Journal of Marine Science and Technology

Volume 19 | Issue 3 Article 12

AN ALIGNMENT APPROACH TO IDENTIFY FISH WITH ENCODING HYDROACOUSTIC SIGNAL TO ACOUSTICS ALPHABET SQUENCES

Shih-Yen Ku
MIB Program, Institute of Statistical Science, Academia Sinica, 128, Academia Rd. Sec. 2, Taipei 115, Taiwan, R.O.C.

Hsueh-Jung Lu
Department of Environmental Biology & Fishery Science, National Taiwan Ocean University, 2 Pei-Ning Rd. Keelung 20224, Taiwan, R.O.C.., hjlu@mail.ntou.edu.tw

Chi-Ting Tseng
Department of Environmental Biology & Fishery Science, National Taiwan Ocean University, 2 Pei-Ning Rd. Keelung 20224, Taiwan, R.O.C..

Ker-Chau Li
MIB Program, Institute of Statistical Science, Academia Sinica, 128, Academia Rd. Sec. 2, Taipei 115, Taiwan, R.O.C.

Long-Jin Wu
Coastal Fishery Research Center, Taiwan Fishery Research Institute, Kaohsiung, Taiwan, R.O.C.

Recommended Citation
Ku, Shih-Yen; Lu, Hsueh-Jung; Tseng, Chi-Ting; Li, Ker-Chau; and Wu, Long-Jin (2011) "AN ALIGNMENT APPROACH TO IDENTIFY FISH WITH ENCODING HYDROACOUSTIC SIGNAL TO ACOUSTICS ALPHABET SQUENCES," *Journal of Marine Science and Technology*: Vol. 19: Iss. 3, Article 12.
DOI: 10.51400/2709-6998.2198
Available at: https://jmstt.ntou.edu.tw/journal/vol19/iss3/12

AN ALIGNMENT APPROACH TO IDENTIFY FISH WITH ENCODING HYDROACOUSTIC SIGNAL TO ACOUSTICS ALPHABET SQUENCES

Acknowledgements

The authors would like to thank Mr. Jinn-Shing Weng and anonymous crews of the research vessel of the Coastal Fishery Resource Research Center, Taiwan Fisheries Research Institute, for their help in at-sea tuna fish sampling. We are also grateful to the National Science Council of the republic of China (ROC) for financial support (NSC97-2625-M-019-002).

AN ALIGNMENT APPROACH TO IDENTIFY FISH WITH ENCODING HYDROACOUSTIC SIGNAL TO ACOUSTICS ALPHABET SQUENCES

Shih-Yen Ku*, Hsueh-Jung Lu**, Chi-Ting Tseng**, Ker-Chau Li*, and Long-Jin Wu***

Key words: target strength, echo sounder, tuna, 1D sequences alignment tool, alphabet design, Self Organizing Map algorithm.

ABSTRACT

1D sequence homologous alignment tool, like FastA (FASTALL) [8] or BLAST (Basic Local Alignment Search Tool) [1], has been widely used in bioinformatics field and perform elegant and fast searching for the sequences developed from the same kinds of species. In other word, it can classify through determining the homologous similarity which is not totally similar in sequences of protein sequences, structure or nucleotide sequences. An approach is proposed in this paper called AA-FAST (abbreviation for Acoustics Alphabet-FAST) which takes advantage of alignment tool and significant sequence encoding method. In this experiment, it could not only determine 4 fish species with similar size and shape but also the motion of them with identical alignment matrix. Besides, it shows that the position containing higher similarity encoding sequence fragment is related to the position of specific fish species and the acoustic features of specific fish species. Other purpose of this paper is to demonstrate how a bioinformatics tool could be applied to the acoustic field.

I. INTRODUCTION

Acoustical identification of fish species is a crucial problem when using quantitative echo sounders to estimate fish abundance and distribution, especially in the tropical and sub-tropical waters where multispecies often co-exist. To resolve the problem, many features from the echoes obtained during surveys were used for target identification such as target strength, school descriptors and multi-frequency echoes [10].

Target strength (TS), a logarithm measure of the proportion of the incident energy backscattered by the target, is the scaling factor for transferring energy of echo integration into abundance of marine organism during acoustic surveys [11]. Echo trace descriptors is generally based on TS information collected by the echo sounder because TS differs between species of different body size [9]. However, TS provides insufficient information on species identification, for example TS of 16 cm capelin and 40 cm Atlantic mackerel (no swim bladder) were similar [14], and the TS of an individual fish differed more than 30 dB when it swam in a different orientation [12]. Therefore, only using the value of TS from a single target is not enough for species identification. Other features from the echo of single target are necessary for further interpretation of TS information.

There should be some inherent features from the echo of different fish; however, to catch the fish's profile from the echo is not an easy work. Many studies on recognizing fishes focused on the relationship between the TS and the features of fishes. For example, Knudsen *et al.* [4] monitored TS of Atlantic salmon in a cage to demonstrate that fish shape is an important factor for TS. Thor *et al.* [15] and Didrikas *et al.* [2] also derived the formula between *in situ* TS and fish lengths for krill, herring and sprat. They hope to find the general equations between body length and TS, with final purpose of determining the general rules between the body length and the TS. However, the equations or correlations they developed were only for some specific type of fishes. It gives us the hint that the TS detected *in situ* or at sea might be noisy and makes determining the fish species very complicated.

In biological studies, the protein sequences and structures are also noisy and not easy to determine. However, there is an efficient and fast alignment tool developed, which could help us to determine the sequences homology in very short time. Besides, the protein sequences alignment tools could also use in structural alignment, there're so many fast alignment tools developed. Yang *et al.* [16] developed the 3D-BLAST (3D-

Paper submitted 10/18/10; revised 11/07/10; accepted 11/24/10. Author for correspondence: Hsueh-Jung Lu (e-mail: hjlu@mail.ntou.edu.tw).

***MIB Program, Institute of Statistical Science, Academia Sinica, 128, Academia Rd. Sec. 2, Taipei 115, Taiwan, R.O.C.*

***Department of Environmental Biology & Fishery Science, National Taiwan Ocean University, 2 Pei-Ning Rd. Keelung 20224, Taiwan, R.O.C..*

****Coastal Fishery Research Center, Taiwan Fishery Research Institute, Kaohsiung, Taiwan, R.O.C.*

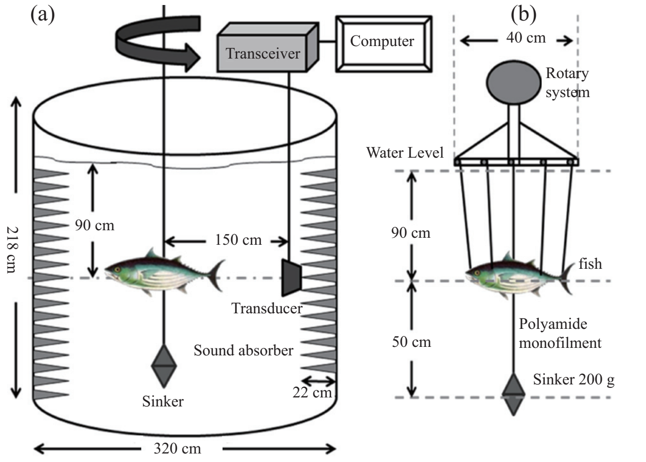


Fig. 1. Construction for TS measurement. (a) system set up for measuring the echoes from single target. (b) the suspension used to support the target.

Basic Local Alignment Searching Tool) to determine the structural similarity from target protein to all proteins in all databases. Lo *et al.* [8] also developed a tool SARST (Sequence Alignment on Ramachandron plot Searching Tool) to determine the similarity of proteins. Ku *et al.* [5] develop faster and easy-to-train pipeline to determine the similarity of proteins and the homology of proteins.

Like protein structural and homology analysis, there are so many uncertain factors that affect feature of the fish echo from single target, such as length, shape, orientation and etc. In this study, we will provide a combinatorial pipeline to determine some basic fish profiles from the acoustic datasets. We will apply Ku's work with a little modification to solve the problem.

II. PREPARING THE DATASETS

1. The Environment

In order to set up the environment and get the acoustic datasets from the sonar detector, we use a tank with sea water to get the datasets (Fig. 1). The diameter of the tank is 3.2 meters and the depth of it is about 5 meters. The transducer is set on a side of the tank. We put the fish at the middle of tank and control its motion with three cotton threads.

2. Preparing the Objects

Fig. 2 shows the pictures of the four fishes aggregated by anchored fishing aggregation device (FAD) in southwestern Taiwan, namely yellowfin tuna (*Thunnus albacore*), skipjack tuna *(Katsuwonus pelamis*), bullet tuna (*Auxis rochei rochei*) and rainbow runner (*Elagatis bipinnulate*). Parameters of body shape were provided in Table 1, including body weight, fork length, and circumference defined as the length surrounding the maximum cross-section. The reason why we choose these fishes is that their shapes are very similar and coexist around FAD.

Table 1. The information of fishes for the experiment.

Common name	Scientific name	Weight (g)	Fork length (cm)	Circumference (cm)
Bullet tuna	*Auxis thazard*	654	34.1	20.6
Skipjack tuna	*Katsuwonus pelamis*	1118	38.4	25.2
Rainbow runner	*Coryphaena hippurus*	1126	49.6	24.8
Yellowfin tuna	*Thunnus albacares*	1104	38.5	25.6

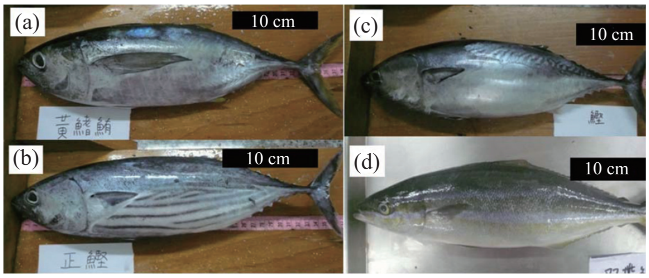


Fig. 2. Fishes used for target strength measurement experiment. (a) Yellowfin tuna (*Thunnus albacore*); (b) Skipjack tuna (*Katsuwonus pelamis*); (c) Bullet tuna (*Auxis rochei rochei*); (d) Rainbow runner (*Elagatis bipinnulate*).

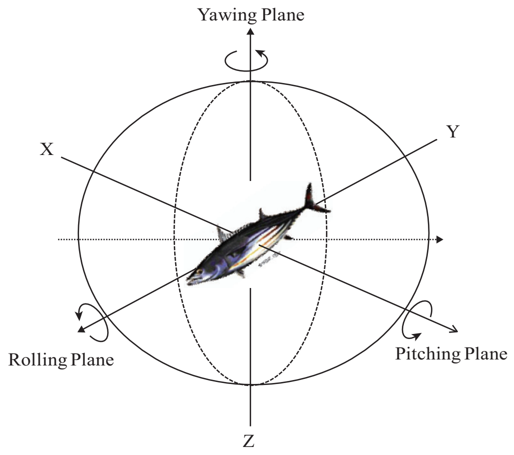
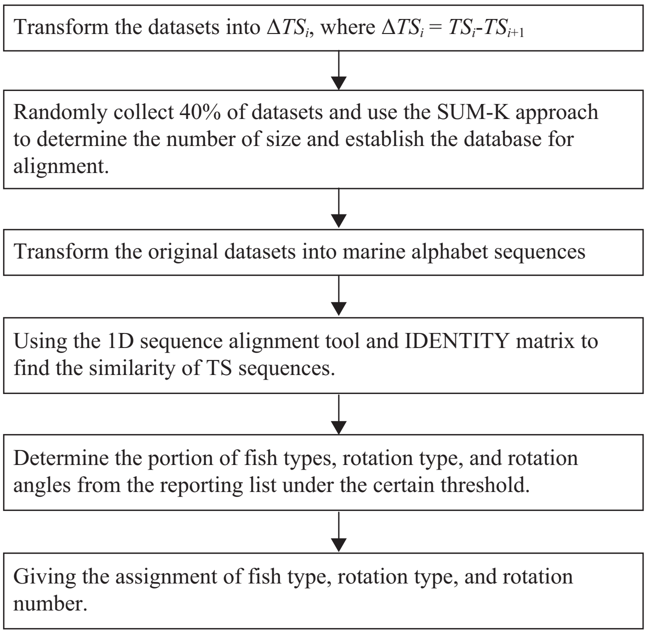


Fig. 3. Three-planes polar diagram of TS measurement of suspended fish versus direction of propagation of sound wave.

3. Setting the Motion of Objects and Training Datasets

To make the experiment more challenging, the aspects of fishes are considered in our experiments. The conceptual viewing of the simulated motion of the fish is demonstrated in Fig. 3. The fish are rotated in three different planes: XY yawing plane, YZ pitching plane, XZ rolling plane. There are 24 rotation angles with 15 degree between two rotation angles. The lasting time to collect the acoustic datasets is 30 seconds.



Transform the datasets into ΔTS_i, where $\Delta TS_i = TS_i - TS_{i+1}$

Randomly collect 40% of datasets and use the SUM-K approach to determine the number of size and establish the database for alignment.

Transform the original datasets into marine alphabet sequences

Using the 1D sequence alignment tool and IDENTITY matrix to find the similarity of TS sequences.

Determine the portion of fish types, rotation type, and rotation angles from the reporting list under the certain threshold.

Giving the assignment of fish type, rotation type, and rotation number.

Fig. 4-1. The system flow of determining the fish's profile.

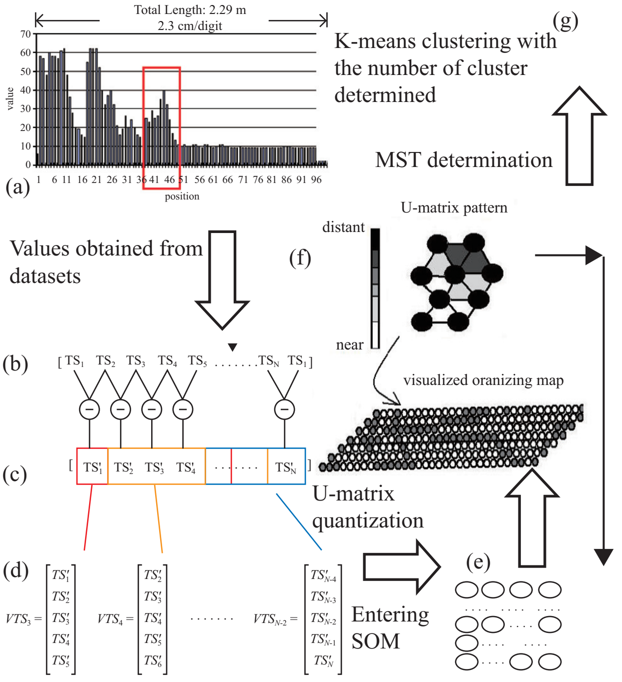


Fig. 4-2. The conceptual viewing of data extraction and SUM-K process. The datasets are from echoes of fish (a) with which the values are obtained (b) and the differences (c) are calculated. The vectors of each dataset (d) are put into the SOM process (e). With U-matrix quantization, the distribution of clusters is shown by SOM map (f) and then with BIC and Minimal Spanning Tree algorithm to determine the number of clusters. After several times of SUM-K process, the final number of clusters is obtained.

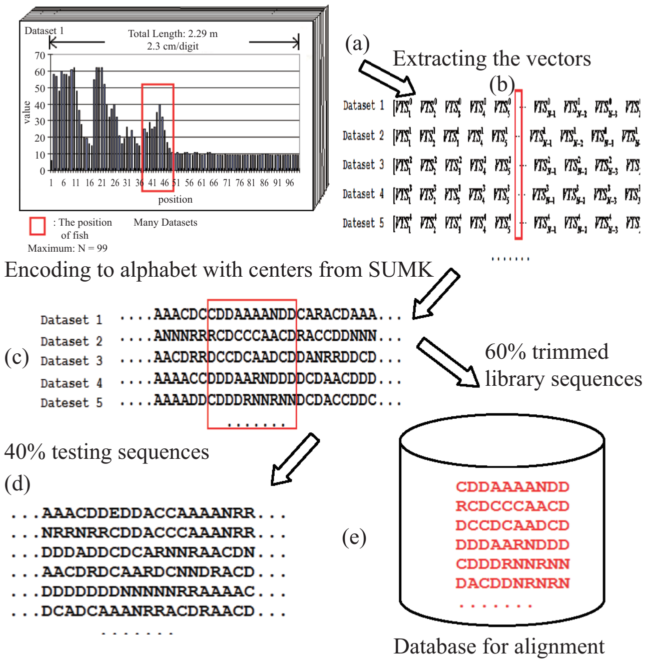

Dataset 1
Total Length: 2.29 m
2.3 cm/digit
value
position
: The position of fish
Many Datasets
Maximum: N = 99
(a)
Extracting the vectors
(b)
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Dataset 5
....
Encoding to alphabet with centers from SUMK
(c)
....AAACDCCDDAAAANDDCARACDAAA...
....ANNNRRRCDCCCAACDRACCDDNNN...
....AACDRRDCCDCAADCDDANRRDDCD...
....AAAACCDDDAARNDDDDCDAACDDD...
....AAAADDCDDDRNNRNNDCDACCDDC...
60% trimmed library sequences
40% testing sequences
(d)
...AAACDDEDDACCAAAANRR...
...NRRNRRCDDACCCAAANRR...
...DDDADDCDCARNNRAACDN...
...AACDRDCAARDCNNDRACD...
...DDDDDDDNNNNNRRAAAAC...
...DCADCAAANRRACDRAACD...
(e)
CDDAAAANDD
RCDCCCAACD
DCCDCAADCD
DDDAARNDDD
CDDDRNNRNN
DACDDNRNRN
Database for alignment

Fig. 4-3. The conceptual viewing of preparing datasets for sequence alignment is demonstrated. Once the transformed centers obtained from SUM-K process, the all original datasets (a) will be transformed to the 99 vectors (b). According to the centers obtained from SUM-K and nearest neighbor assignment, the datasets are transformed to several sequences (c). 60 percent of transformed sequences are trimmed with respect to the position of fish and saved in database for alignment (e) while 40 percent of transformed sequences are taken as the testing data (d). The red rectangle means the position of fish.

By this way, there are significant large datasets generated from the experiments. Totally 34560 acoustic sequences generated from 3 rotation planes × 24 different angles × 2 pings/s × 30 seconds × 2 fishes × 4 species. In our experiments, 40% of these dataset will be used as training datasets and it will be a significant large number for training.

III. PROCESSING APPROACH

1. Overview of Approach

Fig. 4-1 shows the the overview of system flow and Figs. 4-2 and 4-3 show the conceptual view of extracting the vectors from original datasets and determines the number of cluster by SUM-K approach, which is composed by Self-organizing map, U-matrix quantization, Minimal spanning tree, and K-means clustering. Firstly, we randomly picked up 40% of samples from the datasets. For example, if there're 100 sampled echoes under certain condition such like the skipjack and at 30 degrees of pitching angle, we will randomly take 40 samples as training datasets. For each sample in the training dataset, we calculate the difference (ΔTS_i) of echo level from TS_i and TS_{i+1} in ith position and take ΔTS_i as the new value for the ith position which is TS'. The next step is taking the VTS = $[TS'_{i-2}\ TS'_{i-1}\ TS'_{i}\ TS'_{i+1}\ TS'_{i+2}]^T$ as one vector for SUM-K process (See Fig. 4-2). There are N-4 vectors gener

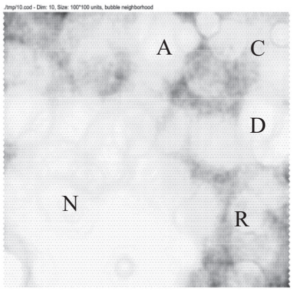


Fig. 5. A resulting SOM map and its related alphabet of echo datasets from SUM-K approach.

ated with sliding window scanning where the length of transformed datasets is N. Then, the SUM-K (Self-organizing map, U-matrix quantization, Minimal spanning tree, and K-means clustering) approach [5, 6] is applied to determine the number of clusters. The SUM-K approach is crucial in our pipeline, since it determines the significant number of clusters. For the first step of SUM-K is using self-organizing map to capture the distribution of vector. During the process of self-organizing map, the vectors tend to find the nearest position and organize a group on the map. In Fig. 5, the white region means the vectors are very close each other and the black region means the vectors are very different from the white vectors. Those vectors placed in the black regions are known as outliers to any white regions. There are five or more clusters to be recognized by eye in Fig. 5. However, it is not convinced to get the number of clusters from observing. Instead, the robust computational method should be taken.

In the second step, U-matrix quantization defines the distance and topology of each position on the map and describes how distant the two groups of vectors on the map is. After defining the distance, the minimal spanning tree algorithm is applied, which is trying to find the nearest distance from one point to all points in one topology and to assign the group of points in the topology to one tree with given criteria. The number of cluster is then determined based on minimal spanning tree algorithm. To carefully find the number of cluster, the process which contains Self Organizing Map, U-matrix, and Minimal Spanning Tree is repeated 200 times with various size of self organizing map and various thresholds for minimal spanning tree, and the BIC (Bayesian Inference Criterion) approach that described in Ku *et al.* [7] is also applied. Fig. 4-2 provides the conceptual viewing of SUM-K approach and the work flow of SUM-K. The purpose of BIC approach is finding the statistically significant number of clusters. Once the number of clusters is determined, the final central vectors of the training datasets were obtained by K-means clustering. The role of K-means clustering is to find the centers and convert the VTS_i to center which is close to VTS_i. For each center, an alphabet is assigned to represent the center. Besides, in order to do the 1D sequence alignment, all datasets will be transformed from the quantity value to the transformed alphabet sequences. For each sample, we will get a transformed sequence.

	A	**R**	**D**	**N**	**C**
A	4	-10	-10	-10	-10
R	-10	4	-10	-10	-10
D	-10	-10	4	-10	-10
N	-10	-10	-10	4	-10
C	-10	-10	-10	-10	4

Fig. 6. The IDENTITY matrix defined by the alignment tools.

2. Preparing the Sequences for Alignment

The key step for our tool is determining the number of clusters that can transform the value of echo levels to the sequences that could be used by the alignment tool. This step is quite like smoothing step that smooths the target strength to the center. Besides, the alphabet could be more flexibile to represent the patterns. Based on the procedure described in previous section, the number of clusters determined by the frequent number of clusters from SUM-K is 5. Since the SUM-K required threshold in the minimal spanning tree to determine the real number of clusters on SOM (Self-Organizing Map) featuring map, the threshold determined showed us there should be 5 clusters on the map with the BIC formula described in Ku *el al.* [7]. Fig. 5 shows how five clusters distributed on one SOM featuring map has been tested in our training process. Based on the centers of encoding alphabets, the original datasets are transformed to sequence. However, the transformed sequences are divided into two groups, 40% of the sequences are taken as testing datasets and 60% of the sequences are trimmed and keep the fragment that contained exactly the fish body. The overview of preparing process is shown as Fig. 4-3.

3. Parameters Used by Alignment

In this study, we only used the IDENTITY matrix (shown as Fig. 6) while we do the 1D sequences alignment. The IDENTITY matrix is not similar to the definition of identical matrix in linear algebra. The elements of IDENTITY matrix will give a positive score while the alphabets of column and row are matching and negative score while the alphabets of column and row are similar. Therefore, it determines the similar sequences to the target sequences, since the similar sequence will lead higher score.

The alignment tool also has its own statistical method measured by p-value. When the p-value is lower, it means the alignment tool will report sequences to the target sequences with higher statistical significance. We examined the lists of significant sequences for each target sequence under the p-value and compared with the target sequences. The statistical measure of the statistical approach is:

$$P(s > x) = 1 - e^{-e^{-\lambda(x-\mu)}} = 1 - e^{-kmne^{-\lambda x}}$$

where μ = ln(Kmn)/λ
K: the parameter fitting in the ungapped alignment
m: the length of the query sequence
n: the length of the library sequence
λ: frequency of aligned words

K is the parameter fitting in the ungapped alignment could be estimated from the alignment matrix and alignment scores, λ is the rate for exponential distribution and represents the frequency of aligned words. From the above formula, the p-value represents the probability of two similar words aligned for a given library and query sequence. Thus, the adjusted alignment score could be defined based on the above formula and in FastA the average score for an unrelated library sequence increases with the logarithm of the length of the library sequence [13]. With this statistics, the alignment score could be adjusted and more related library sequences will be found.

After using the target sequences searching, we will assign the feature according to the most frequent conditions. In addition to the assignment process, we reveal the portion of condition of the selected sequences and the alignment information between two sequences. By these information, we could determine the key feature of a certain condition. Also, we established the sample database for our alignment tool. The sample database was generated from the transformed sequences. According to the alignment results, we could assign the fish species, motion type and the degree of the rotation.

IV. IMPLEMENTATION

The approach is implemented by PERL and PHP program. PHP program provided the website interface and PERL program deal with the transforming the values of echo level for a single target to encode alphabet sequences. All encoded sequences are stored in the database which is supported by MYSQL. FASTA with IDENTITY matrix and k-mean algorithm could be obtained from the internet. Self Organizing Map toolkit could also be obtained from the Kohonan's website [15]. Therefore, our tool can be easily implemented.

V. EXPERIMENTS

1. Evaluate Quality with Existing Datasets

To evaluate the utility of AA-FAST for recognizing the similarity of a query acoustic sequence, we used 60% of previous datasets excluding the training datasets and tried to run the experiment of assignment and picked one acoustic sequence as a query sequence. Then, we do the similar search with AA-FAST alignment tool. There are three types of assignment experiment we did here. For the level I match, the fish species of hit is similar to the type of query one. For the level II match, the fish species and the rotation type of hit is similar to those of query sequences. Finally, we use these three matching level to demonstrate the power of our pipeline. Besides, we changed our threshold from 10^5 to 10^1 and see the precision of precision and recall rate. The quality of similarity searching is based on some common measurement, including the precision, recall, and F-score. In the i_{th} experiment, the precision is defined as A_h^i / A and recall is given as A_h^i / T_h^i, where A_h^i is the number of true hit acoustic datasets in the hit list, T_h^i is the total number of acoustic datasets in the databases, and A means the total number of acoustic dataset. Since the experiments performed N times for one classification, for example the skipjack and at 45 degree of pitching angle, the average precision and recall are defined as $\sum_{i=1}^{N}(A_h^i / T_h^i)/N$ and $\sum_{i=1}^{N}(A_h^i / A)/N$.

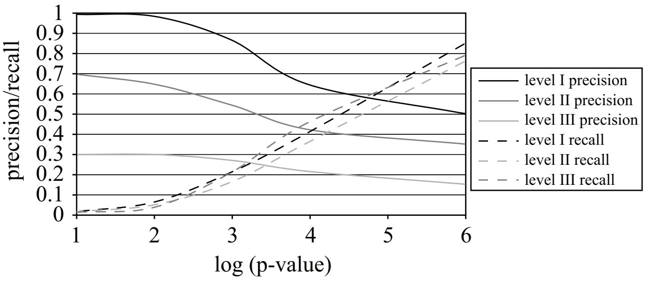


Fig. 7. Precision and recall chart demonstrates the performance of the tool AA-FAST applied in three different matching levels.

Then, the F-score is calculated with the formula: $F - score$ = (2 × *precision* × *recall*)/(*precision* + *recall*).

For acoustic similarity searching, our tool provides the cutoff value to identify the similarity of acoustic dataset with the query dataset. When a lower e-value is used, the portion of true positive is increasing for similarity searching (Fig. 7).

Table 2 shows the relationship among the e-value, precision, recall and F-score under three different matching levels. For the acoustic database searching in level I matching, the precision is 0.87 and the recall is 0.2 when the cutoff value is 10^3. If the cutoff value is 10^1, the precision is 0.99 and the recall is 0.05. Even we loose the cutoff value, the precision is 0.5 and the recall is 0.84. It means that we could find all the similar acoustic datasets and keep a certain precision. For the level II searching, the precision is decreased from 0.99 to 0.65 but the recall is 0.025 when the cutoff value is 10^1. However, for the level III matching, the precision is 0.3 and the recall is 0.015. These results show that we could distinguish the fish species from the acoustic datasets and our tool still works to distinguish the fish species and motion types from the acoustic under the IDENTITY matrix. However, when determining the degree of rotation under certain motion and certain fish species, our tool performed not as well as determined the fish species.

Moreover, our tool could fast distinguish the motion and fish species without deriving any formula like previous work. Besides, it might give us some hint of the pattern that distinguishes from one fish to others.

Table 2. The precision and recall of matches under three different matching levels.

Level I matches			
P-value	Precision	Recall	F-score
10^5	0.50	0.84	0.626
10^3	0.65	0.40	0.493
10^2	0.87	0.20	0.325
10^1	0.99	0.05	0.095
Level II matches			
10^5	0.35	0.75	0.477
10^3	0.42	0.35	0.382
10^2	0.54	0.15	0.235
10^1	0.65	0.035	0.066
Level III matches			
10^5	0.15	0.78	0.251
10^3	0.21	0.45	0.290
10^2	0.27	0.20	0.229
10^1	0.30	0.025	0.046

2. Evaluate with New Datasets

Because in previous evaluation, we use the 60% of datasets to verify our precision, repeat the experiment and collect the acoustic data from our training datasets. It's reasonable that our result is pleasing due to the training datasets and test datasets based on the same resources. In order to verify our tools still working and solving even more challenging problems, we start to apply our method to new datasets with new conditions.

For the new datasets, we obtained new datasets for yellowfin tuna (*Thunnus albacares*) of which average weight is 561 grams. The average size of yellowfin tuna is almost half of previous yellowfin tuna we used to train our tools. For the temperature of environment is also quite different. We obtained the datasets in the winter and the average temperature is 18.7 degree Celsius. However, the average temperature of new datasets is 29.2 degree Celsius . We used these datasets to prove that our tools could determine the fish species and motion when the size of fishes and the temperature is changed.

To know the standard of evaluating results, we use the most frequently appearing fish species in the hit list as our answer when we do the assignment experiments. Like previous experiments, we also give the assignment of three different levels described before. The number of new datasets is 3518. The assignment rates is T_c/T and precision is A_t/A_h which is defined as same as the previous measurement where T is the total number of testing datasets, Tc is the number of correct assignment, At is the true hit acoustic datasets on the hit list, and Ah is the total number of hit list. We add the running time to demonstrate the speed of our searching tool. The environment is Pentium IV 3.2 G personal computer. Besides, we search about 92,160 sequences at one time.

Table 3. The result of new testing datasets.

Match type	Assignment rates	Precision	Running time (s)
Level I	0.90	0.75	1.23
Level II	0.72	0.59	1.21
Level III	0.31	0.22	1.24

To test our datasets, we use different cutoff value in our experiment. For the level I match, we use 10^4; for the level II match, we use 10^5. We use the 10^3 for level III matching. Beside, we also calculate the average running time to measure the speed of our searching tools. The average running time defined as $(\sum_{i=1}^{T} t_i)/T$, where t_i is the running time in ith run, T is the number of total datasets.

The result of new testing datasets is shown in Table 3. It required 1.225 seconds to scan the entire database and do the assignment of fish species. For fish species matching, the tool could perform well since the assignment rate is 0.9 and the precision is 0.75 which means there's only 25 percent incorrect results in the hit list. For the new datasets detected by different environment, the tool still works well in matching the fish species and motion type of fish with a correct assignment rate 0.72. For matching the fish species, motion, and degree of rotation angle, the precision and assignment rate is 0.31 and 0.22. In conclusion, the experiment shows that we could classify the fish species and motion type well within a running time around 1 second.

3. Finding Sharing Patterns

Besides verifying the novel and original datasets, we also apply a simple multiple alignment of the assignment results. We use the members of hit list which match the fish species, defined as the level I matching. We select the five members of the hit list ranging from the first to 2000th of hit lists. Each member is different in rank 200. Here we pick the five members to do the alignment and find the similar parts of sequences.

From the Figs. 8(a), (b), (c), and (d), we could find the sequence pattern of each type of fishes. These patterns mean the target strength pattern of acoustic datasets and also represent the basic acoustic feature of different fishes. For the yellowfin tuna is ANNDDD-DD-D-RRRAADDDA-DDDDD; for the rainbow runner, the pattern of sequence is DND-ANNDDD-DD-DD-DNANAA-D-D-ADDDDD; for the bullet tuna's multiple alignment result, the pattern we observed is DDNCA NDRDDNND-D-RA-D-ADDDDD; for the skipjack tuna, the pattern is AND-DD-D-A-DDD-AD-A-DD-DDDD-D. From the above pattern, we could find the common pattern of rainbow runner and bullet tuna is ADDDDD and the common pattern of rainbow runner and yellowfin tuna is ANNDDD. The most representative pattern from the simple alignment for bullet tuna is DDNCANDRDDNND. For the rainbow runner, there's the representative pattern, which is DNANAAD. For

```
 10       20        30        40         50        60
|--------|---------|----------|----------|----------|
NNRDNANNDDDDRDDRADDNDDNNRRRAADDDARDDDDDNNNNNNNNNN
                                                   Rank
NRNDNANNDDDDRDDNNRDDRDNNRRRAADDDAADDDDDN    2000
DNNNDANNDDDDADDNADDNRDNDRRRAADDDAADDDDDD    1500
NNNNNANNDDDNNDDARDDDNNNNRRRAADDDARDDDDDN    1000
DNRDNANNDDDNNDDRADDNDDDNRRRAADDDARDDDDDN    500
NRRDNANNDDDNNDDRADDNDDDNRRRAADDDARDDDDDN    3
|---------|---------|---------|---------|
         10        20        30        40
```

(a)

```
         10       20        30        40         50
|--------|---------|----------|----------|----------| 60
NDNNNANNDDDNRDDNRDDRDDRDNANAADDDAADDDDDNNNNNNNNNNN
                                                   Rank
RDNDRANNDDDDRDDDANDDDDRDNANAADDDDADDDDDN    2000
NDNDAANNDDDNRDDDNNNRDDNDNANAADDDAADDDDDD    1750
ADNDAANNDDDDADDNNCDRDDRDNANAADNDAADDDDDN    1500
RDNDAANNDDDNRDDDNRDCDDNDNANAADDDAADDDDDN    1000
RDNDAANNDDDNRDDDRDDADDNDNANAADDDNADDDDDN    500
RDNDRANNDDDDRDDNRDDCDDNDNANAADDDAADDDDDN    3
|---------|---------|---------|---------|
          10        20        30        40
```

(b)

```
10       20        30        40         50        60
|--------|---------|----------|----------|----------|
DDDNCANDRDDNNDDRNNADDNDDNADRANDDRADDDDDNNNNNNNNNN
                                                  Rank
NDDNCANDRDDNNDDDAADDDDNNDNNRADDDRADDDDDD    2000
DDDNCANDRDDNNDNNDNADNNDNDADRADDDRADDDDDD    1750
DDDNCANDRDDNNDDDCCADDDDNDADRAADDAADDDDDD    1500
DDDNCANDRDDNNDDNDNADDNDNNNRRADNDRADDDDDD    1000
DDDNCANDRDDNNDDNDNADDNDNNNRRADDDRADDDDDD    500
NDDNCANDRDDNNDDNNNADDNDDNADRANDDNADDDDDN    3
|---------|---------|---------|---------|
          10        20        30        40
```

(c)

```
       10       20        30        40         50        60
|--------|---------|----------|----------|----------|
RDDNAANDDDDCDDNAADDDDNDCDADRAADDRADDDDDDNNNNNNNNNN
                                                   Rank
NNDDAANDDDDCDDNNAADDDNDRDADRADDDARDDDDDD    2000
DDDNCANDRDDNNDDNAADDDDNNDADRANDDRADDDDDD    1750
RDDRRANDDDDCDDNRARDDDNNCDADRADDDAADDDDDD    1500
DDRDDANDDDDCNDDAADDDDNDNDADCAADDRADDDDND    1000
NDNRAANDDDDCDDNNANDDDDDNDADRAADDRADDDDDD    500
NNDDAANDDDDCDDNAADDDDDDRDADRAADDRADDDDDD    3
|---------|---------|---------|---------|
         10        20        30        40
```

(d)

Fig. 8. Multiple sequence alignment results for (a) yellowfin tuna, (b) rainbow runner, (c) bullet tuna and (d) skipjack tuna. The underlined alphabets mean the similar patterns that different sequences are shared and The position 10 to 40 units is where the fish is located. From the figures, the associated pattern could be recognized and these common patterns are shared, even though the ranking number of similarity is high. It means same type of fishes with different postures may share the common associative alphabet patterns.

the yellowfin rainbow runner and yellowfin tuna is ANNDDD. The most representative pattern from the simple alignment for bullet tuna is DDNCANDRDDNND. For the rainbow runner, there's the representative pattern, which is DNANAAD. For the yellowfin tuna, the representative pattern is RRRAADDDA. For the skipjack tuna, the pattern should be associated with little fragment, which is DD-D-A-DDD-AD. For these patterns, we could establish the classification rule to determine the fish species.

```
Query (species: Thunnus albacares;
       motion: yawing)
            40         50
        -----|----------|----------
      Query  NNNNARDDANDDDDND
                                    Rank
      Hit 1  DDNRAADDDDDDADDD  10
      Hit 2  DDNRAADDDDDDADDN  20
      Hit 3  DNDCANNDDDRNDDDD  50
      Hit 4  DDNRAADDDDDDADDA  80
      Hit 5  DDNRAANDDDDDADDD  100
        -----|----------|------
             10         20
```

Fig. 9. Multiple alignments of acoustic sequences with using new testing dataset as a query.

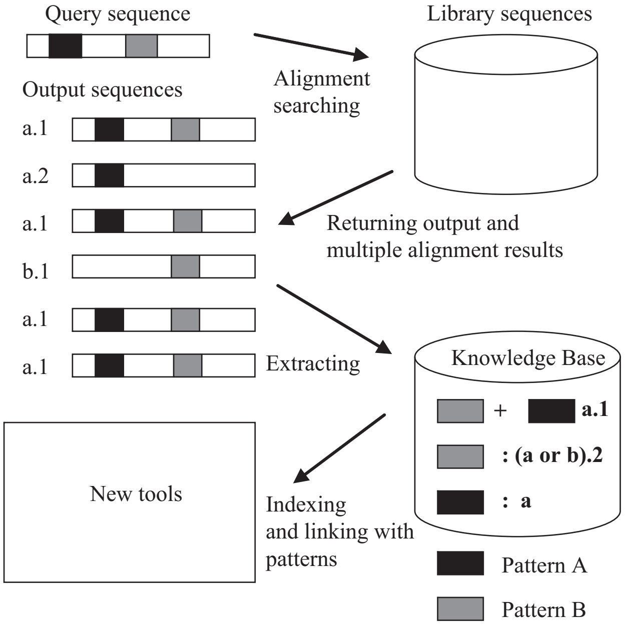


Fig. 10. The conceptual viewing of developing new tools for scanning.

For the new datasets, we could see the general pattern aligned by our current databases and determined the yellowfin tuna's representative pattern that we described before. The underlined alphabets in Fig. 9 indicate the representative region in the query and subject sequences. There are similar alphabets in the same region. The other fragment in the region is variable, but in the same position, it still keeps in the similar alphabet. It means that even if the fish size and water temperature changed, the characteristic of fish would not change. Therefore, the alignment tool developed could determine the fish species even if the variable region may be caused by the temperature and size. However, in order to make our approach more significant we hope to gather more different fish species to enrich our databases and let our alignment tool to recognize various species of fish.

4. Developing Even Fast Approach for Assignment

Fig. 10 demonstrates the conceptual view of developing faster approach. In this figure, we named conditions as a.1, b.2, a.2 which is not actually appeared in our real datasets. One interesting thing happened in this experiment is that after using the transformed alignment searching from the fish echo datasets, we will assign the query sequence of condition a.1 according to the most frequent conditions happened in aligned sequences a.1 in the figure. For the part of the multiple alignments in these conditions, the basic patterns in certain size of fish could be caught and shown as black boxes (Pattern A) and gray boxes (Pattern B) in the figure. Moreover, this information could also characterize the fragment as knowledge base of sequence which has the similar condition. Besides, we will make the link between these patterns and conditions, for example Pattern A plus Pattern B means a.1 condition. This link will provide us to develop faster approach to distinguish the profile of fishes.

VI. CONCLUSION

The alignment tool developed in this study could correctly identify fish species up to 90% in four species coexistence circumstance in a very short time. However, the datasets used in the study were from dead fish under control, which may be different from live fish swimming in open sea. Before the method can be more practically used in fishery survey, live fish experiment is necessary.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Jinn-Shing Weng and anonymous crews of the research vessel of the Coastal Fishery Resource Research Center, Taiwan Fisheries Research Institute, for their help in at-sea tuna fish sampling. We are also grateful to the National Science Council of the republic of China (ROC) for financial support (NSC97-2625-M-019-002).

REFERENCES

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. "Basic local alignment search tool," *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403-410 (1990).
2. Didrikas, T. and Hansson, S., "In situ target strength of the Baltic Sea herring and sprat," *ICES Journal of Marine Science*, Vol. 61, pp. 378-382 (2004).
3. Iivarinen, J., Kohonen, T., Kangas, J., and Kaski, S., "Visualizing the clusters on the self-organizing map," *Proceedings of the Conference on Artificial Intelligence Research*, Finland, pp. 122-126 (1994).
4. Knudsen, F. R., Fosseidengen, J. E., Oppedal, F., Karlsen, Ø., and Ona E., "Hydroacoustic monitoring of fish in sea cages: target strength (TS) measurements on Atlantic salmon (*Salmo salar*)," *Fisheries Research*, Vol. 69, pp. 205-209 (2004).
5. Ku, S. Y. and Hu, Y. J., "A multi-strategy approach to protein structural alphabet design," The 2006 International Conference on Bioinformatics & Computational Biology, Las Vegas, USA, pp. 285-291 (2006).
6. Ku, S. Y., and Hu, Y. J., "Using protein structural alphabet to characterize local structure features," The 2008 International Conference on Bioinformatics & Computational Biology, Las Vegas, USA, pp. 692-697 (2008).
7. Ku, S. Y. and Hu, Y. J., "Protein structure search and local structure characterization," *BMC Bioinformatics,* Vol. 9, p. 349 (2008).
8. Lo, W. C., Huang, P. J., Chang, C. H., and Lyu, P.-C., "Protein structural similarity search by Ramachandran codes," *BMC Bioinformatics*, Vol. 8, p. 307 (2007)
9. Love, R. H., "A model for estimating distribution of fish school target strengths," *Deep Sea Research*, Vol. 28A, pp. 705-725 (1981).
10. Lu, H. J. and Lee, K. T. "Species identification of fish shoals from echograms by an Echo-signal Image Processing System," *Fisheries Research*, Vol. 24, pp. 99-111 (1995).
11. Maclennan, D. N. and Simmond, E. J., *Fisheries Acoustics*, Chapman & Hall, London, p. 325 (1993).
12. Mukai, T., Sano, N., Iida, K., and Sasaki, S., "Studies on dorsal aspect target strength of ten species of fish collected in the South China Sea," *Nippon Suisan Gakkaishi*, Vol. 59, No. 9, pp. 1515-1525 (1993).
13. Pearson, W. R. and Lipman, D. J., "Improved tools for biological sequence comparison," *Proceedings of the National Academy Sciences of the United States of America*, Vol. 85, No. 8, pp. 2444-2448 (1998).
14. Rose, G. A. and Leggett, W. C., "Hydroacoustic signal classification of fish schools by species," *Canadian Journal of Fisheries and Aquatic Sciences*, Vol. 45, pp. 579-604 (1988).
15. Thor, A. K. and Stein, K., "In situ target strength and behaviour of northern krill (*Meganyctiphanes norvegica*)," *ICES Journal of Marine Science*, Vol. 63, pp. 1726-1735 (2006).
16. Yang, J. M. and Tung, C. H., "Protein structure database search and evolutionary classification," *Nucleic Acids Research*, Vol. 34, No. 13, pp. 3646-3659 (2006).
17. Zhang, Z., Pearson, W. R., and Miller, W., "Aligning a DNA sequence with a protein sequence," *Journal of Computational Biology*, Vol. 4, No. 3, pp. 4339-4349 (1997).