JOURNAL OF
ECOLOGY AND ENVIRONMENT
http://www.jecoenv.org

Research Paper
J. Ecol. Environ. 38(1): 95-107, 2015

Subsequent application of self-organizing map and hidden Markov models infer community states of stream benthic macroinvertebrates

Dong-Hwan Kim[1], Tuyen Van Nguyen [2], Muyoung Heo[2] and Tae-Soo Chon[1,*]

[1]Department of Integrated Biological Science, Pusan National University, Busan 609-735, Korea
[2]Department of Physics, Pusan National University, Busan 609-735, Korea

Abstract

Because an ecological community consists of diverse species that vary nonlinearly with environmental variability, its dynamics are complex and difficult to analyze. To investigate temporal variations of benthic macroinvertebrate community, we used the community data that were collected at the sampling site in Baenae Stream near Busan, Korea, which is a clean stream with minimum pollution, from July 2006 to July 2013. First, we used a self-organizing map (SOM) to heuristically derive the states that characterizes the biotic condition of the benthic macroinvertebrate communities in forms of time series data. Next, we applied the hidden Markov model (HMM) to fine-tune the states objectively and to obtain the transition probabilities between the states and the emission probabilities that show the connection of the states with observable events such as the number of species, the diversity measured by Shannon entropy, and the biological water quality index (BMWP). While the number of species apparently addressed the state of the community, the diversity reflected the state changes after the HMM training along with seasonal variations in cyclic manners. The BMWP showed clear characterization of events that correspond to the different states based on the emission probabilities. The environmental factors such as temperature and precipitation also indicated the seasonal and cyclic changes according to the HMM. Though the usage of the HMM alone can guarantee the convergence of the training or the precision of the derived states based on field data in this study, the derivation of the states by the SOM that followed the fine-tuning by the HMM well elucidated the states of the community and could serve as an alternative reference system to reveal the ecological structures in stream communities.

Key words: ecological assessment, emission probability matrix, event sequence, Markov processes, temporal dynamics, transition probability matrix

INTRODUCTION

Ecological processes are not easily observable due to the complexity in responding to numerous environmental conditions. In many cases of ecological research, observed patterns are considered as the result of underlying ecological states embedded in complex ecological processes with different governing rules than simple consequences of the observations. The data are also highly variable due to noise, and redundancy, internal relations and outliers are often observed (Gauch 1982, Jongman et al. 1995). Conventional indices such as Bray-Curtis similarity have been proposed to reveal the temporal dynamics of stream community (Scarsbrook 2002, Collier 2008). In

http://dx.doi.org/10.5141/ecoenv.2015.010

Received 14 January 2015, **Accepted** 28 January 2015

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial Licens (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

***Corresponding Author**
E-mail: tschon@pusan.ac.kr
Tel: +82-51-510-2261

Copyright © 2015 The Ecological Society of Korea. All rights are reserved.

pISSN: 2287-8327 eISSN: 2288-1220

addition multivariate analyses also have been applied to extraction of high dimensional temporal data to obtain ecological integrity in reduced dimension (Legendre and Legendre 1998, Scarsbrook 2002, Beche and Resh 2007). For the former method, however, the parameters are constrained in a sense that community dynamics is only expressed by a single term, whereas the latter method could reveal statistical or overall patterns in complex data, being insufficient in addressing ecological processes objectively.

Markov processes were suitable in addressing complex time series changes depending on time, location, and local or global state frequency (e.g., characterizing clean or pollution states) (Tucker and Anand 2003) in plant (Horn 1975, Yemshanov and Perera 2002) and animal (Usher 1981, Hill et al. 2004) ecology. Markov models that implement Markov processes have been developed in two different manners: a stationary Markov model (SMM) and a hidden Markov model (HMM). Since 1960 (Baum and Petrie 1966, Baum 1972), the HMMs have arisen for overcoming the limitation of the SMM (monotonic increases or decreases), because they are more adaptable to complex situations (Rabiner 1989, Visser et al. 2002). The HMMs are described as "hidden" because the sequence of observation events (e.g., sequences of ecological data) is the result of a stochastic process operating on top of a sequence of hidden states generated by a Markov process (Rabiner 1989). Based on stochastic processes applied to observable events and states concurrently, the discrete states are more clearly addressed in the HMM than in the SMM.

The HMM could be suitably applied to ecological processes and allows the description of the complexity including interacting hierarchical systems and hidden processes underlying the observed ecological dynamics (Tucker and Anand 2004). HMMs are also flexible and relatively easy to implement because efficient algorithms have been developed to estimate the model parameters (Ghahramani and Jordan 1997). The HMM has been feasible for analyzing animal behaviors (MacDonald and Raubenheimer 1995, Liu et al. 2011), and the wide biological disciplines including DNA and protein sequencing analysis (Krogh et al. 1994, Karplus et al. 1997). Although HMMs have been widely applied to ecology, there have only been a few attempts to analyze community dynamics, especially for benthic macroinvertebrates in streams. The HMM was utilized to address vegetation community dynamics in remote sensing (Viovy and Saint 1994), change-points in grassland vegetation (Ver Hoef and Cressie 1997), and coastal wetland processes (Dale et al. 2002).

Prior to applying the HMM, the candidates for states were obtained heuristically by applying a self-organizing map (SOM) to the field data in this study. Based on SOM clustering, transition probabilities were estimated between states according to varying time series data. The SOM is an unsupervised neural network and has been implemented in various ecological studies (Adriaenssens et al. 2007, Chon 2011). The SOM has been a feasible means to address changes in temporal patterns (Voegtlin and Dominey 2001, Simon et al. 2007) including benthic macroinvertebrates (Chon et al. 2000) and fishes (Hyun et al. 2005).

The current study is aimed to finely define the states in temporal dynamics of benthic macroinvertebrate communities with minimum pollution by the HMM based on the initial data heuristically obtained from the SOM. Based on taxonomic diversities, sedentariness in survival range, and long life cycles, benthic macroinvertebrates characteristically respond to the changes of environment from watershed areas (Resh and Rosenberg 1984, Park et al. 2007, Tang et al. 2010). Although the community of benthic macroinvertebrates presents ecological integrity fairly well, community data are highly complex and difficult to analyze because communities consist of numerous species varying in a complex and stochastic manner in response to environmental factors (e.g., high precipitation) in monsoon areas in East Asia (Chon et al. 2001).

In this study, we hypothesized that the temporal changes of the community could be stably expressed in the natural stream environment of monsoon areas. Partial information on community states based on field data was obtained by the SOM to provide an initial transition probability matrix (TPMs) and emission probability matrix (EPMs) for the HMM. Subsequently, the HMM was conducted according to the observable event sequences including biological (e.g., number of species and diversity) and environmental (temperature and precipitation) parameters. Community states preserved in benthic macroinvertebrates in streams were addressed accordingly by responding to natural variability to present structure property residing in community changes more objectively.

MATERIALS AND METHODS

Field sampling

The Nakdong River is the longest river flowing through

the Busan-Daegu Metropolitan areas in the southern part of the Korean peninsula. A sampling site, BCN (coordinates, 35°31′07.72″ N, 128°01′01.97″ E; altitude, 398 m), located in the Baenae Stream, a tributary of the Nakdong River, in the mountainous area near Busan (Tang et al. 2010) was selected for the study. BCN is considered as a reference stream with the minimal environmental impact (BOD, 1.1 mg/L; conductivity, 24.6 μS/cm) for the national Long-Term Ecological Research (LTER) project in Korea. Since 2005, BCN has been a sample site for LTER supported by the Korea's Ministry of Environment.

Sampling was conducted monthly by the Surber sampler (30 × 30 cm, 100 μm mesh) in 5 replications at a sampling site from July 2006 to July 2013. The environmental factors such as temperature and conductivity at sampling site were additionally measured (Table 1) by YSI 30 conductivity-salinity meter (Yellow Springs Instruments, Yellow Springs, OH, USA). The specimens were mostly identified to species or to the lowest possible taxonomical level by following the procedure described by Merritt and Cummins (1996), Brigham et al. (1982), Pennak (1978) and Yoon (1995) for general taxa, Brigham et al. (1982) and Brinkhurst (1986) for Oligochaeta, and Wiggins (1996) for Trichoptera. Chironomidae, however, was not identified at the species level due to the difficulty of classification. After classification of the specimens, species diversity was determined according to the Shannon index (Shannon and Weaver 1949). Biological indices, the number of species in EPT (Ephemeroptera, Plecoptera, and Trichoptera) taxa, biological monitoring working party (BMWP) and average score per taxon (ASPT) (Hawkes 1998), were also obtained to assess the water quality. The average and standard error of parameters in different seasons were presented in Table 1.

Self-organizing map

In the SOM network, the output layer consisted of computation nodes (*j*) in low dimensions (conveniently 2) for presenting the multi-dimensional input data in a comprehensive manner (Park et al. 2003). The vector x_i is considered to be an input layer to the SOM. In the network, each computation node, *j,* is connected to each node, *i,* of the input layer. The connectivity is represented as the weights, $w_{ij}(t)$, adaptively changing in each iteration of calculation, *t*. Each neuron of the network computes the summed distance between the weights and the distance $d_j(t)$ at output node *j,* and the network is calculated as shown below.

$$d_j(t) = \sum_{i=0}^{S-1} (x_i - w_{ij}(t))^2 \quad [1]$$

In this study the data matrix for the SOM consisted of 77 sample units (monthly collection from July 2006 to July 2013) and 69 variables (total number of species collected during the survey period). In order to reduce the great difference of numerical values in densities for SOM training, the input data, i.e., densities (individual/m^2) plus one individual, were transformed by common logarithm.

After training, the Ward's linkage method (Ward 1963) was applied to the weights of the SOM to cluster the patterned nodes. The initialization and training processes followed suggestions by the SOM Toolbox to allow optimization in algorithm (Vesanto et al. 2000). To evaluate the map quality, the quantization error for the resolution and the topographical error for the topology preservation were used to indicate the accuracy of mapping (Céréghino and Park 2008, Kohonen 2001, Park et al. 2003). More details in training and clustering were performed according to Park et al. (2003) under the MATLAB 6.1 environment (The Mathworks, Inc., Natick, MA, USA).

Table 1. Biological indices and environmental parameters, represented as mean (±SE), at BCN in different seasons

Season	*n*	No. of species	Shannon Diversity	Dominance	EPT%	BMWP	ASPT	Conductivity (μs/cm)	Precipitation (mm/month)	Temperature (°C)
Spring	18	26.8 (±1.3)	2.5 (±0.2)^{ab}	65 (±3.3)^{ab}	76.4 (±1.3)	112.5 (±4.3, 88.4-143.6)	7.7 (±0.1, 7.1-8.3)	24.4 (±1.3)	87.5 (±11)^{a}	12.9 (±1.2)^{b}
Summer	23	25.7 (±1.5)	2.2 (±0.1)^{a}	72.3 (±2.6)^{a}	69.8 (±1.5)	110.1 (±4.9, 66.8-147.6)	7.6 (±0.1, 6.7-8.4)	24.1 (±0.9)	217.9 (±30.3)^{b}	24.8 (±0.5)^{a}
Fall	19	23.2 (±2)	2.6 (±0.1)^{ab}	64.5 (±2.7)^{ab}	73.1 (±2.4)	99.5 (±6.7, 42.3-142.7)	7.6 (±0.2, 6.0-8.5)	24.5 (±1.3)	70.7 (±19.5)^{a}	16.1 (±1.3)^{b}
Winter	17	26.1 (±1.5)	2.8 (±0.2)^{b}	56.7 (±4)^{b}	74.6 (±1.7)	114 (±4.2, 42.3-142.7)	7.6 (±0.1, 7.0-8.0)	25.7 (±1.7)	18.3 (±4.1)^{c}	1.3 (±0.5)^{c}

The different alphabets indicate significant differences among different seasons according to the Tukey test ($P < 0.05$).

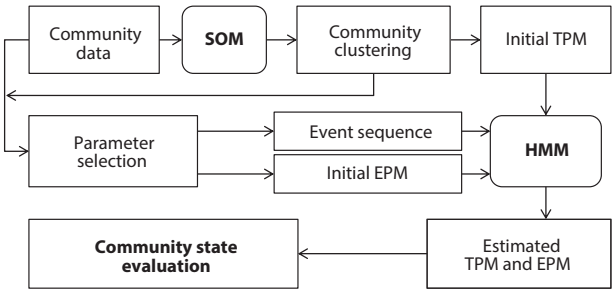

Fig. 1. Flowchart for estimating benthic macroinvertebrate community states in streams based on the self-organizing map (SOM) and hidden Markov model (HMM).

Hidden Markov model

Markov chain calculates the transition probabilities between community states from one time step to the next (Wootton 2001b, Tuker and Anand 2004). When the system has a finite set of discrete states (the number of states was estimated by SOM), these transition probabilities can be presented in a transition probability matrix where a_{ij} represents the transition probability from state i at time t to state j at time t+1 (Winston 1994). We computed the transition probability matrix using community states initially defined by the SOM (Fig. 1).

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \quad [2]$$

An HMM can be characterized by five elements (Rabiner 1989):

(1) A number of states N, $x \in \{1, \ldots, N\}$;

(2) A number of events K, $k \in \{1, \ldots, K\}$;

(3) Initial state probabilities, $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \le i \le$ N;

(4) State-transition probabilities, $A = \{a_{ij}\} = \{P(x_t = j \mid x_{t-1} = i)\}$ for $1 \le i, j \le N$;

(5) Discrete output probabilities, B = $\{b_i(k)\} = \{P(o_t = k \mid x_t = i)\}$ for $1 \le i \le$ N and $1 \le k \le K$,

where in (5), $P(o_t = k \mid x_t = i)$ indicates the probability of observation o_t at time t to be the event k on the condition that the state variable x_t at time t is equal to i. With the HMM formulation, transitions between the states can be estimated from the transition probability matrix (i.e., TPM) and the emission probability matrix (B) (i.e., EPM).

The transition probability matrix according to the SOM was used as initial TPM for HMM (Fig. 1). Transition probabilities across different states based on the SOM were accumulated and average transition probabilities between different community states were obtained. Subsequently we selected the biological indices (number of species, diversity and BMWP) and environmental factors (temperature and precipitation) as the observation events, because these parameters could represent the states of community with simple measurements (Fig. 1). Each parameter was divided into three levels based on the difference in SOM clustering analysis. Then three levels were considered as different categories defining each event (Table 2). Temporal dynamics of each event in the monthly scale were considered as event sequences. The initial EPM for HMM was further obtained by calculating probabilities of events corresponding to states given by the SOM in time series data.

Initial TPMs, EPMs and event sequences were also given to HMM as input data for training. Subsequently TPM and EPM were estimated according to the Baum-Welch algorithm for each event parameter (Juang and Rabiner 1991, Rabiner 1989, Durbin et al. 1998) (Fig. 1). The algorithm halted when the matrices in two successive iterations were within a tolerance value (0.1) for the discrete HMM (Taheri et al. 2005). If the algorithm failed to reach this tolerance within a maximum number of iterations, the default value for termination was 100 iterations. Based on the solutions presented in Rabiner (1989), the process was conducted with the programs provided in the HMM toolbox in MATLAB 2009 (The Mathworks, Inc.).

Table 2. Category of biological and environmental parameters in three levels

Parameters	Abbreviation	Category		
		1	2	3
Number of species	ES	0≤…<21.0	21.0≤…<30.0	30.0≤
Shannon Diversity	ED	0≤…<2.3	2.3≤…<2.8	2.8≤
BMWP	EB	0≤…<100.0	100.0≤…<120.0	120.0≤
Temperature (°C)	ET	0≤…<9.0	9.0≤…<22.0	22.0≤
Precipitation (mm/month)	EP	0≤…<35.0	35.0≤…<100.0	100.0≤

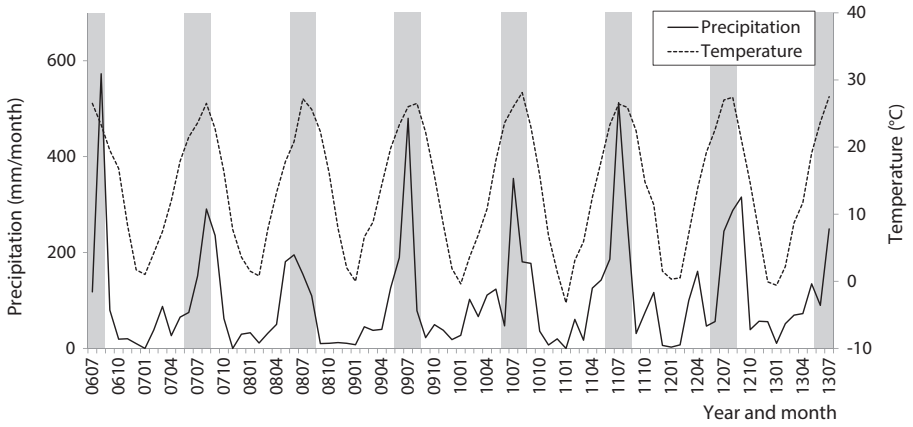

Fig. 2. Monthly changes in precipitation (mm/month; solid line) and temperature (°C; dotted line) at the sampling site, BCN, Gyungsang Province, Korea from July 2006 to 2013. The grey bars indicate rainy season. The first two digits on *x*-axis stand for the year of sampling after 2000, and the second two digits represent months for collection (e.g., 0607, July in 2006).

RESULTS

Environmental conditions

While sampling the community data, we also recorded monthly changes in precipitation and temperature as presented in Fig. 2. Those environmental factors varied at the sampling site according to the seasonal monsoon climate in Korea: high temperature (24.8°C) and precipitation (217.9 mm/month) in summer, and low temperature (1.3°C) and rainfall (18.3 mm/month) in winter (Fig. 2 and Table 1). The environmental factors showed distinct and unique features in summer and winter, while those in spring and fall were somewhat similar. Diversity was substantially low with high dominance in the summer. Biological water quality indices, however, EPT% (the percent EPT species in community), BMWP and ASPT, were relatively constant in seasonal changes (Table 1).

Patterning by the self-organizing map

To reveal patterns underlying the community data, we initially analyzed the monthly data of community abundance by training the SOM (Fig. 3a). In Fig. 3b, the clusters were divided into two distinct groups on the SOM based on the Ward's linkage method: two clusters of 1 and 2 in the upper part and other two clusters of 3 and 4 in the bottom part of the map. Cluster 1 contained the majority samples in summer and fall samples (Fig. 3a). In cluster 2, winter samples dominated. Spring and some other winter samples were spread over all parts of the SOM, with a weak bias toward the bottom part of the SOM (clusters 3 and 4).

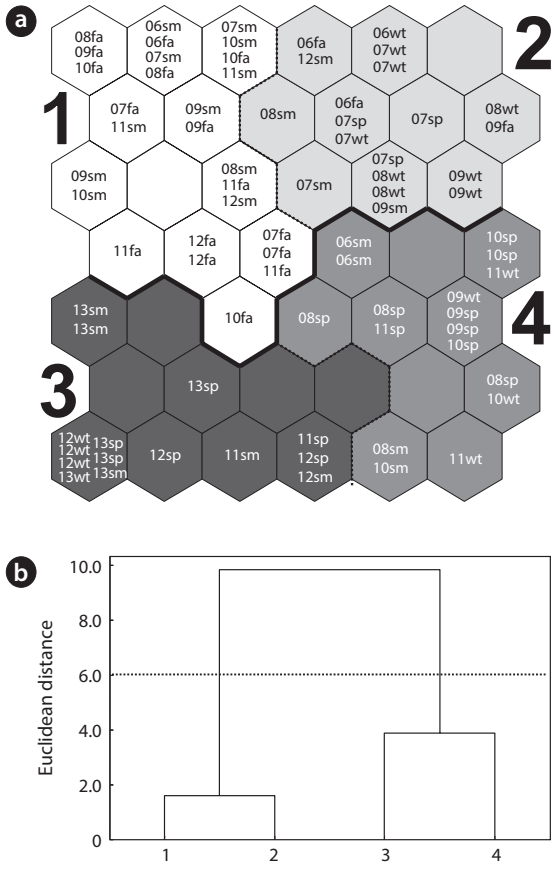

Fig. 3. Clustering by self-organizing map (SOM) on benthic macroinvertebrate communities collected at BCN. (a) SOM map with 4 clusters and (b) dendrogram based on the Ward's linkage method. The first two digits in (a) stand for the name of sample unit each node stand for the year of sampling after 2000 and the second two digits represent seasons (spring, sp; summer, sm; fall, fa; winter, wt; e.g., 06sm, summer in 2006).

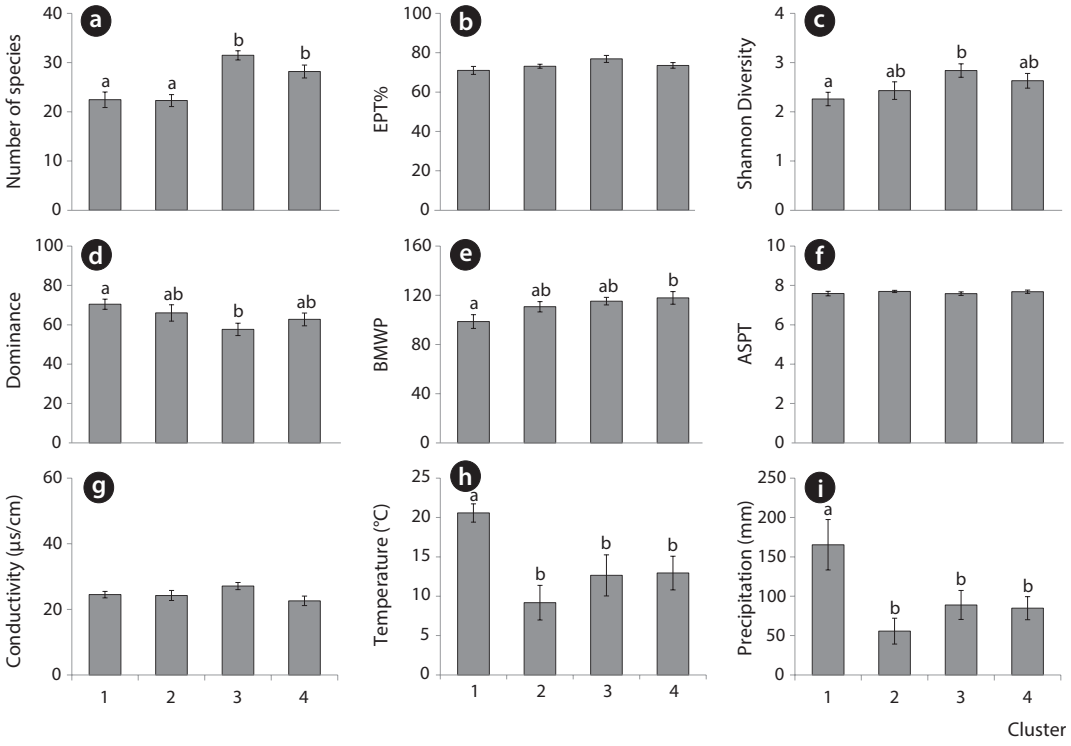

Fig. 4. Biological and environmental parameters matching different clusters on the self- organizing map (SOM) (as Fig. 3a). (a) Number of species, (b) EPT%, (c) Shannon diversity, (d) dominance, (e) the biological water quality index (BMWP), (f) average score per taxon (ASPT), (g) conductivity, (h) precipitation, and (i) temperature. Vertical bars and lines indicate average and standard error, respectively. Different alphabets present significance among different clusters according to the Tukey test ($P < 0.05$).

Fig. 4 shows the average values of biological indices and environmental parameters for each cluster. The number of species was significantly high in clusters 3 and 4 in the lower area of the SOM along the vertical gradient (Fig. 4a) (ANOVA, $P < 0.05$). Low diversity and high dominance were associated with cluster 1 and vice versa in cluster 3 (Fig. 4c and 4d). BMWP values were significantly high in cluster 4, which presented the spring season, but they were low in cluster 1, which matched summer and fall seasons (Fig. 4e). EPT% and ASPT, however, were not differentiated between clusters (Fig. 4b and 4f).

In particular, the environmental factors could characterize each area in the SOM. Temperature and precipitation were significantly high in cluster 1 consisting of the community data in summer and fall on the upper group of the SOM (Fig. 4h and 4i). Communities in cluster 2 presented low precipitation and temperature that characterized winter samples although statistical significance was not shown here. Conductivities were invariably low across clusters (Fig. 4g). Overall, communities in cluster 1 that were characterized with low diversity were influenced by high precipitation and temperatures, whereas communities in cluster 3 that were featured with high diversity and low dominance were affected by mid precipitation and temperatures.

Initial transition probability matrix

We hypothesized that patterns of the community data clustered by SOM (i.e., extraction from field data) could be useful to provide initial information for estimating the states of the community. These patterns could be provided to HMM as initial state transition probabilities, assuming that the clusters derived by the SOM training could represent the states of community in the field condition accordingly (Fig. 1). After SOM recognition, the probabili-

Table 3. Initial transition probability matrix (TPM) between different states (S) at time *t* and *t*+1 according to the self-organizing map (SOM)

		t+1			
		S1	**S2**	**S3**	**S4**
t	S1	0.667	0.148	0.037	0.148
	S2	0.222	0.611	0.000	0.167
	S3	0.071	0.071	0.857	0.000
	S4	0.176	0.118	0.118	0.588

ties to transit from one state (cluster) to the other state (cluster) were computed from the time series data. The obtained initial TPM is shown in Table 3. High remaining probabilities in the same states were observed for all the states (diagonal elements in Table 3) and it confirmed the robustness of probabilities. State 3 (S3) showed the highest remaining probability (0.857) whereas relatively low remaining probability was observed in state 4 (S4), showing 0.588. The probabilities transiting between different states have low values that ranged from 0.1 to 0.2. It was noteworthy that the S3 showed exceptionally low transition probabilities less than 0.1 except transition from S3 to S3 (S4→S3). Asymmetric transition probabilities were observed in relation to S3, for instance S3→S2 (0.072) and S4→S3 (0.118) (Table 3).

Estimated transition probability matrices and emission probability matrices

The biological parameters presenting the structure of the community (the number of species and the diversity) and the water quality (BMWP) were also obtained from the community data as observable events for HMM (see the "Hidden Markov model" section in Materials and Methods) (Table 2). The time series data were used as input event sequence to train the HMM. We also categorized each event depending on its value (i.e., low, medium, and high) (Table 2). The emission probability matrix (EPM) was determined by calculating the probabilities of all the events assigned to each state in time series data (Fig. 1). The initial EPM is shown in Table 4 for the selected parameters, the number of species, the diversity, and the BMWP. The TPMs and the EPMs were accordingly estimated through the HMM (Fig. 1) for different event variables (Rabiner 1989) (Table 4). The sum of probabilities for each matrix was normalized to unity; the sum of the probabilities in each row was set to 1.0 in total. Initial and estimated TPMs for three events were statistically not different (Table 4) based on the Kolmogorov-Smirnov tests (Quach et al. 2013) ($P > 0.05$). According to the paired t-test, the initial and estimated matrices were also in a similar range in Tables 3 and 4 (TPM: $P = 0.999$, 0.999, and

Table 4. Estimated TPMs and initial and estimated EPMs for biological and environmental parameters

No. of species

Estimated TPM

		t+1			
		S1	S2	S3	S4
t	S1	0.945	0.055	0	0
	S2	0	0.946	0	0.054
	S3	0	0	1	0
	S4	0	0	0.045	0.955

Initial EPM

		event		
		ES1	ES2	ES3
state	S′1	0.519	0.222	0.259
	S′2	0.444	0.500	0.056
	S′3	0	0.333	0.667
	S′4	0.176	0.471	0.353

		event		
		ES1	ES2	ES3
state	S1	0.616	0.348	0.036
	S2	0.423	0.424	0.153
	S3	0	0.348	0.652
	S4	0.292	0.339	0.369

Diversity

Estimated TPM

		t+1			
		S1	S2	S3	S4
t	S1	0.729	0	0	0.271
	S2	0.537	0.463	0	0
	S3	0	0.414	0.586	0
	S4	0	0	0.585	0.415

Initial EPM

		event		
		ED1	ED2	ED3
state	S′1	0.444	0.333	0.222
	S′2	0.278	0.333	0.389
	S′3	0.200	0.133	0.667
	S′4	0.353	0.176	0.471

		event		
		ED1	ED2	ED3
state	S1	0.802	0.095	0.103
	S2	0.168	0.521	0.311
	S3	0	0	1
	S4	0	0.698	0.302

BMWP

Estimated TPM

		t+1			
		S1	S2	S3	S4
t	S1	0.365	0	0	0.635
	S2	0.125	0.875	0	0
	S3	0	0	1	0
	S4	0	0.231	0.231	0.538

Initial EPM

		event		
		EB1	EB2	EB3
state	S′1	0.519	0.222	0.259
	S′2	0.167	0.611	0.222
	S′3	0.200	0.333	0.467
	S′4	0.235	0.235	0.529

		event		
		EB1	EB2	EB3
state	S1	1	0	0
	S2	0	1	0
	S3	0.296	0.288	0.416
	S4	0.769	0	0.231

"S" on column indicates the state at time t and "S" on row indicates the state at time *t*+1; S with apostrophe (′) indicates the initial state.

0.999 with the number of species, the diversity, and the BMWP, respectively, and EPM: P = 0.999, 0.999, and 0.999 with the number of species, the diversity, and the BMWP, respectively).

The estimated TPMs for the number of species had a strong tendency to show high probabilities to remain in the same state (Table 4). All probabilities in the diagonal elements were close to unity, ranging from 0.945 to 1.000. The estimated probability to remain in S4 remarkably increased to 0.955 from the initial transition probability of 0.588 (Table 3 and 4). The estimated EPM also characterized well the states of the community that were determined by the SOM (Table 4). In state 1, the frequency of low species number increased, ES1 (0.616). This indicated that state 1 comparatively depicted the stressful conditions of communities with low species number. State 3 showed the highest frequency with high species number, ES3 (0.652). Other states of 2 and 4 were characterized by mixed categories showing evenly high probability with low and/or intermediate species number for state 2, and evenly high probability with intermediate and/or high species number for state 4 (Table 4). The event characterization according to estimated EPMs for different states was also in accordance with the results from the SOM and field experience (Fig. 4).

The estimated transition probability for the diversity appeared lower diagonal elements than that for the number of species (Table 4). The probabilities for S2→S1 (0.537) and S4→S3 (0.585) were higher than those to remain in the same states S2 (0.463) and S4 (0.415), respectively. The estimated EPM could also characterize the community states (Table 4). In general, the emission probabilities seemed similar to the number of species. However, event categories were more strongly correlated with the states. State 1 corresponded to low diversity, ED1 (0.802), with high emission probability. The highest emission probability of 1.0 was associated with high diversity, ED3, at state 3. The emission probabilities corresponding to the diversity were differentiated more clearly (i.e., larger distance between maximum and minimum probabilities) than the number of species (Table 4). The estimated TPMs based on the BMWP showed the states were reluctant to remain in the same states except S3 (1.0) and had tendency of transition to other states more, especially S1→S4 (0.635) (Table 4). The estimated emission probabilities, however, were well separated, including 1.0 for the case of S1:EB1 and S2:EB2. Comparing with the estimated EPM, state 3 presented higher tendency to the high BMWP (EB3) than state 4.

Differences between the initial and the estimated values of the TPMs (left panel) and the EPMs (right panel) are further shown in Fig. 5a-5c. Training with the number of species substantially increased the probabilities to remain in the same states (i.e., the thick black rings in Fig. 5a). The diversity, however, resulted in the substantial reduction in the probabilities to remain in the same states (i.e., the white rings in Fig. 5b). For the transition probabilities to other states (the off-diagonal elements in the TPMs), only the diversity showed distinct changes with cyclic pattern.

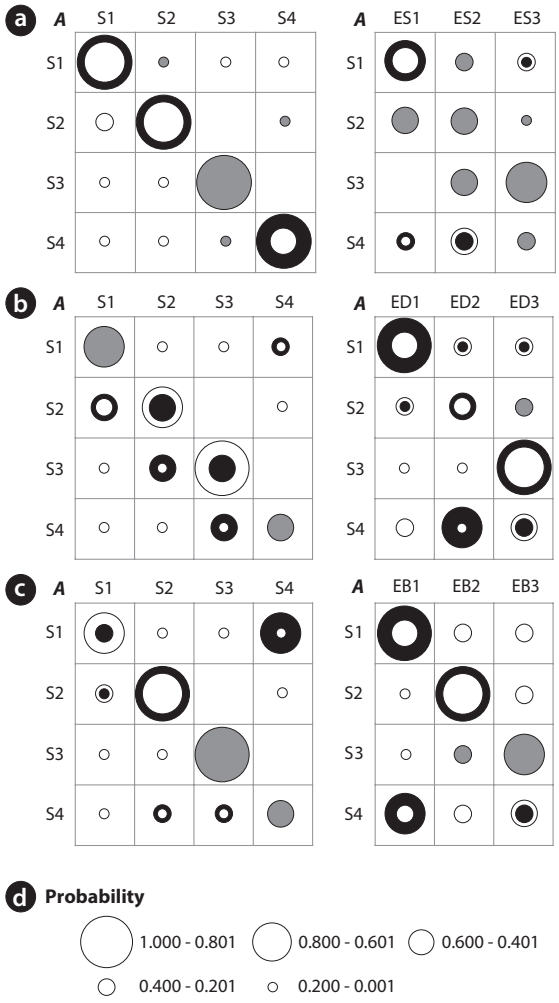

Fig. 5. Comparison of the initial and the estimated values of the TPMs (left panel) and the EPMs (right panel) based on the HMM. In TPM, "S" on column indicates the state at time t and "S" on row indicates the state at time t+1. The sizes of circles mark the probabilities. The white circle indicates the initial probability whereas black circle stands for estimated probability. If white ring surrounds the black circle, the estimated probabilities are lower than initial probabilities, whereas the black ring surrounds the white circle in case the estimated probabilities are higher. Grey circles present the same level of probability criteria between initial and estimated values. (a) estimated values of TPMs (left panel) and EPMs (right panel) of No. of species, (b) Shannon diversity, (c) the biological water quality index (BMWP), and (d) probability criteria.

Consequently it can be conjectured that the probabilities of state changes were more strongly expressed with diversity whereas discreteness in self-transition probabilities were more clearly addressed in the case of the number of species (Fig. 5). The EPMs also accordingly characterized the states as shown in Figure 5. Regarding the number of species, state 3 matched high number of species (ES3) whereas state 1 corresponded to low number of species (ES1). States 2 and 4 showed a mixed abundance of low-intermediate (ES1, ES2) and intermediate-high (ES2, ES3) levels of the number of species, respectively. For the diversity, the estimated probabilities substantially increased for S1:ED1, S3:ED3, and especially S4:ED2 (Fig. 5b). It means that the states of 1, 4 and 3 were likely to have the low (ED1), the middle (ED2) and the high diversity (ED3), respectively. State 2 was evenly associated with intermediate and high diversity (Fig. 5b), and was considered as a connection between two states S1 (low diversity) and S3 (high diversity).

We further checked if the water quality data that was empirically measured (i.e., BMWP) could preserve the transition probabilities based on the HMM training (Fig. 5c). A substantial change was observed regarding S1 in the estimated TPM. Probability for S1→S4 was markedly increased, and transition probabilities regarding S4→S3, S4→S2 and S2→S1, were also observed although the values were relatively low in TPM for BMWP. State 4 tended to be both state 2 (the middle value of BMWP) and state 3 (the high value of BMWP), and it indicated that state 4 was a link between the states of the low and high BMWP. The estimated EPM based on BMWP showed a similar trend to the case of the diversity. The maximum emission probabilities clearly appeared in all states, which showed the distinct association between the water quality and the states of the community (Fig. 5c). For instance, state 1 showed higher emission probability to the low BMWP than to the low diversity. However, state 2 was exceptionally characterized by the middle value of BMWP at EB2 whereas middle value of the diversity was shown at state 4 with ED2 (Fig. 5). Consequently, the EPM elaborated by training with the BMWP was more clearly characterized with high frequency of EB1, EB2 and EB3 for S1, S2 and S3, respectively. S4 was split in matching two event categories, EB1 followed by EB3. This demonstrated that the empirically determined BMWP could complement changes of community state that were partially observed in fields. The overall results suggested that the diverse aspects in states could be revealed according to different biological parameters.

DISCUSSION

To diagnose the integrity and the healthy condition of the communities from the field data, we developed a novel method to indicate the temporal states of communities by combining two computational methods of the SOM and HMM. By using the heuristic SOM as input data, stochastic processes residing in community development processes were accordingly elucidated to demonstrate a structure property of community. In most studies, the initial transition probabilities were given as either uniform or random distribution to train the HMM by assuming that the states were totally unknown. However, this approach does not guarantee to achieve the convergence to obtain the estimated TPMs in this study. It was mainly originated by the short and partial sequences of the field data (77 sequences for monthly sampling in this study). This study demonstrated how to combine the short information obtained from field data by applying the SOM and the HMM consecutively to address ecological processes observed in benthic macroinvertebrate in field condition in depth. First, by clustering the SOM, we could combine many pieces of information that describe the different aspects of the community and derive the state of the community. Second, by applying the HMM, we could find the impact of the different pieces of information on each state objectively and the strong consistency among different parameters as well (Fig. 1). We also showed that the inferred time series data of the states could depict the seasonal dynamics of the community in detail (Figs. 4 and 5).

This paper showed that the different aspects of the community states could be concurrently measured and analyzed with respect to different observable events such as the number of species, the diversity, and the BMWP. While the number of species was distinctively specialized in defining states (discreteness in community states from the high diagonal elements in TPM), the diversity showed the cyclic development of communities (Fig. 5 and Table 4). Remarkably, the estimated TPM derived by the diversity could efficiently address the seasonal changes of states. According to the temperature for each state shown in Figure 4h based on the SOM, we could find that state 1 and 2 represent the summer and winter seasons, respectively. The development of community with cyclic pattern were presented, S1 (the low diversity in summer)→S4 (the variable diversity in transition)→S3 (the highest diversity in transition)→S2 (the variable diversity in winter)→S1 (Fig. 5b and Table 4). Two states of S2 and S4 associated with intermediate diversity were presented as a connection between high diversity and low diversity. The former

one indicated "before" high diversity while the latter one showed "after" high diversity condition. In other words, the increasing and decreasing trends of diversity in temporal dynamics of communities were efficiently revealed based on HMM.

The observable events (i.e., the number of species and the diversity) could further present different aspects of community states consequently. The reason why the number of species is efficient in defining state discreteness while diversity is more addressable in presenting state transition is currently unknown. One factor for consideration would be the data type. The number of species mainly delivers the presence or absence information of species and similar values tended to be observed in the same seasons. Consequently, the difference between the numbers of species could be more clearly associated with defining discreteness in states. Diversity, however, presented entropy of communities and more continuous values were obtained compared with the number of species. According to the definition of Shannon diversity index, abundance of all species contributed to determining diversity values, resulting in more continuous values in community samples. This continuity and variability observed in diversity may be more feasible in addressing transition to different states. However, the detailed mechanism is unknown currently and more tests are required with additional field data in the future.

BMWP also delivered meaningful information through TPM and EPM in characterizing community states (Fig. 5c). The empirically determined water quality indices were indeed capable of characterizing community states by showing enhanced values of EPM (i.e., higher maximum values) and lower minimum values (Fig. 5c). BMWP is effective in revealing states of benthic macroinvertebrates in streams with minimum pollution and can be used as a reference system for evaluating biological indices in response to environmental variability. In this study, BMWP was checked for defining community states instead of ASPT. BMWP appeared to be better at differentiating polluted states in the weakly polluted condition because it contained more precise information regarding the richness of pollution-sensitive species in the family level, whereas ASPT (average of BMWP) tended to be more representative of the overall tolerance in pollution impact. Testing of more water quality indices is required under different impacts of anthropogenic disturbance for biomonitoring in the future. To the best of authors' knowledge, state definition of communities based on by SOM and HMM has not been reported regarding biological parameters.

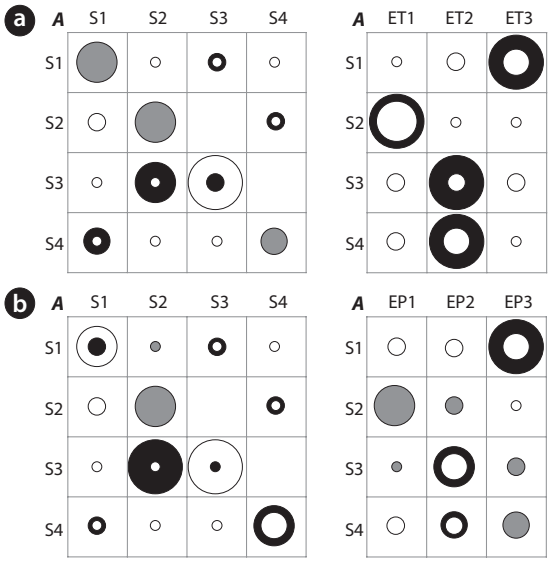

Fig. 6. Comparison of initial and estimated TPMs and EPMs according to (a) temperature and (b) precipitation. Details of the graphs are explained in Fig. 5.

We further tested if the parameters in HMM could be still obtained when the environmental factors such as temperature and precipitation were used as events. Fig. 6 showed the TPMs and the EPMs when the HMMs were trained with environmental factors. Although the state was determined initially according to ecological community data by the SOM, the TPMs were still estimated accordingly based on environmental factors and comparable with those obtained according to biological parameters. The initial and the estimated values of TPMs and EPMs by the environmental events also fell into a similar range according to the Kolmogorov-Smirnov tests ($P > 0.05$) and paired t-test (Tables 3 and 4) (TPM: $P = 0.999$ and $P = 0.999$ with temperature and precipitation, respectively, and EPM: $P = 0.999$ and $P = 1.000$ with temperature and precipitation).

The transition probabilities in the diagonal were high and the corresponding EPMs were accordingly characterized. However, the difference was also observed when compared with the biological parameters. Regarding temperature, the probability to remain in state 3 was not as high in the estimated TPM as in the initial TPM (Fig. 6). However, the transition probability of S3 to S2 increased correspondingly and it suggested that state 3 tend to be state 2 with strong tendency. The EPM was also different according to each state. State 1 was presented by high temperature and precipitation (large black circle at S1:ET3 and S1:EP3), matching summer (Fig. 4). The frequency of low temperatures was the most dominant at S2

and the intermediate temperatures were more abundantly observed in S3. Event probabilities for S4 were similar to S3 with intermediate temperatures (Fig. 6). The TPM of precipitation was also similar to that of temperature. Along with EPMs and information based on the SOM, two TPMs from the environmental parameters showed the state transition patterns clearly (Fig. 6). Cyclic pattern was observed; S1 (high temperature and precipitation)→S3 (intermediate temperature and precipitation)→S2 (low temperature and precipitation)→S4 (intermediate temperature and precipitation)→S1. It is noteworthy that the state changes that were originally based on community patterning (Fig. 3) could be still preserved when the states were presented by environmental changes. However, the cyclic pattern of precipitation was rather diffusive and not clearly observed; S1→S2 or S3, S3→S2→S4→S1.

Overall, the HMM based on temperature and precipitation put more emphasis on the seasonality. The cyclic pattern based on the diversity appeared as S1→S4→S3→S2→S1. But the cyclic pattern according to the environmental factors was observed as S1→S3→S2→S4→S1, which was more directly related to the seasonality. This could be useful for predicting and detecting the seasonal dynamics in stream ecosystems including summer flooding and winter drought along with corresponding community data.

In this study, Chironomidae was not included for analysis due to difficulty of taxonomic identification at species level. Overall, however, the Chironomidae community would similarity reflect seasonal changes, providing a similar trend in community states in general. However, it should be also noted that Chironomidae would be feasible in presenting in niche division in species abundance distribution (Tang et al. 2010), and an extra aspect may be revealed regarding community organization. Further research is warranted in this regard in the future.

In conclusion, the HMMs were feasible in elucidating the dynamics of the community states in combination with the SOMs that provide initial data for HMM. Initially the community abundance data were heuristically classified by the SOM. Along with initial TPM and EPM, field data for observable events were provided to HMM. In case that the field data were relatively short, the partial information obtained from the field data could be useful for addressing the states of the benthic macroinvertebrate communities in streams in more details. The states were identified efficiently to explain the status communities in temporal dynamics in streams with minimal pollution. The number of species described each state more clearly whereas the diversity reflected the changing states of the cyclic community development. The environmental factors such as temperature and precipitation could also reveal the seasonal cyclic changes of the communities. Overall, the biological and environmental parameters were also well characterized by EPMs. The SOMs and HMMs were efficiently combined to address community state changes in temporal scale, and the combined model could be a reference system to reveal ecological processes in stream communities.

ACKNOWLEDGMENTS

This work was supported for two years by Pusan National University Research Grant.

LITERATURE CITED

Adriaenssens V, Verdonschot PFM, Goethals PLM, De Pauw N. 2007. Application of clustering techniques for the characterization of macroinvertebrate communities to support river restoration management. Aquat Ecol 41: 387-398.

Baum LE. 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3: 1-8.

Baum LE, Petrie T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37: 1554-1563.

Beche LA, Resh VH. 2007. Short-term climatic trends affect the temporal variability of macroinvertebrates in California 'Mediterranean' streams. Freshw Biol 52: 2317-2339.

Brigham AR, Brigham WU, Gnilka A. 1982. Aquatic Insects and Oligochaetes of North and South Carolina. Midwest Aquatic Enterprises, Mahomet, IL

Brinkhurst RO. 1986. Guide to the Freshwater Aquatic Microdrile Oligochaetes of North America. Department of Fisheries and Oceans, Ottawa.

Céréghino R, Park YS. 2009. Review of self-organizing Map (SOM) approach in water resources: commentary. Environ Model Softw 24: 945-947.

Collier KJ. 2008. Temporal patterns in the stability, persistence and condition of stream macroinvertebrate communities: relationships with catchment land-use and regional climate. Freshw Biol 53: 603-616.

Chon TS, Park YS, Park JH. 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecol Model 132: 151-166.

Chon TS. 2011. Self-Organizing Maps applied to ecological sciences. Ecol Inform 6: 50-61.

Chon TS, Kwak IS, Park YS, Kim TH, Kim Y. 2001. Patterning and short-term predictions of benthic macroinvertebrate community dynamics by using a recurrent artificial neural network. Ecol Model 146: 181-193.

Dale MB, Dale PER, Li C, Biswas G. 2002. Assessing impacts of small perturbations using a model-based approach. Ecol Model 156: 185-199.

Durbin R, Eddy S, Krogh A, Mitchison G. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids . Cambridge University Press, Cambridge.

Gauch HG Jr. 1982. Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge.

Ghahramani Z, Jordan MI. 1997. Factorial hidden Markov models. Mach Learn 29: 245-275.

Hawkes HA. 1998. Origin and development of the biological monitoring working party score system. Water Res 32: 964-968.

Hill MF, Witman JD, Caswell H. 2004. Markov chain analysis of succession in a rocky subtidal community. Am Nat 164: E46-E61.

Ver Hoef JM, Cressie N. 1997. Using hidden Markov chains and empirical Bayes change-point estimation for transect data. Environ Ecol Stat 4: 247-264.

Horn HS. 1975. Markovian processes of forest succession. In: Ecology and Evolution of Communities (Cody ML, Diamond JM, eds). Harvard University Press, Cambridge, MA, pp 196-211.

Hyun K, Song MY, Kim S, Chon TS. 2005. Using an artificial neural network to patternize long-term fisheries data from South Korea. Aquat Sci 67: 382-389.

Jongman RHG, ter Braak CJF, van Tongerenm OFR. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

Juang BH, Rabiner LR. 1991. Hidden Markov models for speech recognition. Technometrics 33: 251-272.

Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. 1997. Predicting protein structure using hidden Markov models. Proteins Suppl 1: 134-139.

Kohonen T. 2001. Self-Organizing Maps. 3rd ed. Springer-Verlag, Heidelberg.

Krogh A, Brown M, Mian I, Sjölander K, Haussler D. 1994. Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 235: 1501-1531.

Legendre P, Legendre L. 1998. Numerical Ecology. 2nd ed. Elsevier, Amsterdam.

Liu Y, Lee SH, Chon TS. 2011. Analysis of behavioral changes of zebrafish (Danio rerio) in response to formaldehyde using Self-organizing map and a hidden Markov model. Ecol Model 222: 2191-2201.

MacDonald IL, Raubenheimer D. 1995. Hidden Markov models and animal behaviour. Biomet J 37: 701-712.

Merritt RW, Cummins KW. 1996. An Introduction to the Aquatic Insects of North America. Kendall Hunt, Dubuque, IA.

Park YS, Céréghino R, Compin A, Lek S. 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol Model 160: 265-280.

Park YS, Song MY, Park YC, Oh KH, Choe E, Chon TS. 2007. Community patterns of benthic macroinvertebrates collected on the national scale in Korea. Ecol Model 203: 26-33.

Pennak RW. 1978. Freshwater Invertebrates of the United States. 2nd ed. Wieley, New York, NY.

Quach QK, Chon TS, Kim HS, Nguyen TV. 2013. One and two-individual movements of fish after chemical exposure. J Korean Phys Soc 63: 18-27.

Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. P IEEE 77: 257-286.

Resh VH, Rosenberg DM. 1984. The Ecology of Aquatic Insects. Praeger Publishers, New York, NY.

Scarsbrook MR. 2002. Persistence and stability of lotic invertebrate communities in New Zealand. Freshw Biol 47: 417-431.

Shannon CE, Weaver W. 1949. The Mathematical Theory of Information. University of Illinois Press, Urbana, IL.

Simon G, Lee JA, Cottrell M, Verleysen M. 2007. Forecasting the CATS benchmark with the double vector quantization method. Neurocomputing 70: 2400-2409.

Taheri A, Tarihi MR, Abad HBB, Bababeyk H. 2005. Fuzzy hidden Markov models for speech recognition on based FEM Algorithm. In: The Second World Enformatika Conference (Cemal, A, ed.). WEC'05, Feb 25-27, 2005. Enformatika, Canakkale, Turkey, pp 59-61.

Tang H, Song MY, Cho WS, Park YS, Chon TS. 2010. Species abundance distribution of benthic chironomids and other macroinvertebrates across different levels of pollution in streams. Ann Limnol - Int J Lim 46: 53-66.

Tucker BC, Anand M. 2003. The use of matrix models to detect natural and pollution-induced forest gradients. Comm Ecol 4: 89-110.

Tucker BC, Anand M. 2004. On the use of stationary versus hidden Markov models to detect simple versus complex ecological dynamics. Ecol Model 185: 177-193.

Usher MB. 1981. Modelling ecological succession, with par-

ticular reference to Markovian models. In: Vegetation Dynamics in Grasslans, Healthlands and Mediterranean Ligneous Formations (Poissonet P, Romane F, Austin MA, van der Maarel E, Schmidt W, eds). Springer Netherlands, Amsterdam, pp11-18.

Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. 2000. SOM Toolbox for Matlab 5. Helsinki University of Technology, Espoo.

Visser I, Raijmakers MEJ, Molenaar PCM. 2002. Fitting hidden Markov models to psycological data. Sci Prog 10: 185-199.

Viovy N, Saint G. 1994. Hidden Markov models applied to vegetation dynamics analysis using satellite remote sensing. IEEE Trans Geosci Remote Sens, 32: 906-917.

Voegtlin T, Dominey PF. 2001. Learning high-degree sequences in a linear network. In: Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on Jul 15-19. IEEE, Washington, DC, pp 940-944.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58: 236-244.

Winston WL. 1994. Operations Research: Applications and Algorithms, 3rd ed. Duxbury Press, Belmont, CA.

Wiggins GB. 1996. Larvae of the North American Caddisfly Genera (Trichoptera). University of Toronto Press, Toronto.

Wootton JT. 2001b. Local interactions predict large-scale pattern in empirically derived cellular automata. Nature 413: 841-844.

Yemshanov D, Perera AH. 2002. A spatially explicit stochastic model to simulate boreal forest cover transitions: general structure and properties. Ecol Model 150: 189-209.

Yoon IB. 1995. Aquatic Insects of Korea. Jeonghaengsa, Seoul. (in Korean)