ATP*3 by the SSM (**Figure 8**).

In x-ray crystallography of γ - ε complex of ATP synthase in *E. coli* and bovine, presumably, the 200th amino acid and the adjacent amino acids of γ - subunit (Atp3p) locating the foot-position could be interacted with ε - subunit (Atp15p) [19,20]. The prediction by SSM might be in accord with the results of these literatures for X-ray crystallography. The experiment to confirm the interactive regions of Atp15p and Atp3p analyzed by SSM is under the progress.

SSM was the analytical method to identify the base numbers (position from 5'-ATG = the start codon) of the interactive regions (sites) of the reference- and the target-protein. However there were not many examples where the interactive regions with the base numbers were identified for the reference and target proteins in the yeast genome databases such as SGD etc. Therefore we could not select many examples for the SSM analyses and showed all examples we have in this manuscript.

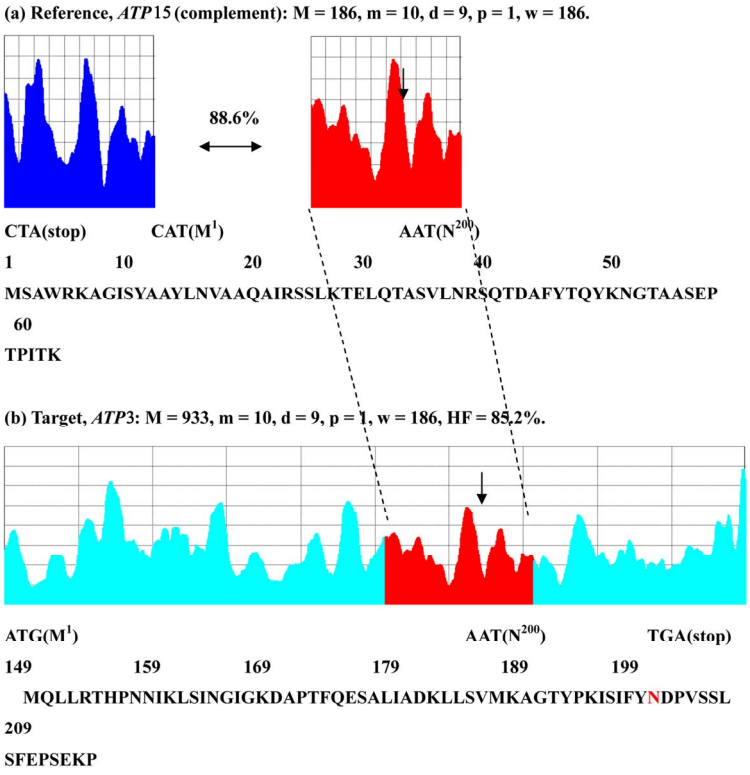

Figure 8. Sequence spectra of *ATP*15 and *ATP*3 genes (d = 9, m = 10，p = 1). The reverse-complement of the base sequence gave more homologous than the normal base sequence could be shown in the interaction *ATP15* (Atp15p)/*ATP*3 (Atp3p). (a) Coding region of *ATP*15 (Atp15p, M = 186, the active region was shown in purple-blue). (b) Coding region of *ATP*3 (Atp3p, target: M = 933). The region most similar to the reference is shown in red (HF = 88.6%). It has been shown that the amino acids sequences (shown in red letter) of Atp15p interacts with the amino acids sequences (shown in red region) of Atp3p [19-21]. The scales of axes in (a) and (b) are the same. The amino acid sequences of Atp15p and Atp3p neighboring the interactive regions were shown in figures, respectively. The arrowhead and the red letter amino acid residue, N^{200} of Atp3p might be interacted with Atp15 from X-ray crystallography [19,20].

The results in this paper could be sufficient to confirm the validity of SSM method because the probability to identify the interactive regions was very small by coincidence. For instance, in the case of *MAS*1 (Mas1p)/*MAS*2 (Mas2p), *MAS*2 was composed of about 1,400 nt, which meant that the identification probability by coincidence was lower than 1/7 (= 200 / 1400) under the condition of the homology window width w = 200 nt. The probabilities of other examples in this manuscript were following.

*PHO*4/*PHO*80, lower than 2/9 (= 200/900);
*RPB*12/*RPB*2, 1/20 (= 200/4000);
*GCR*2/*GCR*1, 1/15 (= 200/3000);
*SLA*2/*SLA*1, 1/20 (= 200/4000);
*PHO*2/*PHO*4, 1/15 (= *200*/*3000*);
*PHO*2/*SWI*5, 1/10 (= 200/2000);
*ATP*15/*ATP*3, 1/5 (= 200/1000);
*GAL*1/*GA*4, 1/15 (= 200/3000);
*GAL*4/*GAL*10, 2/7 (= 200/700);
*GAL*4/*GAL*2, 1/7 (= 200/1000);
*GAL*4/*GAL*7, 1/4 (= 200/800);

Therefore the results in this paper made sense statistically to confirm the validity of the proposed method. In addition the positions of the key amino acids were identical in the identified interactive regions in case of the examples of MAS and PHO proteins. This fact definitely reinforced the proposed method.

Finally we predicted the interactive regions of many proteins which were chosen randomly from 16 different chromosomes of *S. cerevisiae* [22], and summarize the prediction results in **Table 1** to demonstrate the effectiveness of SSM. For the examples in **Table 1** we used the same analytical conditions, m = 10, d = 9, p = 1 and w = 200, and predicted the interactive regions both of the reference and target proteins. However the proposed method in this paper was based on the condition that the interactive region of the reference protein was known and that of the target protein was unknown. Therefore some of these prediction results might be revised in our future work because the identification ability of SSM was not strong at present when the interactive regions both of the reference and target proteins were unknown. We are improving SSM to apply these cases now.

Table 1. Possible interactive region. The upper column indicated the 1st, and the lower column indicated the 2nd interactive region, respectively. *1) Conditions, m = 10, d = 9, p = 1, w = 200; *2) Reference gene; *3) Chromosome located the reference gene; *4) Amino acid residues of the reference protein; *5) Interactive region of the reference protein predicted by SSM; *6) Target gene; *7) Chromosome located the target gene; *8) Amino acid residues of the target protein; *9) Interactive region of the target protein; *10) Homology factor between the target to the reference protein; *11) Either protein was used as the reverse-complement base sequence.

Reference*2	Chromosome*3	Amino acids*4	Interactive reagion*5	Target*6	Chromosome*7	Amino Acids*8	Interactive region*9	HF (%)*10	Complement*11
GDH3	1	457	272-338	GDH1	15	454	116-182	93.7	
			52-118				52-118	92.3	
CDC24	1	854	183-249	ACT1	6	478	83-149	94.7	
			234-300				94-160	94.2	○
PHO11	1	467	374-440	PHO5	2	467	374-440	94.3	
			88-154				144-210	93	○
ATP2	10	511	170-236	ATP3	2	311	57-123	93.1	○
			300-366				20-86	92.6	
SUP45	2	437	311-377	RPS12	15	143	4-70	92.1	
			188-254				(-7)-59	91.3	
YDJ1	14	409	292-358	PRD1	3	712	508-574	94	
			67-133				552-618	93.6	○
GCD2	7	651	550-616	GCD7	12	381	274-341	94.9	○
			221-287				(-21)-45	91.4	
PHO87	3	923	433-499	SPL2	8	148	24-90	92.1	
			564-630				60-126	92	○
HXT15	4	567	323-389	GAL2	12	574	52-118	96.9	○
			483-549				399-465	96.8	○
NAB2	7	525	69-135	SNF3	4	884	480-544	95.4	
			243-309				175-241	95.2	○
ECM10	5	644	140-206	SSA1	1	642	237-303	94	
			70-136				395-461	93.4	○
HEM1	4	548	16-82	LCB2	4	561	296-362	95	○
			269-335				447-513	94.1	○
POL4	3	582	169-235	CCA1	5	546	218-284	97	○
			67-133				103-169	93.9	
GUT1	8	709	99-165	XKS1	7	600	24-90	94.5	
			649-(715)				(-12)-54	93.7	
YAP1	13	650	335-401	CAD1	4	409	96-162	93.9	
			531-597				217-283	93.3	

3.2. Mutual Interaction of Protein-DNA

This section clarified that the homology of sequence spectra was also related to the mutual interaction between protein and DNA. The interactions of the transcription factor *GAL*4 [23] and the promoters of *GAL* genes (UAS$_{Gal}$ signal, *GAL*1, *GAL*10, *GAL*2 and *GAL*7) [24-26] were taken as an example. **Figure 9** showed the sequence spectra of the upstream region of *GAL*1 as the reference (a) and the reverse-complement base sequence of the coding region of *GAL*4 as the target (b). We employed the upstream region of *GAL*1 to demonstrate the effectiveness of the method although its base size was 668 wh- ich was a little large for the reference region. In **Figure 9** the red region was the first identified region of *GAL*4. Surprisingly this red region is completely identical to the DNA binding region of *GAL*4 with the zinc finger motif, and the purple-blue region is the promoter region of *GA- L*1. This means that in this case the proposed method perfectly identified both the interactive reference (in purple-blue) and target regions (in red) at the same time despite the different objects, the protein region for *GAL*4 and the DNA region for *GAL*1.

Thus interactive analysis might be applied to other *GAL* genes, *GAL*10, *GAL*2, and *GAL*7, which their promoter regions were also interacted with the N-terminal DNA binding domain (zinc-finger domain) of *GAL*4 (Gal4p).

Figure 10 showed all the promoter regions identified by SSM with the DNA binding region of the Gal4p (the reverse-complement base sequence) in **Figure 9** as the reference region (in purple-blue). In this figure the reference region of *GAL*4 was fixed to arrange the layout of

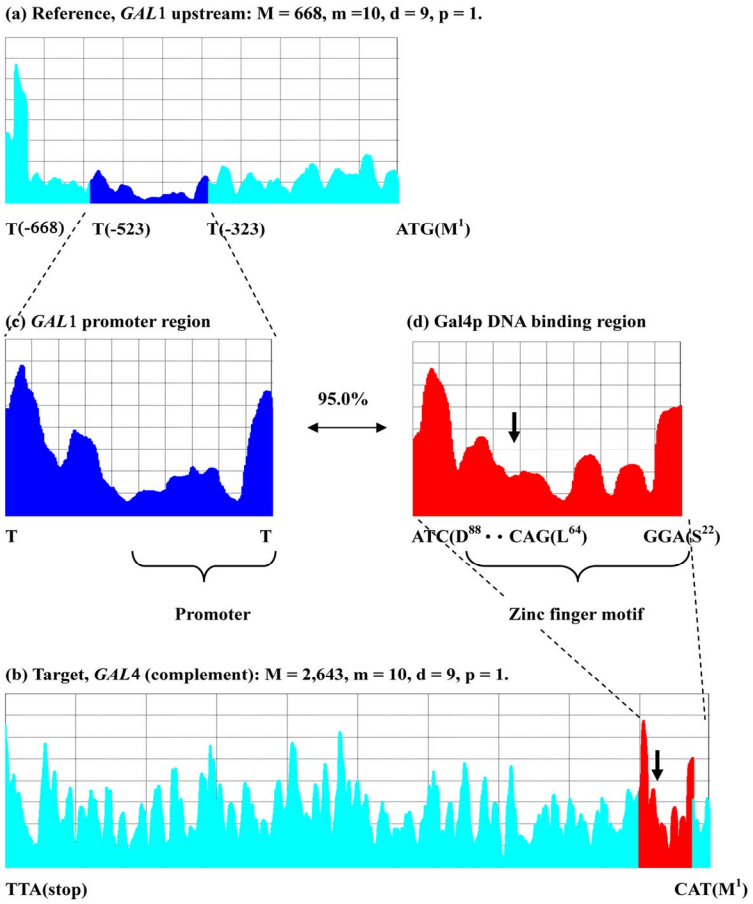

Figure 9. Sequence spectra of *GAL*1 and *GAL*4 genes (d = 9, m = 10， p = 1). (a) Upstream region of *GAL1* (668 nt) was used as the reference (in purple-blue). The arrowheads were indicated several promoter sequences. (b) DNA binding region of *GAL*4 (reverse-complement sequence of *GAL*4 (Gal4p, M = 2,643) was useful in comparison with *GAL*1 gene. The first 107 amino acids at the N-terminus of Gal4p, which is involved in DNA binding (shown in red, ref. 23), were used as the target. The bold arrowhead of Gal4p was indicated the position of L^{64}.

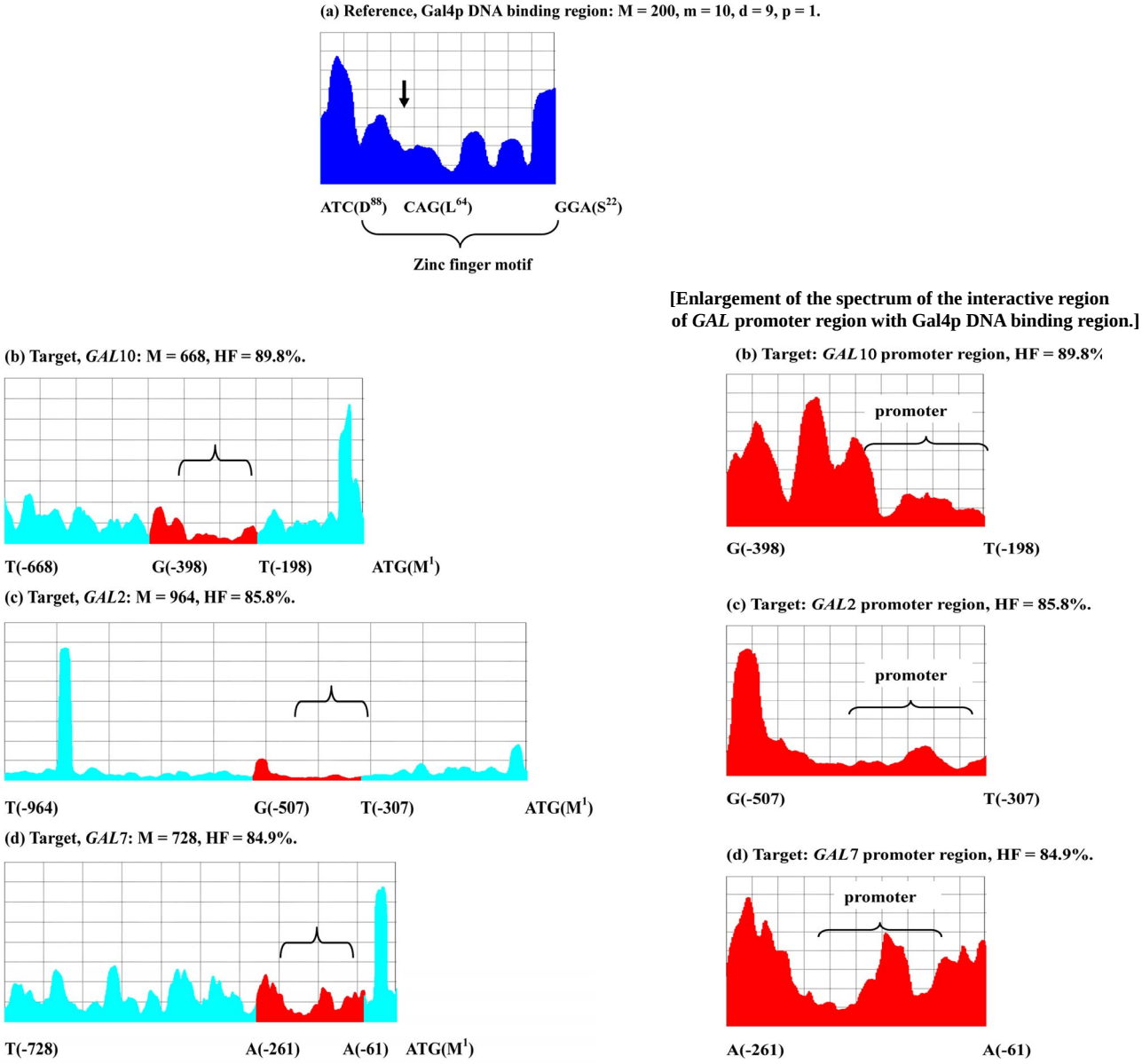

Figure 10. Sequence spectra of other *GAL* genes (d = 9, m = 10, p = 1). (a) DNA binding region of *GAL*4 (reverse-complement sequence of *GAL*4 (Gal4p, M = 200) was used as the reference (shown in purple-blue), and other *GAL* genes upstream, *GAL*10, *GAL*2 and *GAL*7 were as the target to search their promoter regions (the arrowheads were indicated several promoter sequences). (b) Upstream region of *GAL*10 (target: M = 668: HF = 89.8%). (c) Upstream region of *GAL*2 (target: M = 964: HF = 85.8%). (d) Upstream region of *GAL*7 (target: M = 728: HF = 84.9%). The bracket in each *GAL* gene indicated the promoter regions (upstream activator sequences, UAS_{Gal}) binding with the zinc finger motif of Gal4p [23-26]. The UAS_{Gal} signals (arrowhead) of each *GAL* gene were concentrated in the similar region shown in red. The red regions in (b), (c) and (d) were the most similar regions. The base numbers on the abscissa were matched in each panel either to the coding or upstream region. The bold arrowhead of Gal4p was indicated the position of L^{64}.

identified regions for the promoter. It was clear from this figure that the promoter sites in the red regions over lapped with each other. We obtained similar results for *PH- O* genes (data not shown).

3.3. Crucial Problems and Discussions

Our results raised various crucial problems below which were definitely related to fundamental principles of life. However we had to admit that we did not have perfect answer to these problems at the moment. Therefore our discussions below had some uncertain hypotheses.

[Question 1] Why was the sequence spectrum associated with functions of protein and DNA?

Originally the sequence spectrum was devised to ex-

amine the generation-rules in genome, and succeeded in visualizing the rules of reverse-complement symmetry, multiple fractality and so on. Therefore the fact that the sequence spectrum was associated with the functions of protein and DNA led to the fact that the generation-rules could govern not only the static base sequence in genome as the blueprint of life but also the dynamic phenomena of proteins and DNAs as the principle of life mechanism.

[Question 2] Why was the homology of sequence spectrum closely associated with the interaction of proteins?

A possible answer to this problem was that the sequence spectrum could reflect the higher order structure of proteins. The interacting region was considered to consist of the specific sequence of amino acids. This specificity of the amino acid sequence could be reflected to the appearance frequency of the base sequence corresponding to the amino acid sequence. The homology of the sequence spectrum could be interpreted to be an affinity of the interactive regions of the proteins.

[Question 3] Why was the homology of sequence spectrum closely associated with the interaction of protein and DNA?

Similarly to the problem [**Question 2**], a possible answer to this problem was that the sequence spectrum could reflect the higher order structure of protein and DNA. However, this fact would raise another crucial problem. Why could the sequence spectrum reflect the higher order of both protein and DNA in the same manner which was totally different objects? In order to answer this problem, it was definitely necessary to examine the relation between the higher order structures of protein and DNA (or RNA). Our results implied that there could exist a close structural relation between them. For instance, it was well known that a domain of EF-G factor protein emulated amino acyl-tRNA [26]. It could be even possible that the structure of protein could inherit the structure of its original DNA in genome because inheritance could be most simple answer for this problem. SSM basically could detect the interacting regions of gene DNAs through the homology of the sequence spectrum, and this automatically could lead to detect the interacting regions of proteins translated from the gene DNAs through the structure inheritance. We suspected that tRNA and codon table gave an important clue on this issue because tRNA were directly associated with the amino acid of protein and the triplet codon of DNA. Moreover the sequence spectrums of tRNA and protein possess the similar relation. For instance the GTP binding protein *RAS2* [27,28] and Gly(GGG)-tRNA which were both related to guanine(G) in common were similar in the sequence spectrum [2].

4. CONCLUSIONS

The conclusions obtained in this study were summarized as follows.

1) The homology of the sequence spectrum was closely associated with the interaction of protein and DNA.

2) The SSM was a suitable prediction method to identify interacting regions regardless of the biological macromolecules: DNA, RNA and protein.

3) The SSM was so fast and useful that it did not require a super computer but rather a personal computer.

4) The generation-rules in genome could govern not only the static base sequence in genome but also the dynamic phenomena of proteins and DNAs.

5) The sequence spectrum could reflect the higher order structure of protein and DNA.

6) There could be a close relation between the structures of protein and DNA.

The proposed method by SSM should be improved to identify or predict both the reference and target regions at the same time in any cases. This project is now ongoing in our laboratory and we will report on this subject in the next paper.

REFERENCES

[1] Takeda, M. and Nakahara, M. (2009) Structural features of the nucleotide sequences of genomes. *Journal of Computer Aided Chemistry*, **10**, 38-52.

[2] Nakahara, M. and Takeda, M. (2010) Characterization of the sequence spectrum of DNA based on the appearance frequency of the nucleotide sequences of the genome-A new method for analysis of genome structure. *Journal Biomedical Science and Engineering*, **3**, 340-350.

[3] Geli, V., Yang, M., Suda, K., Lustig, A. and Schatz, G. (1990) The MAS-encoded processing protease of yeast mitochondria. Overproduction and characterization of its two nonidentical subunits. *Journal of Biological Chemistry*, **265(31)**, 19216-19222.

[4] West, A.H., Clark, D.J., Martin, J., Neupert, W., Hartl, F.U. and Horwich, A.L. (1992) Two related genes encoding extremely hydrophobic proteins suppress a lethal mutation in the yeast mitochondrial processing enhancing protein. *Journal of Biological Chemistry*, **267(34)**, 24625-24633.

[5] Ito, A. (1999) Mitochondrial processing peptidase: multiple-site recognition of precursor proteins. *Biochemical and Biophysical Research Communication*, **265(3)**, 611-616.

[6] Nagao, Y., Kitada, S., Kojima, K., Toh, H., Kuhara, S., Ogishima, T. and Ito, A. (2000) Glycine-rich region of mitochondrial processing peptidase α-subunit is essential for binding and cleavage of the precursor proteins. *Journal of Biological Chemistry*, **275**, 34552-34556.

[7] Ogawa, N. and Oshima, Y. (1990) Functional domains of a positive regulatory protein, *PHO4*, for transcriptional control of the phosphatase region in *Saccharomyces cer-*

evisiae. *Molecular and Cellular Biology*, **10(5)**, 2224-2236.

[8] Okada, H. and Toh-e, A. (1992) A novel mutation occurring in the *PHO80* gene suppresses the *PHO4c* mutations of *Saccharomyces cerevisiae*. *Current Genetics*, **21(2)**, 95-99.

[9] Cramer, P., Bushnell, D.A. and Kornberg, R.D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 Angstrom resolution. *Science* **292(5523)**, 1863-1876.

[10] Baker, H.V. (1991) *GCR1* of *Saccharomyces cerevisiae* encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC sequence motif. *Proceeding National Academy of Sciences of the United States of America*, **88(21)**, 9443-9447.

[11] Uemura, H. and Jigami, Y. (1992) Role of *GCR2* in transcriptional activation of yeast glycolytic genes. *Molecular and Cellular Biology*, **12(9)**, 3834-3842.

[12] Deminoff, S.J., Tornow, J. and Santangelo, G.M. (1995) Unigenic evolution: A novel genetic method localizes a putative leucine zipper that mediate dimerization of the *Saccharomyces cerevisiae* regulator Gcr1p. *Genetics,* **141(4)**, 1263-1274.

[13] Deminoff, S.J. and Santangelo, G.M. (2001) Rap1p requires Gcr1p and Gcr2p homodimers to activate ribosomal protein and glycolytic genes, respectively. *Genetics,* **158(1)**, 133-143.

[14] Gourlay, C.W., Dewar, H., Warren, D.T., Costa, R., Satish, N. and Ayscough, K.R. (2003) An interaction between Sla1p and Sla2p plays a role in regulating actin dyn- amics and endocytosis in budding yeast. *Journal of Cell Science*, **116(12)**, 2551-2564.

[15] Liu, C., Yang, Z., Yang, J., Xia, Z., and Ao, S. (2000) Regulation of the yeast transcription factor PHO2 activity by phosphorylation. *Journal of Biological Chemistry*, **275(41)**, 31972-31978.

[16] Yang, J. and Ao, S.Z. (1996) Interaction of the yeast PHO2 protein or its mutants with the PHO5 UAS *in vitro*. *Sheng Wu Hua Xue Yu Sheng Wu Li Xue Bao* (Shanhai) **28(3)**, 316-320.

[17] Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y. and Hakoshima, T. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: Flanking base recognition. *EMBO Journal,* **16(15)**, 4689-4697.

[18] Bhoite, L.T. and Stillman, D.J. (1998) Residues in the Swi5 zinc finger protein that mediate cooperative DNA binding with the Pho2 homeodomain protein. *Molecular and Cellular Biology*, **18(11)**, 6436-6446.

[19] Rodgers, A.J. and Wilse, M.C. (2000) Structure of the gamma-epsilon complex of ATP synthase. *Nature Structural Biology*, **7(2000)**, 1051-1054.

[20] Montgomery, G.C., Lesile, A.G. and Walker, J.E. (2000) The structure of the central stalk in bovine F(1)-ATPase at 2.4 A resolution. *Nature Structural Biology*, **7(11)**, 1055-1061.

[21] Tsumuraya, M., Furuike, S., Adachi, K., Kinoshita, K. jr. and Yoshida, M. (2009) Effect of ε subunit on the rotation of thermophilic Bacillus F_1-ATPase. *FEBS Letters*, **583(7)**, 1121-1126.

[22] Saccharomyce G.D (2010) (http://www.yeastgenome.org/).

[23] Ding, W.V. and Johnston, S.A. (1997) The DNA binding and activation domains of Gal4p are sufficient for conveying its regulatory signals. *Molecular and Cellular Biology*, **17(5)**, 2538-2549.

[24] Johnston, M. and Davis, R.W. (1984) Sequences that regulate the divergent *GAL1-GAL10* promoter in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, **4(11)**, 1440-1448.

[25] Lorch, Y. and Kornberg, R.D. (1985) A region flanking the *GAL7* gene and binding site for *GAL4* protein as upstream activating sequences in yeast. *Journal of Molecular Biology*, **186(4)**, 821-824.

[26] Tajima, M., Nogi, Y. and Fukazawa, T. (1986) Duplicate upstream activating sequences in the promoter region of the *Saccharomyces cerevisiae GAL7* gene. *Molecular and Cellular Biology*, **6(1)**, 246-256.

[27] Nissen, P., Kjeldgaard, M., Thirup, S., Polekhina, G., Reshetnikova, L., Clark, B.F. and Nyborg, J. (1995) Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. *Science,* **270(5241)**, 1464-1472.

[28] Kataoka, T., Powers, S., McGill, C., Fasano, O., Strathern, J., Broach, J. and Wigler, M. (1984) Genetic analysis of yeast *RAS*1 and *RAS*2 genes. *Cell,* **37(2)**, 437- 445.

[29] Mabuchi, T., Ichimura, Y., Takeda, M. and Douglas, M.G. (2000) *ASC*1/*RAS*2 suppresses the growth defect on glycerol caused by the *atp*1-2 mutation in the yeast *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, **275(14)**, 10492-10497.

