Pacific Symposium on Biocomputing 15:88-97(2010)

IMPROVEMENT OF STRUCTURE CONSERVATION INDEX WITH CENTROID ESTIMATORS

YOHEI OKADA

Department of Biosciences and Informatics, Keio University,
3–14–1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223–8522, Japan
E-mail: okada@dna.bio.keio.ac.jp

KENGO SATO

Japan Biological Informatics Consortium (JBIC),
2–45 Aomi, Koto-ku, Tokyo 135–8073, Japan
E-mail: sato-kengo@aist.go.jp

YASUBUMI SAKAKIBARA

Department of Biosciences and Informatics, Keio University,
3–14–1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223–8522, Japan
E-mail: yasu@bio.keio.ac.jp

RNAz, a support vector machine (SVM) approach for identifying functional non-coding RNAs (ncRNAs), has been proven to be one of the most accurate tools for this goal. Among the measurements used in RNAz, the Structure Conservation Index (SCI) which evaluates the evolutionary conservation of RNA secondary structures in terms of folding energies, has been reported to have an extremely high discrimination capability. However, for practical use of RNAz on the genome-wide search, a relatively high false discovery rate has unfortunately been estimated. It is conceivable that multiple alignments produced by a standard aligner that does not consider any secondary structures are not suitable for identifying ncRNAs in some cases and incur high false discovery rate. In this study, we propose C-SCI, an improved measurement based on the SCI applying γ-centroid estimators to incorporate the robustness against low quality multiple alignments. Our experiments show that the C-SCI achieves higher accuracy than the original SCI for not only human-curated structural alignments but also low quality alignments produced by CLUSTAL W. Furthermore, the accuracy of the C-SCI on CLUSTAL W alignments is comparable with that of the original SCI on structural alignments generated with RAF for which 4.7-fold expensive computational time is required on average.

Keywords: structure conservation index; centroid estimators; non-coding RNAs

1. Introduction

Many studies have recently discovered essential roles of non-protein-coding functional RNAs (ncRNAs) in cells such as translation, post-transcriptional gene regulation and maturation of rRNAs, tRNAs and mRNAs.[1,2] Therefore, to identify ncRNAs in genomes and analyze their functions is a crucial task for not only molecular cell biology but also bioinformatics.

It is well-known that such biological functions of ncRNAs are deeply related to their secondary structures which consist of hydrogen-bonded base-pairs including the Watson-Crick base-pairs (`A-U` and `G-C`), the wobble base-pairs (`G-U`) and other non-canonical base-pairs. These base-pairs stabilize the structure of RNAs in terms of the free energy. Thus, the secondary structure with the minimum free energy (MFE) has been regarded as the most reliable prediction of RNA secondary structures.

However, MFE alone could not be an appropriate measure for identifying ncRNAs since the free energy is heavily biased by the nucleotide composition.[3] Therefore, several comparative approaches for identifying ncRNAs have been proposed.[4–10] For this purpose, Washietl *et al.* have developed RNAz which uses a support vector machine (SVM) approach, and have proposed the Structure Conservation Index (SCI) as a feature to measure the evolutionary conservation in terms of secondary structures.[7] Assuming that MFE for the consensus secondary structure is close to that for each sequence if a given multiple alignment is structurally conserved, the SCI is defined as the rate of MFE for the common secondary structure to averaged MFE for

Pacific Symposium on Biocomputing 15:88-97(2010)

each sequence. MFEs for each sequence and the common secondary structure are calculated by RNAfold and RNAalifold both part of the Vienna RNA packages,[11,12] respectively.

RNAz with the SCI has been proven to be the one of the most accurate tools for identifying ncRNAs.[7,13] However, for practical use of RNAz on the genome-wide search, a relatively high false discovery rate has unfortunately been estimated.[14] It is conceivable that multiple alignments produced by a standard aligner that does not consider any secondary structures are not suitable for identifying ncRNAs in some cases and incur high false discovery rate. Wang *et al.* have also suggested that the genome-wide alignments in the UCSC Genome Browser[15] produced by MULTIZ[16] should be improved in some regions for identifying ncRNAs.[17] To improve the accuracy, two strategies can be considered: the one is to employ a structural aligner such as RAF[18] to produce high quality alignments, and the other is to develop a more robust method against low quality alignments. Since the former strategy will consume impractical execution time for structural alignments, this study takes the latter strategy.

Recently, CentroidFold which employs γ-centroid estimators for predicting RNA secondary structures has been developed, and has been shown to be more accurate than other existing tools.[9] Especially, CentroidFold can predict much more accurate common secondary structures for low quality multiple alignments produced by CLUSTAL W[19] than RNAalifold.

In this study, we propose C-SCI, an improved measurement based on the SCI applying γ-centroid estimators for (common) secondary structure prediction, instead of RNAfold and RNAalifold, to incorporate the robustness against low quality multiple alignments. Our experiments show that the C-SCI achieves higher accuracy than the original SCI for not only human-curated structural alignments but also low quality alignments produced by CLUSTAL W. Furthermore, the accuracy of the C-SCI on CLUSTAL W alignments is comparable with that of the original SCI on RAF alignments for which 4.7-fold expensive computational time is required on average.

2. Method

2.1. *Structure Conservation Index*

The Structure Conservation Index (SCI) evaluates secondary structure conservation of a given multiple alignment of RNAs in terms of the minimum free energy (MFE). We denote with $\mathcal{S}(x)$ the entire folding space of a single sequence x and denote with $\mathcal{S}(A)$ the entire consensus folding space of an alignment A. The SCI is defined as

$$SCI(A) = \frac{E_{Align}(y_A^{MFE})}{\frac{1}{\#A}\sum_{x\in A} E(y_x^{MFE})}, \tag{1}$$

where $\#A$ is the number of sequences in the alignment A. For a single sequence x, $E(y)$ denotes the free energy of a secondary structure $y \in \mathcal{S}(x)$, and $y_x^{MFE} = \arg\min_{y\in\mathcal{S}(x)} E(y)$ is defined to be the MFE structure of x calculated by RNAfold.[11] Similarly, for an alignment A, $E_{Align}(y)$ is the free energy of a consensus structure $y \in \mathcal{S}(A)$, and $y_A^{MFE} = \arg\min_{y\in\mathcal{S}(A)} E_{Align}(A)$ is the consensus MFE structure of A calculated by RNAalifold.[12] The free energy of a consensus structure is defined as the average of the energy contributions of the single sequences plus covariance scores for bonuses of compensatory and consistent co-mutation in the alignment.

The consensus MFE alone could be used to identify functional RNAs likelihood of functional RNAs in terms of thermodynamic stability of consensus folded structures. However, it is difficult to make straightforward use of it, since the folding energy is heavily biased by the nucleotide composition and the length of the alignment. The SCI solved this problem by normalizing $E_{Align}(y_A^{MFE})$ with the average of $E(y_x^{MFE})$ for all $x \in A$. From a different view, the SCI reflects the idea that for a well-conserved alignment the structure of each sequence resembles each other and the consensus structure resembles all of them, so $E_{Align}(y_A^{MFE})$ would have as low value as $E(y_x^{MFE})$, otherwise $E_{Align}(y_A^{MFE})$ would not. The SCI is near 0 for an alignment that is not structurally conserved, whereas the SCI is near 1 or above for an alignment that is structurally

Pacific Symposium on Biocomputing 15:88-97(2010)

conserved. Especially, if the alignment is structurally well-conserved and compensatory and consistent mutation often occurs, the SCI may be above 1.

As shown in the definition (1) of the SCI, the SCI obviously depends on the accuracy of common secondary structure prediction, which is also deeply influenced by the quality of multiple alignments of RNAs. This fact is supported by a previous study[20] and our results shown in Sec. 3. For the genome-wide search, high quality alignments that consider RNA secondary structures cannot be obtained easily due to the computational cost for calculating structural alignments. Therefore, a robust method that does not require high quality alignments is required.

2.2. *γ-Centroid Estimator*

CentroidFold implements a γ-centroid estimator which predicts secondary structures with the maximum expected accuracy by a kind of posterior decoding methods on the base-pairing probability matrix. CentroidFold employs a gain function between a true secondary structure θ and a predicted secondary structure y on x defined as

$$G_\gamma(\theta, y) = \sum_{1 \le i < j \le |x|} \{\gamma I(y_{ij} = 1) I(\theta_{ij} = 1) + I(y_{ij} = 0) I(\theta_{ij} = 0)\}, \tag{2}$$

where γ is a weight for base-pairs, y_{ij} is 1 if the i-th and j-th nucleotides form a base-pair in y, and $I(\mathit{condition})$ is the indicator function, which takes 1 or 0 relying on whether *condition* is true or false. The gain function (2) is equal to the weighted sum of the number of true positives and the number of true negatives of base-pairs. CentroidFold predicts a secondary structure $y \in \mathcal{S}(x)$ which maximizes the expectation of $G_\gamma(\theta, y)$ with respect to an ensemble of all possible secondary structure $\mathcal{S}(x)$ which is distributed under a posterior distribution $p(\theta|x)$,

$$\begin{aligned} \mathbb{E}_{p(\theta|x)}[G_\gamma(\theta, y)] &= \sum_{\theta \in \mathcal{S}(x)} G_\gamma(\theta, y) p(\theta|x) \\ &= \sum_{1 \le i < j \le |x|} ((\gamma + 1) p_{ij} - 1) I(y_{ij} = 1) + C, \end{aligned} \tag{3}$$

where C is a constant independent of y, and $p_{ij} = \mathbb{E}_{p(\theta|x)}[\theta_{ij}]$ is the base-pairing probability that the i-th and j-th bases form a base-pair. The optimal secondary structure $\hat{y} = \arg\max_{y \in \mathcal{S}(x)} \mathbb{E}_{p(\theta|x)}[G_\gamma(\theta, y)]$ can be calculated efficiently by using the following DP algorithm:

$$M_{i,j} = \max \begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + (\gamma + 1) p_{ij} - 1 \\ \max_k [M_{i,k-1} + M_{k,j}] \end{cases}, \tag{4}$$

and tracing back from $M_{1,|x|}$ to calculate $\hat{y}$. The model of the posterior distribution $p(\theta|x)$ can be chosen from various implementations including the McCaskill model[21] based on the Boltzmann free energy and the CONTRAfold model[22] based on a machine learning technique.

CentroidFold can also predict a common secondary structure of a multiple alignment of RNAs by using averaged γ-centroid estimators. The optimal common secondary structure which maximizes the sum of the expected gain (2) for all $x \in A$, that is,

$$\hat{y} = \arg \max_{y \in \mathcal{S}(A)} \sum_{x \in A} \mathbb{E}_{p(\theta|x)}[G_\gamma(\theta, y)]$$

can similarly be calculated by using (4) with the averaged base-pairing probability[23] defined as

$$\bar{p}_{ij} = \frac{1}{\#A} \sum_{x \in A} \mathbb{E}_{p(\theta|x)}[\theta_{ij}],$$

Pacific Symposium on Biocomputing 15:88-97(2010)

instead of p_{ij}.

The weight γ in the definition (2) controls the number of predicted base-pairs, that is, the trade-off between specificity and sensitivity of predicted base-pairs. If $\gamma = 1$, this estimator is equivalent to the centroid estimator.[24,25]

CentroidFold has been shown to be more accurate than other existing tools.[9] Especially, CentroidFold can predict much more accurate common secondary structures than RNAalifold for low quality multiple alignments produced by CLUSTAL W.

2.3. *The C-SCI*

Now, we propose an improved measurement of secondary structure conservation based on the SCI by employing CentroidFold for (common) secondary structure prediction, instead of RNAfold and RNAalifold, to incorporate the robustness against low quality multiple alignments.

At first, we predict the consensus centroid structure for an alignment A, denoted by y_A^C, and centroid structures for each sequence $x \in A$, denoted by y_x^C, by using CentroidFold. For a single sequence, we map a predicted structure onto each sequence x and calculate its free energy $E(y_x^C)$ for all of the sequences. For an alignment, we map a predicted consensus structure onto each sequence x and get rid of gaps and corresponding parts of the structure. In removing a gap, if the part of structure corresponding to the gap is represented as unpaired, the compartment is removed, whereas if the corresponding part is represented as paired, the compartment is removed and its pair is removed or converted to unpaired depending on whether the pair corresponds gap or not. To calculate the energy, we use RNAeval[11] with the predicted structure on the sequence. The free energy of a consensus secondary structure is calculated from the averaged free energy for all sequences and the covariance score which is implemented according to RNAalifold.[12]

Then, the C-SCI is calculated as follows:

$$C\text{-}SCI(A) = \frac{E_{Align}(y_A^C)}{\frac{1}{\#A}\sum_{x \in A} E(y_x^C)}. \tag{5}$$

The C-SCI has two parameters which affect the discrimination capability of the C-SCI. We denote γ_A as the parameter γ for predicting consensus secondary structures on multiple alignments, and γ_S as γ for predicting secondary structures on single sequences. These parameters were determined by 10-fold cross-validation with the grid search on $\gamma \in \{2^k : -10 \leq k \leq 10, k \in \mathbb{Z}\}$ for γ_A and γ_S. The detail of how the 10-fold cross-validation was performed is written in the section 3.1. Furthermore, the C-SCI has a modification that if the predicted structure is unstable and the energy has positive value, the energy is treated as 0. This is because C-SCI may get a high value regardless of secondary structure conservation, if the numerator and the denominator of C-SCI are positive.

3. Result

3.1. *Evaluation*

To confirm the discrimination capability of the C-SCI, we performed the experiments along with the previous study[26] on BRAliBase 2.1 data set,[27] which is constituted with 18,990 reference alignments of 36 RNA families and the same number of the corresponding sets of sequences which are not aligned. Reference alignments included in BRAliBase 2.1 are human-curated alignments which are made from Rfam database[28] aiming for evaluating structural alignments. We also produced multiple alignments using CLUSTAL W[19] version 1.83 with standard settings to investigate the discrimination capability on low quality alignments. For each alignment, we generated negative controls by utilizing `shuffle-aln.pl`.[29] This program shuffles columns of a given alignment to destroy its secondary structure, while maintaining gap patterns, nucleotide compositions and sequence length. We generated five negative controls for each alignment. These alignments were binned according to their normalized Shannon entropy by the size of 0.05. The normalized Shannon entropy is defined as the average of the Shannon entropy for the individual column over all columns in the

Pacific Symposium on Biocomputing 15:88-97(2010)

Table 1. Detail information about reference alignments.

entropy	number of alignments						average pairwise sequence identity					
	2	3	5	7	10	15	2	3	5	7	10	15
0.10	827	111	11	2	0	0	92.3	93.8	94.7	94.9	—	—
0.15	922	329	48	27	16	6	87.5	90.9	93.1	93.6	93.9	94.0
0.20	974	502	148	50	16	8	83.1	87.4	89.9	91.0	91.7	92.8
0.25	432	479	253	158	58	18	77.6	84.1	87.2	88.3	89.2	89.4
0.30	391	178	262	138	108	65	72.6	80.6	84.6	86.0	87.0	87.7
0.35	456	108	71	95	71	47	67.4	76.6	81.8	83.7	84.8	85.5
0.40	554	134	32	23	16	14	62.6	72.7	78.3	80.6	82.4	83.3
0.45	588	194	48	10	8	3	57.4	68.5	74.2	75.8	79.5	77.8
0.50	559	195	68	38	13	5	52.6	64.5	69.9	71.0	72.1	75.0
0.55	739	194	83	53	27	16	47.5	61.1	65.8	67.0	69.3	69.5
0.60	797	196	82	44	34	20	42.5	58.0	64.1	65.0	66.6	67.2
0.65	589	234	61	21	10	3	37.7	55.1	63.9	62.7	64.2	63.4
0.70	478	320	43	10	5	2	32.5	51.7	61.6	63.1	64.6	63.8
0.75	244	274	39	18	2	1	27.8	48.2	57.9	62.9	59.6	58.9
0.80	126	313	71	17	8	6	22.6	44.3	54.6	56.4	61.6	59.5
0.85	37	326	117	22	11	2	18.2	40.9	53.2	56.1	59.4	60.5
0.90	2	227	139	39	12	2	14.1	37.1	49.9	55.1	56.6	56.2
0.95	0	130	125	68	24	2	—	33.8	46.4	51.2	54.2	58.9
1.00	0	131	168	62	25	13	—	31.0	43.6	49.2	51.8	53.8
1.05	0	41	141	79	46	18	—	29.2	41.6	45.3	49.4	52.2
1.10	0	4	100	99	34	18	—	26.2	39.1	42.7	45.9	51.1
1.15	0	2	61	92	48	25	—	24.0	37.5	40.5	42.7	44.2

In the content of "number of alignments", each column corresponds to the number of alignments constituted with the designated number of sequences (2, 3, 5, 7, 10 or 15 sequences). Similarly in the content of "average pairwise sequence identity", each column means the average of average pairwise sequence identity in the alignments with the designated number of sequences.

alignment whose length is $|A|$,

$$H = -\frac{1}{|A|}\sum_{i=1}^{|A|}\sum_{j\in\Sigma} p_j^i \log_2 p_j^i, \tag{6}$$

where j is in the alphabet $\Sigma = \{\mathtt{A}, \mathtt{U}, \mathtt{G}, \mathtt{C}, \mathtt{-}\}$ constituted with the four nucleotides and the gap character "-", and p_j^i is the probability observing the character j in column i. We used the alignments in the bins from 0.1 to 1.15 of normalized Shannon entropy according to Ref. 26. The number of alignments and the average of averaged pairwise sequence identity (APSI) on reference alignments for each normalized Shannon entropy bin are summarized in Tab. 1. This shows that higher entropy regions tend to include the alignments with lower APSI or with larger number of sequences. Therefore, most of alignments with small number of sequences appear in low entropy region.

To evaluate the performance of the various strategies, we performed the receiver operating characteristic (ROC) curve analysis. An ROC curve is a plot of true positive rate versus false positive rate in varying the discrimination threshold of a classifier. The area under the ROC curve (AUC) is used for evaluation of the discrimination: the shift of AUC to 1 means better discrimination capability. Calculation of AUC for each entropy subset was done by using ROCR package.[30]

In our study, we compared the C-SCI with the original SCI and the measurement "base-pairing distance" (pairwise, consensus), which have been reported to achieve as high AUC as the SCI.[26] Base-pairing distance is a measurement to compare two single structures by using the Hamming distance. Here "pairwise" means the comparison of each structure of a single sequence with each other, and "consensus" means the comparison of each structure of a single sequence with the consensus structure. For the structures compared in base-pairing distance, we adopted MFE structures. The SCI and base-pairing distance were implemented by using RNAfold with options "`-d2`" and RNAalifold with options "`-d2`". RNAalifold has recently been updated by replacing the simple covariance scores with a more sophisticated RIBOSUM[31]-like scoring matrices.[32] However, it has been reported that the new covariance scoring matrices failed to improve the accuracy of

Pacific Symposium on Biocomputing 15:88-97(2010)

Table 2. Averaged AUC of each measurement.

Method	Reference	CLUSTAL W
C-SCI (McCaskill model)	0.950	0.899
C-SCI (CONTRAfold model)	0.955	0.912
SCI	0.927	0.853
Base-pair distance (consensus)	0.905	0.849
Base-pair distance (pairwise)	0.900	0.854

the SCI although this update improved the accuracy of common secondary structure predictions. Therefore, we employed the previous covariance scores described in Ref. 12. For implementing the C-SCI, we used CentroidFold version 0.0.4 for predicting secondary structures, and RNAeval[11] from Vienna RNA Package version 1.7.2 for calculating the free energy of predicted structures. The C-SCI has two parameters γ_A and γ_S, and we performed 10-fold cross-validation for each bin of normalized Shannon entropy. We determined γ_A and γ_S which maximize AUC on the 90% of the dataset in each bin, and calculate AUC on the rest 10% of the dataset. As for evaluation we adopted the average of 10 AUCs.

3.2. *Discrimination capability*

Figure 1 shows the results of AUC analysis of the C-SCI (McCaskill model, CONTRAfold model), the SCI and base-pair distance (consensus, pairwise) on reference alignments and CLUSTAL W alignments for each bin of normalized Shannon entropy, indicating that the C-SCI achieved the highest AUC, especially on low entropy region. Table 2 shows the summarized result by averaging AUC values in all bins. This indicates that the C-SCI achieves higher AUC on both alignments than the other measurements. Especially with CONTRAfold model, the C-SCI achieved the highest AUC in the C-SCI variants. The parameters γ_A and γ_S used in 10-fold cross-validation are written in Table S1 in the Supplemental material.

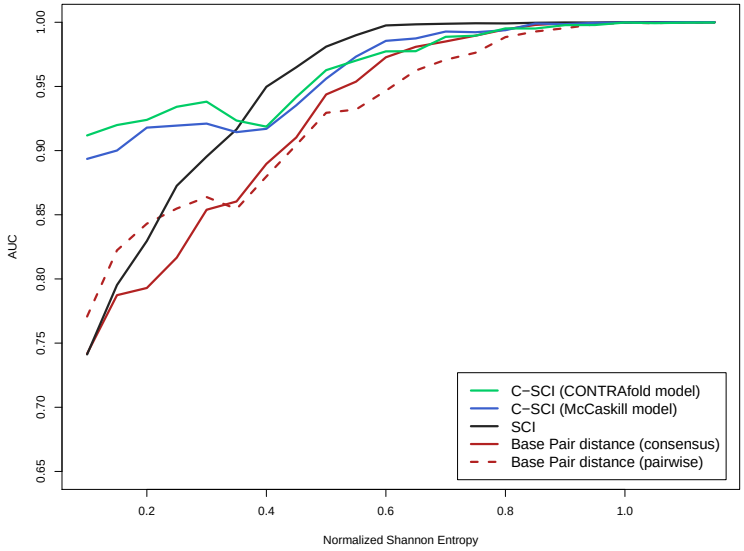
Furthermore, to investigate the reason why on the low entropy region the C-SCI could achieve extremely higher AUC than the other measurements on both alignments, we plotted the behavior of the median value of the score on reference alignments for each bin of normalized Shannon entropy. To clarify the difference between the SCI and the C-SCI, we show the result of two measurements: the SCI and the C-SCI with CONTRAfold model which has the highest averaged AUC as shown in Fig. 2. For the SCI, the score distribution of the positive data and that of the negative data is so close that even 25%-quantile of the positives and 75%-quantile of the negatives overlap on low entropy region. On the other hand, the C-SCI could clearly separate these score distributions well with γ_A and γ_S optimized by 10-fold cross-validation for each bin of normalized Shannon entropy. This suggests that the C-SCI has higher discriminant power than the SCI, especially, on low entropy region.

3.3. *Computational complexity*

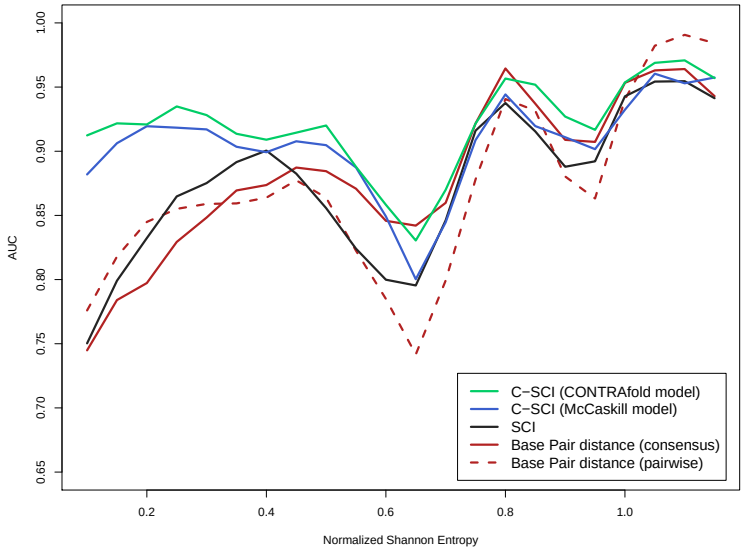
To address the genome-wide search, the computational cost is a serious problem. It is obvious that the use of reference alignments which are structurally corrected by human curation is impractical although all the measurements achieve high accuracy on reference alignments. Two alternative approaches are to use structural aligners which can align RNA sequences with conserving their secondary structures, and to use the standard aligners like CLUSTAL W. We produced the structural alignments using RAF[18] version 1.00 with default settings. RAF is one of the most efficient structural aligners based on the Sankoff algorithm[33] which simultaneously aligns and folds given RNA sequences. All the experiments was executed on a Linux machine with AMD Opteron 2200SE (2.8GHz).

As shown in Tab. 3, all the measurements on RAF alignments achieve as high accuracy as those on reference alignments, and much higher than those on CLUSTAL W alignments. However, huge computational time is required for producing structural alignments even by RAF (the elapsed time: 1.86 seconds on average), which is known as the most efficient structural aligner, comparing with CLUSTAL W (0.0147 seconds on average). On the other hand, our approach the C-SCI, has an advanced property of robustness against low

Pacific Symposium on Biocomputing 15:88-97(2010)



(a) reference alignments

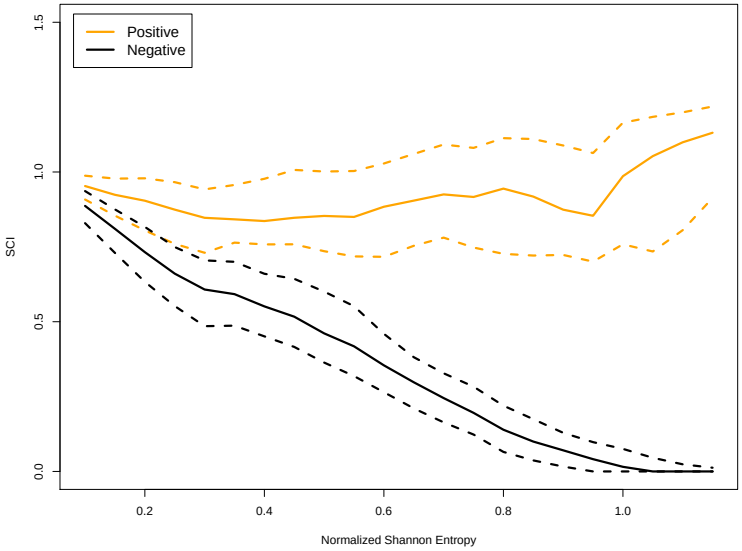


(b) CLUSTAL W alignments

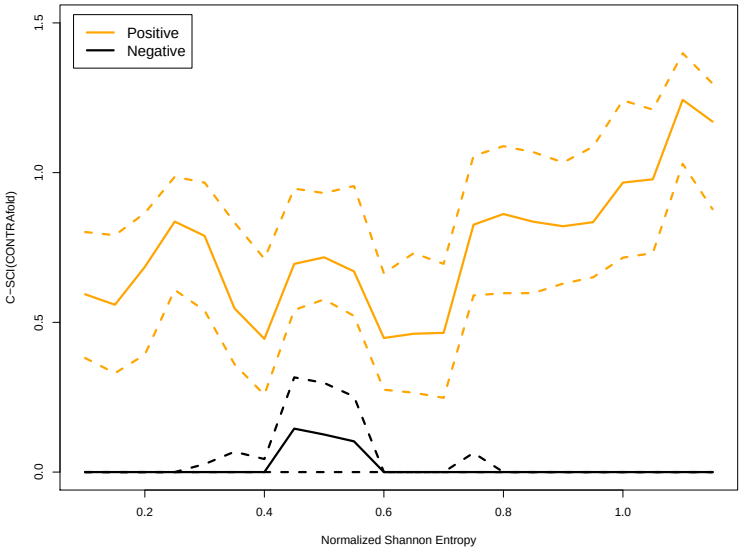
Fig. 1. The discrimination capacity of the C-SCI, the SCI and base-pair distance in AUC on reference alignments and CLUSTAL W alignments for each bin of normalized Shannon entropy.

quality alignments. In fact, Tab. 3 indicates that averaged AUC of the C-SCI with CONTRAfold model on CLUSTAL W alignments is comparable with that of the SCI on RAF alignments. Furthermore, the elapsed time for calculating the SCI through RAF alignments was 1.99 seconds for each alignment on average, whereas that of the C-SCI with CONTRAfold model through CLUSTAL W alignments was only 0.426 seconds for each alignment on average. In case that structural alignments might be unavailable such as the genome-wide search, the C-SCI is practical to use and is expected to have as high discriminant power as the SCI on structural alignments.

Pacific Symposium on Biocomputing 15:88-97(2010)



(a) SCI



(b) C-SCI (CONTRAfold model)

Fig. 2. The behavior of median of the SCI and the C-SCI distribution along with the entropy on reference alignments. In each of positive and negative line sets, the solid line means the median of the distribution and the lower and upper dashed lines mean 25%-quantile and 75%-quantile of the distribution.

4. Discussion

We proposed the C-SCI, an improved measurement of secondary structure conservation, and examined its performance. The result was summarized in Tab. 2, which shows that the C-SCI is much more discriminative than the SCI and other measurements. This is because the C-SCI outperforms others in low entropy area as shown in Fig. 1. By the observation that important genes have high sequence identities between related species on the sequence alignment, the alignments with high sequence identity can be in the major part of data on which calculation of the measurement are performed. Therefore, the improvement of the accuracy on low entropy will be of great benefit. Table 2 also shows that the C-SCI with CONTRAfold model exceeds the C-SCI with McCaskill model. This is because CONTRAfold model has more appropriate parameters to

Pacific Symposium on Biocomputing 15:88-97(2010)

Table 3. Calculation time and AUC of each measurement. The result of calculation time is shown on the second scale.

	RAF			CLUSTAL W		
Method	AUC	Time[a]	Total time[b]	AUC	Time[a]	Total time[b]
C-SCI (McCaskill model)	0.953	0.241	2.10	0.899	0.222	0.237
C-SCI (CONTRAfold model)	0.957	0.445	2.30	0.912	0.411	0.426
SCI	0.923	0.130	1.99	0.853	0.135	0.150
Base-pair distance (consensus)	0.908	0.195	2.06	0.849	0.188	0.203
Base-pair distance (pairwise)	0.901	0.179	2.04	0.854	0.166	0.181

Time[a]: elapsed time for calculating the SCI or the C-SCI only. Total time[b]: total elapsed time for aligning sequences and calculating the SCI or the C-SCI.

estimate a secondary structure.

To examine the reason why the improvement on low entropy region occurs, we further calculated the score distribution of positive data and negative data of each measurement and showed that the C-SCI could separate these data more clearly in Fig. 2. This also shows that the median of the C-SCI on negative controls gets close to 0, whereas that of the SCI does not in low entropy region. We can discuss two things: why the C-SCI on negative controls tends to get close to 0 and why the C-SCI exhibits higher discrimination capability than the SCI. For the former question, we suppose that this is because the tendency that the consensus secondary structure is the open chain or an unstable structure is strong on the negative controls whereas not on the positive data with the proper γ_A and γ_S. For the latter one, we suppose that this is because MFE of consensus structure does not increase so much by shuffling columns, whereas the energy of the consensus structure calculated by a γ-centroid estimator increase significantly. Note that we cannot exclude the possibility that the shuffling algorithm used in our experiments does not work uniformly for all the bin of the entropy to preserve gap patterns and conservation patterns of columns. The number of possible pairs of columns to be shuffled depends on the gap patterns and conservation patterns which are reflected in the entropy. Further investigation should be done by using more sophisticated negative controls generating algorithms such as `SISSIz`[10] which can preserve dinucleotide composition in alignments in expectation.

Moreover, we investigated the computational time for calculating measurement and alignment shown in Tab. 3. This shows that the C-SCI on CLUSTAL W alignment is expected to be as discriminative as the SCI on structural alignment although the C-SCI through CLUSTAL W alignment is 4.7 times faster. Hence the C-SCI is practical, also considering that calculating structural alignments is not reasonable for genome-wide search. We can conclude that the C-SCI is computationally practical to use as well as much more discriminative than the SCI.

Supplemental materials

The parameters γ_A and γ_S optimized by 10-fold cross-validation for each bin of normalized Shannon entropy in CLUSTAL W alignments are written in the supplemental material. See the following file.
http://www.dna.bio.keio.ac.jp/~okada/psb2010/supplemental1.pdf

Acknowledgements

This work was supported in part by a grant from "Functional RNA Project" funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan, and was also supported in part by Grant-in-Aid for Scientific Research on Priority Area "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan. We thank Michiaki Hamada and Yutaka Saito for fruitful discussions.

References

1. S. R. Eddy, *Nat Rev Genet* **2**, 919(Dec 2001).
2. J. S. Mattick and I. V. Makunin, *Hum Mol Genet* **15 Spec No 1**, R17(Apr 2006).

Pacific Symposium on Biocomputing 15:88-97(2010)

3. E. Rivas and S. R. Eddy, *Bioinformatics* **16**, 583(Jul 2000).
4. E. Rivas and S. R. Eddy, *BMC Bioinformatics* **2**, p. 8 (2001).
5. D. di Bernardo, T. Down and T. Hubbard, *Bioinformatics* **19**, 1606(Sep 2003).
6. A. Coventry, D. J. Kleitman and B. Berger, *Proc Natl Acad Sci U S A* **101**, 12102(Aug 2004).
7. S. Washietl, I. L. Hofacker and P. F. Stadler, *Proc Natl Acad Sci U S A* **102**, 2454(Feb 2005).
8. J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller and D. Haussler, *PLoS Comput Biol* **2**, p. e33(Apr 2006).
9. M. Hamada, H. Kiryu, K. Sato, T. Mituyama and K. Asai, *Bioinformatics* **25**, 465(Feb 2009).
10. T. Gesell and S. Washietl, *BMC Bioinformatics* **9**, p. 248 (2008).
11. I. L. Hofacker, *Nucleic Acids Res* **31**, 3429(Jul 2003).
12. I. L. Hofacker, M. Fekete and P. F. Stadler, *J Mol Biol* **319**, 1059(Jun 2002).
13. S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer and P. F. Stadler, *Nat Biotechnol* **23**, 1383(Nov 2005).
14. S. Washietl, J. S. Pedersen, J. O. Korbel, C. Stocsits, A. R. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigò, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker and P. F. Stadler, *Genome Res* **17**, 852(Jun 2007).
15. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler, *Genome Res* **12**, 996(Jun 2002).
16. M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler and W. Miller, *Genome Res* **14**, 708(Apr 2004).
17. A. X. Wang, W. L. Ruzzo and M. Tompa, *BMC Bioinformatics* **8**, p. 417 (2007).
18. C. B. Do, C.-S. Foo and S. Batzoglou, *Bioinformatics* **24**, i68(Jul 2008).
19. J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res* **22**, 4673(Nov 1994).
20. I. H. A.R. Gruber, Stephan H Bernhart and S. Washietl, *BMC Bioinformatics* **9**, p. 122 (2008).
21. J. S. McCaskill, *Biopolymers* **29**, 1105 (1990).
22. C. B. Do, D. A. Woods and S. Batzoglou, *Bioinformatics* **22**, e90(Jul 2006).
23. H. Kiryu, T. Kin and K. Asai, *Bioinformatics* **23**, 434(Feb 2007).
24. Y. Ding, C. Y. Chan and C. E. Lawrence, *RNA* **11**, 1157(Aug 2005).
25. L. E. Carvalho and C. E. Lawrence, *Proc Natl Acad Sci U S A* **105**, 3209(Mar 2008).
26. A. R. Gruber, S. H. Bernhart, I. L. Hofacker and S. Washietl, *BMC Bioinformatics* **9**, p. 122 (2008).
27. A. Wilm, I. Mainz and G. Steger, *Algorithms Mol Biol* **1**, p. 19 (2006).
28. S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy and A. Bateman, *Nucleic Acids Res* **33**, D121(Jan 2005).
29. S. Washietl and I. L. Hofacker, *J Mol Biol* **342**, 19(Sep 2004).
30. T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics* **21**, 3940(Oct 2005).
31. R. J. Klein and S. R. Eddy, *BMC Bioinformatics* **4**, p. 44(Sep 2003).
32. S. Bernhart, I. Hofacker, S. Will, A. Gruber and P. Stadler, *BMC Bioinformatics* **9**, p. 474(Nov 2008).
33. D. Sankoff, *SIAM Journal on Applied Mathematics* **45**, 810 (1985).