Alignment-Free Classification of G-Protein-Coupled Receptors Using Self-Organizing Maps

Joji M. Otaki,*,†,§ Akihito Mori,‡ Yoshimasa Itoh,‡ Takashi Nakayama,‡ and Haruhiko Yamamoto†

Department of Biological Sciences and Department of Information and Computer Science, Kanagawa University, 2946 Tsuchiya, Hiratsuka, Kanagawa 259-1293, Japan

Received September 8, 2005

Proteins are classified mainly on the basis of alignments of amino acid sequences. Drug discovery processes based on pharmacologically important proteins such as G-protein-coupled receptors (GPCRs) may be facilitated if more information is extracted directly from the primary sequences. Here, we investigate an alignment-free approach to protein classification using self-organizing maps (SOMs), a kind of artificial neural network, which needs only primary sequences of proteins and determines their relative locations in a two-dimensional lattice of neurons through an adaptive process. We first showed that a set of 1397 aligned samples of Class A GPCRs can be classified by our SOM program into 15 conventional categories with 99.2% accuracy. Similarly, a nonaligned raw sequence data set of 4116 samples was categorized into 15 conventional families with 97.8% accuracy in a cross-validation test. Orphan GPCRs were also classified appropriately using the result of the SOM learning. A supposedly diverse family of olfactory receptors formed the most distinctive cluster in the map, whereas amine and peptide families exhibited diffuse distributions. A feature of this kind in the map can be interpreted to reflect hierarchical family composition. Interestingly, some orphan receptors that were categorized as olfactory were somatosensory chemoreceptors. These results suggest the applicability and potential of the SOM program to classification prediction and knowledge discovery from protein sequences.

1. INTRODUCTION

Post-genome biology increasingly faces a challenge in how to classify a large number of unknown and known proteins in order to highlight their structural and functional properties. One of the important groups of proteins in drug discovery is G-protein-coupled receptors (GPCRs),[1−3] which comprise the largest superfamily, at least in mammals.[4−8] A GPCR molecule spans the membrane with seven transmembrane domains connected by extracellular and intracellular loops. Yet, many orphan GPCRs remain to be classified and characterized. In dealing with a wide variety of GPCRs, the alignment process necessitates the exclusion of loops and terminals of GPCR molecules, restricting the alignments to transmembrane domains. However, it has been pointed out that the lengths of loops and terminals contain significant biological information.[9,10] Furthermore, the structural and functional importance of extracellular loops was exemplified by the structural study of rhodopsin[11] together with biochemical methods.[12−14] Thus, developing an accurate classification system for GPCR raw sequences is of high importance.

Several lines of research have proposed new methods for an alignment-independent classification of GPCRs: the clustered database approach,[15] phylogenetic analysis,[16] length analysis,[9,10] support vector machines,[17] alignment-independent methods using principal component analysis (PCA),[18] and bagging classification trees.[19] Especially noteworthy is the method with PCA,[18] in which amino acid sequences are transformed into vectors using modified autocross-covariance terms obtained from principal components of physicochemical properties of amino acids. This approach was shown to be able to classify 535 new Class A GPCR samples precisely into 12 conventional categories with the exception of 14 misclassified samples, yielding 97.4% precision.[18] Together with a demand in proteomics for processing a large number of various protein samples, whose information is mostly limited to their primary sequences, a method may be highly useful that can extract information of each protein even from the primary sequence only, visually express their relations in the large population of proteins, and also learn new sequences as available information expands.

Here, we developed yet another alignment-free approach to protein classification using a type of artificial neural network called a self-organizing map, or SOM.[20] SOM is a type of competitive neural network, which performs unsupervised learning. Whereas PCA produces excellent results in dealing with linear relations among data sets, SOM can cope with nonlinear relations among data sets of interest. SOM can meaningfully locate samples from an input space to particular "neurons" in a two-dimensional lattice through an adaptive process. That is, SOM uncovers hidden patterns that are not recognizable to human eyes and translates them into recognizable patterns. The SOM method has been applied to various problems in biological sciences such as microarray data analysis and genomics,[21,22] comparative

* Corresponding author phone: +81-463-59-4111; fax: +81-463-58-9684; e-mail: otaki@bio.kanagawa-u.ac.jp.
† Department of Biological Sciences.
‡ Department of Information and Computer Science.
§ Present address: Department of Biology, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan. Phone and fax: +81-98-895-8557; e-mail: otaki@sci.u-ryukyu.ac.jp.

10.1021/ci050382y CCC: $33.50
Published on Web 04/05/2006

proteome analysis,[23,24] analysis of protein sequence similarity including segments of secondary structures,[25,26] and feature extraction for artificial neural networks.[25,26] Review papers by Schneider and co-workers[27,28] give a good perspective on SOM applications to proteome analysis. The ability of SOM to cluster protein sequences into families has already been demonstrated for 1758 human proteins,[23,24] in which the number of neurons used was rather small (15×15) and a protein sequence was represented in terms of a dipeptide matrix. It gave a feature map whose clustering corresponds to the families in question, but the map was rather general and was not sufficient for a detailed classification of the GPCR family. The classification program of GPCRs using SOM reported in this paper employs the representation method that followed Lapinsh et al.[18] and a fairly large output space of neurons. We achieved a total accuracy of family classification of 97.8%, which indicates that our SOM-based method could be used as an important tool for classifying GPCRs and other proteins, being an alternative or complementary method to other approaches.

2. METHODS AND RESULTS

2.1. Constitution of SOM. The constitution of the SOM that we employed is as follows: (a) The output space consists of 50×50 ($= 2500$) neurons posted to the nodes on a square lattice. Each neuron is fully connected to the input space and has a synaptic weight vector for the connection. The learning process is to update the weight vectors to adapt to input vectors. (b) Both a square map and a toroidal rectangular map were implemented, though square maps were described in this paper because there was no significant difference with respect to classification accuracy. (c) The so-called Batch Map method was used for learning, in which all data were fed at once at the beginning of the learning process. (d) The shape of the topological neighborhood is a circle or a part of a circle whose area reduces linearly according to the progress of the learning process. (e) The learning process is terminated when it converges, that is, when no changes are observed in any weight vectors.

The Batch Map procedure was carried out as follows:[20] (1) Initialize the weight vectors by appropriate random values. (2) Input all the input vectors; find a winner for each input vector, and store the input vector into the winner and its neighboring neurons. The winner is defined as a neuron having the best-matching weight vector with the input vector. Multiple input vectors could be stored in a single neuron. (3) Update the weight vector of each neuron by replacing it with the mean vector of input vectors that were temporarily memorized in the neuron and its neighboring neurons. (4) Repeat steps 2 and 3 with a gradual reduction of the topological neighborhood, until the map converges.

The initial values of weight vectors in our system were chosen randomly from input vectors. That is, 2500 values of initial weight vectors were a subset of the input data set. The number of training cycles until convergence was about 55 in the sense of Batch Map iteration. When the learning process converged, a feature map was obtained in the output space, which is a kind of clustering corresponding to some input categories. The mean quantization error was about 2.2 on the basis of the Euclidean distance metric. A particular neuron in a particular cluster becomes sensitive to a given input. Similar inputs are located in similar places in the feature map. Thus, the feature map can be used for classifying new data to some existing categories or to a new category near some known categories.

The feature map used for classification, however, was slightly different from a usual SOM output cluster map. Although the output map of SOM learning shows some clustering, it is not assured that the clusters correspond to GPCR families straightforwardly. Distances between weight vectors are often analyzed in order to give an objective border between clusters. The U-matrix (unified distance matrix) representation[29,30] is one of the well-known criteria for this purpose, yet the U-matrix we applied to this problem did not give a clear border (data not shown). We, thus, employed another approach to get proper clusters. That is, we mapped all of the training data on the neuron plane together with the family label and formed a family map (see section 2.3). The classification of GPCR sequences was performed on the basis of the family map, which is actually a kind of nearest-neighbor (NN) method in the weight vector space. Thus, the approach we employed for GPCR classification is to be called a SOM-NN-based method.

We developed all of the programs by ourselves using Java. Accordingly, we could implement various functions to help protein analysis in addition to SOM learning in the system.

2.2. Representation of GPCR. Because the input space of a SOM is a feature vector space, it is necessary to represent GPCR sequence data as a vector of a fixed dimension that reflects the characteristics of the GPCR family. Several approaches to yield such a representation may be sought. First, the use of aligned data is straightforward, because aligned data are mostly sequences of a fixed length. The alignment of protein sequences is based on a similarity search between the sequences, so we employed this alignment-based approach as a baseline experiment. Second, converting nonaligned raw amino acid sequences to feature vectors of a fixed length is preferable, if possible. Several ideas for this conversion have been proposed: simple amino acid composition,[31] dipeptide (i.e., bi-gram) occurrences in a sequence,[23,24] autocorrelation vectors,[27,32] fractal encoding,[33] autocross-covariance terms obtained from z scales of amino acids,[18] and so on. We used the autocross-covariance terms of z scales to make a feature vector for GPCR representation, following Lapinsh et al.[18] The z scales are principal components of physicochemical properties of amino acids using PCA,[34−36] in which 26 descriptors of physicochemical properties of amino acids are used. These descriptors are as follows: molecular weight; van der Waals volume; heat of formation; energy of the highest occupied molecular orbital; energy of the lowest unoccupied molecular orbital; log P; α polarizability; absolute electronegativity; absolute hardness; total molecular surface area; polar molecular surface area; number of hydrogen bond donors; number of hydrogen bond acceptors; indicator of positive charge in the side chain; indicator of negative charge in the side chain; NMR α-proton shift at pD $= 2$, 7, and 12.5; and seven descriptors representing thin-layer chromatographic mobilities using different stationary and mobile phases. Values z_1 to z_5 are the first five principal components derived from these 26 descriptors and largely represent hydrophobicity, steric properties, polarity (z_1-z_3), and electronic effects (z_4 and z_5).

Table 1. Number of GPCR Sequences in Each Family Collected from GPCRDB Release 8.0 as of February 2004 Using a Data Gathering Program.

family	number of sequences
amine	389
peptide	1014
hormone protein	50
(rhod)opsin	361
olfactory	1942
prostanoid	59
nucleotide-like	81
cannabinoid	13
platelet-activating factor	9
gonadotropin-releasing hormone	41
thyrotropin-releasing hormone and secretagogue	30
melatonin	17
viral	66
lysosphingolipid and LPA(EDG)	37
leukotriene B4	7
class A orphan/other	181
total	4297

Data Sets. Two kinds of data sets were prepared: one for alignment-assisted SOM and the other for alignment-free SOM. For the former, we used Class A (rhodopsin-like) GPCR sequences that were already aligned with the length of 248 amino acid residues using the "What If" program that executes a profile-based alignment.[17] They were obtained from GPCRDB release 8.0 as of February 2004 (http://www.gpcr.org/7tm/).[37] The total number of data records was 1397 excluding orphan receptors. For the latter, we downloaded each entry of GPCR from GPCRDB release 8.0, as of February 2004, using a data-gathering program developed by ourselves, because the raw data are not stored in the form of a single packaged file but are scattered in many places. The number of total samples collected was 4297 including orphan receptors. The downloaded GPCR sequences had various lengths from 209 to 1360 amino acids. The number of sequences in each family of the data set for alignment-free SOM is shown in Table 1. The data set is available at the author's Web site (http://www.nn.info.kanagawa-u.ac.jp). The input vectors to SOM learning were calculated from the sequence data according to the procedure described later in this section. The calculation of the input vectors is embedded in the program, so the data set of input vectors is not available directly, but the program is easy to write.

Alignment-Assisted SOM. Three kinds of representation were used.

(1) Symbolic representation. An amino acid symbol in a sequence directly corresponds to an element of the feature vector. That is, each amino acid sequence was treated straightforwardly as an input vector. The learning process is somewhat different from the usual one in that the winner was determined as the neuron that had the maximal number of elements in common between its weight vector and the input vector and in that a weight vector was updated so that each element was replaced with the amino acid at the same position that occurred most frequently in the neighborhood.

(2) Z3 representation. Three z scales (z_1−z_3) were assigned for each amino acid, which gives three kinds of sequences of numerical values. The three sequences were concatenated in an orderly fashion, giving a feature vector of 744 ($= 3 \times 248$) dimensions. The z-scale value for an alignment gap was set to zero.

(3) Z5 representation. This is the same as the Z3 representation explained above, except that five z scales (z_1−z_5) were employed instead of three. Here, an input vector had 1240 ($= 5 \times 248$) dimensions.

Alignment-Free SOM. As described above, nonaligned raw amino acid sequences were transformed into feature vectors with a fixed dimension. That is, a GPCR sequence was first transformed to the vector which had multiple z-scale values for one element. Here, the number of z scales, D, is three or five. Specifically, each amino acid residue is expressed as (z_1, z_2, z_3) or (z_1, z_2, z_3, z_4, z_5). After the sequence data was transformed to z scales, the autocross-covariance (ACC) term $c_{d1d2}(l)$ is calculated as

$$c_{d1d2}(l) = \sum_{i=1}^{n-l} \frac{(v_{d1,i} - \bar{v}_{d1})(v_{d2,i+l} - \bar{v}_{d2})}{(n-l)^p} \qquad (1)$$

where $l = 1...L$ (L is the maximal lag), $n =$ the total number of amino acids in the sequence, $d_k = 1...D$, $v_{jk} =$ the value of z_j of an amino acid in a sequence at position k, and $p =$ the degree of normalization. Because ACC is a function of l, an autocross-covariance matrix $\mathbf{C}(l)$ can be expressed as a $D \times D$ matrix. The feature vector was constructed first by linearly joining each row of $\mathbf{C}(l)$ as ($c_{11}...c_{1D}$, $c_{21}...c_{2D}$, ..., $c_{D1}...c_{DD}$), producing partial vectors, and then further concatenating them as [$\mathbf{C}(1)$, $\mathbf{C}(2)$, ..., $\mathbf{C}(L)$], finally producing a single feature vector with LD^2 dimensions. When these ACC terms were used, the following two kinds of representation were employed: (1) the ACC3 representation, where GPCR data are converted to a vector of ACC terms using three z scales (i.e., $D = 3$), and (2) the ACC5 representation, where GPCR data are converted to a vector of ACC terms using five z scales (i.e., $D = 5$).

There are two parameters in calculating ACC terms: maximum lag L and the degree of normalization p. The optimal parameter values were experimentally determined in terms of total classification accuracy by applying various combinations of them, where the range of maximum lag L was set from 10 to 105, and the degree of normalization, p, was set at 0, 0.5, 1.0, or 2.0. Although it does not cover all of the combinations of lag ranges, the rough tendency of classification accuracy was observed; thus, we checked the range between 0 and 30 as shown in Figure 1 (see section 2.5). It shows that the accuracy becomes higher according to the maximal lag increases and reaches a plateau at a maximal lag of about 10, and lag 23 seems to be the optimal lag. However, we employed lag 18 as the maximal lag, considering computation time. The feature vector thus obtained is a variation of correlation vectors that have been used in earlier works referred to above, and it is unique in that it is generated by concatenating many ACC terms in the range of specified lags.[18]

2.3. Family Identification and Evaluation of Performance. Family Map. After the learning process, a given neuron or its closely located neurons in the output space become responsive to similar input data, which are called responding neurons here. Thus, after the learning process, clusters of the responding neurons, called a feature map, which correspond to the clusters (a group of similar input

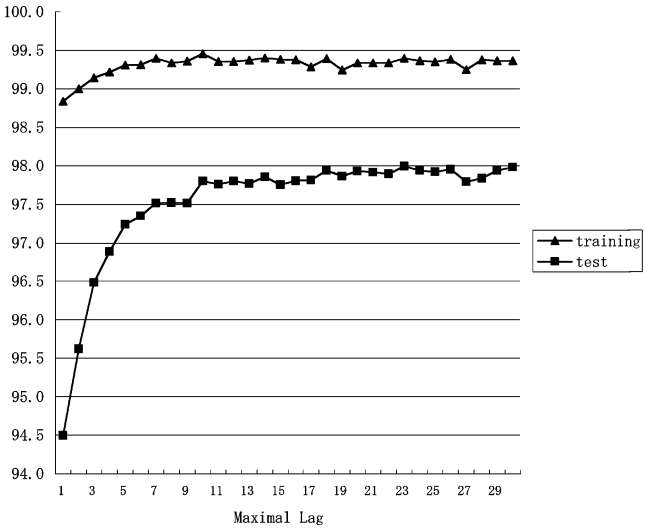

Figure 1. Behavior of total accuracy. Total accuracy becomes high according to the extension of maximal lag and reaches a plateau at about maximal lag 10.

data) of the training data, are formed in the output space. The sequences in the same family are expected to form a cluster, because they are to be similar to one another. Inversely, when one inputs a new sequence sample to the feature map, a neuron in the map responds, and the cluster to which this responding neuron belongs is considered to be the family for that sample. Here, because we know a belonging family for every test sample, we are able to calculate the classification accuracy after identifying families for all of the test samples.

However, in this classification approach, it is not assured that the clusters visually recognized on the SOM output map correspond to families straightforwardly. One reasonable method to ensure the reliability of clustering is to provide meaningful borders among clusters based on distances between weight vectors of neurons. The U-matrix method[29,30] is one of them. Another approach to the classification is to employ known families of training data as a label of each neuron. This is possible because the belonging families of training data are known in this case. Although SOM performs unsupervised learning, the labels of training data are preserved in the transformed space of weight vectors, which are projected onto a two-dimensional neuron plane. Therefore, we are able to construct a feature map so that each responding neuron is painted with the color that represents a corresponding family (15 colors are used for 15 families). We call this colored feature map a family map. This color-assigning process, not inherently required for SOM, was introduced for the sake of visualization of clustering (family areas) and classification.

Some neurons may respond to multiple data from different families. To get a family map, it is necessary to determine a family area that contains the most frequent "activator" family samples. For this purpose, we devised a family vector $\mathbf{f}^k = (f_1^k, f_2^k, ..., f_i^k, ..., f_{15}^k)$, where the superscript k indicates the kth neuron and subscript i indicates the ith family ($i = 1-15$). The element f_i^k is the number of responses to the ith family for the kth neuron and its neighboring neurons. That is, f_i^k indicates the number of activating data samples in each family for the kth neuron over all of the rounds of the batch SOM.

If there is just a single nonzero component in the final family vector for a given neuron, that neuron is considered to be a recognizing neuron for that particular family. If there are two or more nonzero components in the final family vector for a given neuron, it is considered to be a recognizing neuron for the family with the highest nonzero component. That is, we simply chose arg $\max_i\{f_i^k\}$ in determining the family category. In case two or more components are the same value, the family of the smallest members is chosen as a family category.

To be sure, on the basis of inherent characters of SOM, the existence of the cross-family recognizing neuron indicates that those data, and hence those GPCRs, that activated the neuron are similar to one another overall despite having been classified into different conventional families. However, for the sake of clarity and simplicity, we maintained a one-to-one correspondence between the neuron and the family and produced a colored family map, assuming that the conventional classification into 15 families is essentially "correct".

As mentioned in the previous section, the classification method using the family map is actually a NN method in the weight vector space of neurons, where a family label of a new sequence (test data) is to be the one that the nearest known data belongs to. Furthermore, the classification method would become k-nearest neighbor if the top k elements of a family vector were used. As a matter of fact, the nearest-neighbor method can be applied to the input vector space directly, and the result is comparable to the one using SOM (data not shown in this paper).

There are two types of family maps: an overall map and a final map. An overall map was formed by assigning a family color to every neuron on the map to reflect all of the learning process, as shown in Figure 2. In a final map, family areas were composed of only responding neurons to which any of the training data were mapped after the SOM learning was completed. Each family area of a final map is included in a corresponding family area of an overall map, as shown in Figure 3 (see sections 2.4 and 2.5). The classification was performed using a final map.

Evaluation of Classification Performance. The classification power of the SOM against the conventional categories is evaluated in three ways by calculating selectivity and sensitivity (or precision and recall in the field of information retrieval). The recall of a family is defined as follows:

$$\text{Family Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} = \frac{\text{number of data correctly classified to the family}}{\text{number of data in the conventional family}} \quad (2)$$

In eq 2, tp and fn mean true positive and false negative with respect to the family in question, respectively. That is, the numerator (tp) indicates the number of correct responses of neurons in an output family area in reference to the conventional family classification (input family) when one inputs all of the data available after the family map was made. The fn indicates input data that are misclassified to other families. Here, the expression "the number of correct responses" is defined by the final family map, which is composed of the neurons that finally responded to all of the training data after completion of the learning.

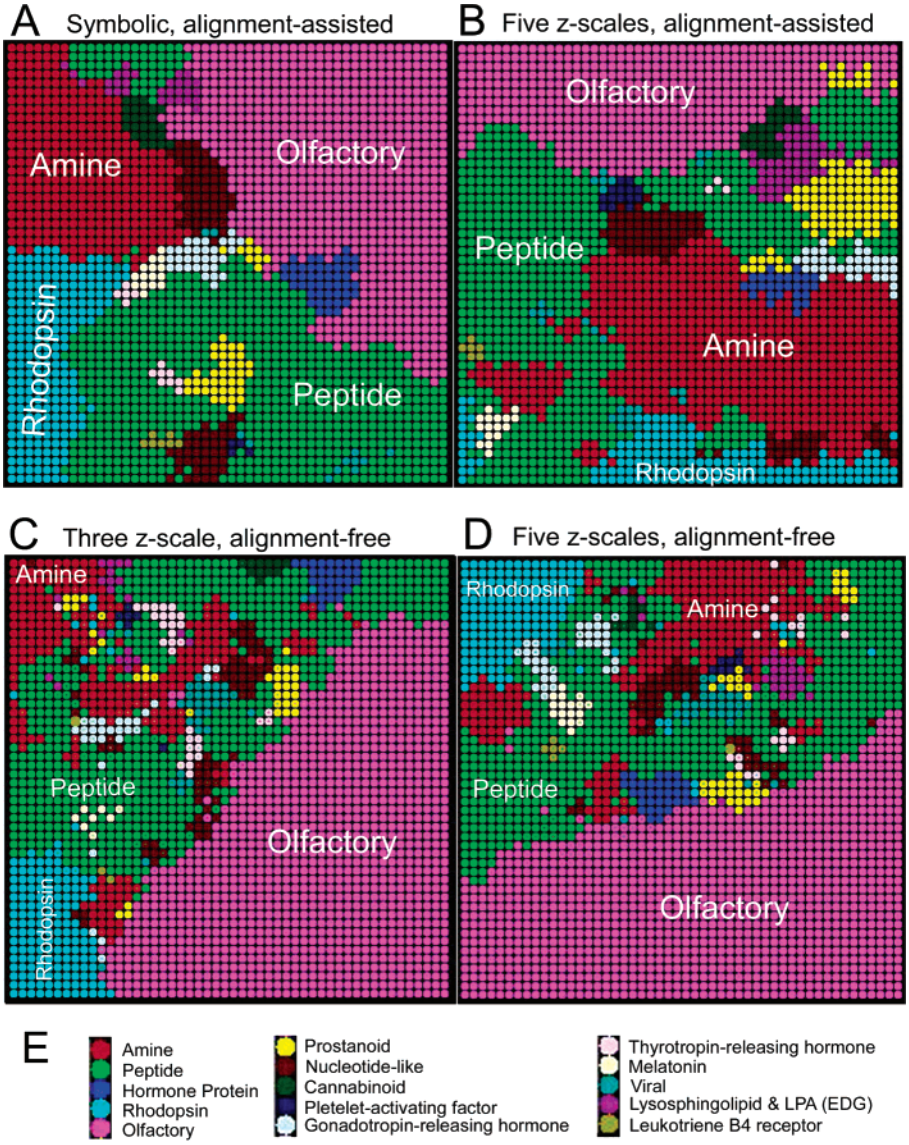

Figure 2. Alignment-assisted and alignment-free SOM outputs (overall family maps). (A) Alignment-assisted SOM output using symbolic representation. (B) Alignment-assisted SOM output using Z5 representation. (C) Alignment-free SOM output using ACC3 representation. (D) Alignment-free SOM output using ACC5 representation. (E) Color code for each family.

The precision of a family is defined as follows:

$$\text{Family Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} = \frac{\text{number of data correctly classified to the family}}{\text{number of data mapped to the family area}} \quad (3)$$

While the numerator of eq 3 is the same as that of the recall above, fp means false positive, which indicates input data not to be classified to the family in question, that is, noise. That is, the denominator expresses the number of all of the data that activate the neurons in the family area. Thus, eq 3 represents the precision of classification in terms of the family.

The total accuracy of classification for all of the input data is defined as the ratio of the number of correctly classified data to the number of all input data:

$$\text{Total Accuracy} = \frac{\text{number of data correctly classified}}{\text{number of input data}} \quad (4)$$

The sum of family recalls agrees with that of family precisions and gives the total accuracy above.

Cross Validation. Cross validation was performed by randomly choosing a test set instead of a *k*-fold approach. Specifically, half of the sequences of each family were chosen randomly, and a test set was generated by putting them together. The rest of the data set was used as a training set. This procedure was repeated 30 times for cross validation. The test and training sets were arranged to sample half of the members from each family in order to avoid overfitting of the learning by SOM. Because the smallest size of the families is seven at present, the randomized method would be better than *k*-fold cross validation when the repetition exceeds seven times (30 times in this case). This randomized method of cross validation was applied to all of the experiments described in this paper.

2.4. Alignment-Assisted SOM. To test the ability of our alignment-free SOM program to classify GPCR samples, we first tried alignment-assisted SOM as a baseline experiment. The downloaded Class A aligned GPCR samples[37] had

Figure 3. Each family area of 15 GPCR families on a SOM map shown separately. Each family's final map is included in its overall map. The dark shaded areas are overall maps of families, and colored areas are the corresponding final maps. Color codes are the same as with Figure 2.

already been classified into 15 conventional families. Among them, four families, that is, amine, peptide, rhodopsin, and olfactory receptors, are relatively large in number, comprising many subfamilies and sub-subfamilies.

Each family of the downloaded data set was divided into two groups, training and test data sets. The total number of training and test data excluding orphan receptors was 703 and 694, respectively. The training set was selected randomly from the data set, and the rest was used for the test set. A total of four test and training sets were arranged for cross validation for each combination of symbolic, Z3, and Z5 representations. Using these training sets, the learning was carried out, and the results were tested for the three kinds of representation of GPCR: symbolic, Z3, and Z5 representations, as explained above. The total accuracy of classification shown in Table 2 (top three lines) was calculated as a mean value of those four trials. As shown, the classification accuracy for the test set was 99.3% for the symbolic representation, 99.0% for the Z3 representation, and 99.2% for the Z5 representation. All three representations yielded reasonably high and similar values of accuracy, but the Z5 representation seems good in regard to the standard deviation. Figure 2 shows overall family maps constructed by one of the trials. For instance, the area of the olfactory family

Table 2. Total Accuracy of Classification

representation	number of families	training data		test data	
		accuracy[a]	SD[b]	accuracy[a]	SD[b]
symbolic	15	99.63	0.06	99.28	0.17
Z3	15	99.43	0.05	99.02	0.26
Z5	15	99.83	0.01	99.23	0.09
ACC3	15	99.35	0.17	97.49	0.40
ACC5	15	99.41	0.23	97.84	0.34
ACC3 with orphan	27 (15 + 12)	99.19	0.17	97.10	0.38
ACC5 with orphan	27 (15 + 12)	99.13	0.18	97.26	0.44

[a] Total accuracy is the mean values of four independent trials for alignment-assisted learning (symbolic, Z3, and Z5 representations) and 30 independent trials for alignment-free learning. [b] SD means standard deviation.

seemed very large, which suggests that the olfactory family contains many subfamilies. It is noteworthy that only a single sample of olfactory receptor O6K3_HUMAN (OR family 6, subfamily K) was misclassified as a rhodopsin family member in the symbolic map, whereas no sample was misclassified using Z3 or Z5 representation, indicating the high discrimination power of our program. On the other hand, some nonolfactory receptors were misclassified as olfactory (Table 3). As for the Z5 representation, for instance, the recall

Table 3. Nonolfactory Receptors Misclassified as Olfactory Receptors (i.e., False Positive) in the Alignment-Assisted SOM Using Symbolic, Z3, and Z5 Representations.

sample ID	conventional family	representation	identity
REIS-TODPA	rhodopsin	symbolic, Z3 and Z5	rhodopsin (Japanese flying squid)
ETB2_HUMAN	peptide	Z3	human endothelin B receptor-like
GP37_HUMAN	peptide	Z3	human endothelin B receptor-like
NTR2_HUMAN	peptide	Z3	human neurotensin receptor type 2
VU51_HSV6U	viral	symbolic	human herpes virus
VG74_HSVSA	viral	Z3	human herpes virus

Table 4. Recall (Sensitivity) of Classification for Each Family in the Alignment-Free SOM Using ACC5

	training data		test data	
family	recall[a]	SD[b]	recall[a]	SD[b]
amine	98.91	1.02	97.22	1.68
peptide	98.94	0.46	97.30	0.80
hormone protein	100.0	0.00	100.0	0.00
(rhod)opsin	99.59	0.54	96.87	1.25
olfactory	99.96	0.07	99.86	0.12
prostanoid	99.78	0.85	98.74	4.00
nucleotide-like	96.34	2.92	83.58	7.12
cannabinoid	100.0	0.00	95.00	15.26
platelet-activating factor	100.0	0.00	99.17	4.56
gonadotropin-releasing hormone	99.52	1.92	98.83	2.52
thyrotropin-releasing hormone and secretagogue	96.00	6.46	84.67	8.60
melatonin	99.63	2.03	95.42	12.05
viral	97.27	3.12	79.09	8.91
lysosphingolipid and LPA(EDG)	98.60	3.07	95.37	7.45
leukotriene B4	97.50	7.63	87.78	26.96
total accuracy	99.41	0.23	97.84	0.34

[a] The value is a mean of 30 trials. [b] SD means standard deviation.

Table 5. Precision (Selectivity) of Classification for Each Family in the Alignment-Free SOM Using ACC5

	training data		test data	
family	precision[a]	SD[b]	precision[a]	SD[b]
amine	98.71	0.89	94.35	2.36
peptide	99.62	0.39	97.05	0.64
hormone protein	100.0	0.00	96.88	5.57
(rhod)opsin	99.31	0.67	98.51	1.72
olfactory	100.0	0.02	99.88	0.19
prostanoid	97.98	2.27	93.82	6.36
nucleotide-like	96.40	2.39	89.63	6.80
cannabinoid	97.50	9.51	94.84	14.3
platelet-activating factor	98.89	4.23	94.70	12.8
gonadotropin-releasing hormone	99.10	2.15	96.67	4.58
thyrotropin-releasing hormone and secretagogue	90.77	6.94	86.42	10.4
melatonin	99.00	3.05	95.12	9.12
viral	97.37	3.30	95.87	6.23
lysosphingolipid and LPA(EDG)	96.11	4.35	94.75	7.20
leukotriene B4	91.79	12.7	85.89	21.2
total accuracy	99.41	0.23	97.84	0.34

[a] The value is a mean of 30 trials. [b] SD means standard deviation.

and precision of the olfactory family are 100% (194/194) and 99.5% (194/195), respectively.

2.5. Alignment-Free SOM. The total number of GPCR samples was 4297, including 181 orphan receptors. Half of these samples were selected randomly to make a training set, and the rest were used for a test set, as shown in Table 4. A total of 30 independent training sets were arranged for cross validation for each combination of ACC3 and ACC5 representation and with and without orphan receptors. The learning process was carried out for these four kinds of representation of GPCRs. Cross validation was performed by executing learning and test phases for 30 different training and test data sets.

Figure 1 shows the behavior of the total accuracy of classification in the variation of maximal lag in the case of ACC5 representation. It shows that the accuracy becomes higher in accordance with the maximal lag and reaches a plateau at a maximal lag of about 10. A large maximal lag, such as 100, gives a rather lower accuracy (data not shown in the figure). Thus, we employed a maximal lag of 18 according to this preliminary experiment. The total accuracy of classification shown in Table 2 (each of bottom four lines) was calculated as a mean value of those 30 learning and test trials. As shown in Tables 4 and 5, although different families had a different recall and precision (or sensitivity and selectivity), the total accuracy of 97.8% in the case of the ACC5 representation is almost the same as or superior to prior reported results.[18,19]

Figure 3 shows family maps of 15 families separately, in which final maps are included in overall maps. The dark shaded areas are overall maps of families, and colored areas correspond to final maps. Any combinations of families can be displayed as family maps according to the user's designation, and a family vector of a winning neuron is obtained, so that the users could easily grasp the relationship among GPCR sequences intuitively.

Here, we focused on the family maps that were obtained from the ACC5 representation. In the map (Figures 2D and 3), olfactory receptors, the largest family, occupied the largest continuous area, about one-third of the entire map. An almost single cluster was observed, which was a sharp contrast to other receptors such as peptide receptors. Olfactory receptors have several distinctive motifs, being different from other GPCRs.[38] Although properly identified as olfactory receptors, some samples were exceptionally placed at distant neurons in the map, whose sample IDs were O93549 (goldfish), Q9DGH4 (Atlantic salmon), Q9I835 (common carp), Q9I8Y8 (zebrafish), and Q9PSJ4 (channel catfish). They all belonged to fish receptors. Because fish-like receptors are known to be different from other vertebrate receptors,[38] these receptors may have distinctive sequence signatures.

In the olfactory receptor family, one to seven receptors were misclassified (i.e., false negative) among 1942 samples in each trial (Table 6), achieving a 99.8% recall (i.e., sensitivity) on average in 30 trials. One olfactory receptor, Q9YHY2, was frequently misclassified as an amine receptor. Indeed, it is an "olfactory amine" receptor in the primitive vertebrate European river lamprey.[39] Other receptors from Japanese Medaka fish may well be amine or peptide receptors. Although some human and mouse receptors were also misclassified for no obvious reasons, this type of "misclassification" was unavoidable, because the conven-

Table 6. Olfactory Receptors Misclassified as Other Receptors (i.e., False Negative) in the Alignment-Free SOM Using ACC5

sample ID	misclassifed family	identity
Q9PVP4	amine or peptide	Medaka fish E4 olfactory receptor
Q9PVP5	amine or peptide	Medaka fish E3 olfactory receptor
Q9PVP6	amine or peptide	Medaka fish E2 olfactory receptor
Q9PVP7	amine or peptide	Medaka fish E1 olfactory receptor
Q9PVW1	peptide	Medaka fish mfOR2
Q9PVW2	peptide	Medaka fish mfOR1
Q9YHY2	amine	European river lamprey
Q8NG79	peptide	human olfactory receptor
Q7TRW4	peptide	mouse Olfr419
Q8VFU6	nucleotide-like	mouse Olfr826 or MOR210-2
Q8VFY5	peptide	mouse MOR267-6

tional categories do not always reflect pharmacological properties of receptors. Specific classification results depended on the combination of training and test sets, and naturally, misclassified samples were different from trial to trial.

On the other hand, 21 nonolfactory receptors were classified as olfactory (i.e., false positive) in, overall, 30 trials, achieving a 99.7% precision (i.e., selectivity) on average (Table 7). Many misclassified samples belonged to chemokine receptors mainly used in the immune system, possibly suggesting some molecular relations between olfactory and chemokine receptors. Some of the other misclassified samples were from invertebrates or viruses.

The rhodopsin family showed a main single cluster (Figures 2D and 3). Some rhodopsin neurons were, however, isolated in the middle of peptide and other families. Peptide receptors, on the other hand, were located in a relatively wide but nearly continuous area (Figures 2D and 3). This probably reflects the fact that the peptide receptor family is by far the most divergent family in the Class A GPCRs, comprising 26 subfamilies identified in the GPCRDB. Similarly, amine receptors were also distributed relatively widely and indeed contain 7 subfamilies, making it the second most "variable" family. Clear distinction between amine and peptide areas was difficult to make, because they intermingled with each other. Other smaller families, that is, hormone protein, prostanoid, nucleotide-like, cannabinoid, platelet-activating factor, gonadotropin-releasing hormone, tyrotropin-releasing hormone and secretagougue, melatonin, viral, lysosphingolipid and LPA (EDG), and leukotriene B4 receptors, seemed to be located in neurons that were mostly surrounded by peptide or amine neurons (Figures 2D and 3).

As shown in Figure 3, most families dispersed in several separate areas, although some families were extracted as a single cluster. The separation of a family area, which was resulted from the fairly high resolution of the neuron plane, seemed to reflect the hierarchical constitution of the family. Figure 4 shows subfamilies contained in family areas. The families with a single area, that is, hormone protein, olfactory, and gonadotropin-releasing hormone, located their subfamilies separately within their family areas, although this tendency was not clear in other families. Note that some neurons found in a subfamily map, but not in a corresponding final map, are mistakenly mapped GPCR sequences. Two nodes with a beige color in the upper part of the olfactory map and a blue node apart from the area of gonadotropin-releasing hormone are such examples.

2.6. Orphan Receptor Classification. In GPCRDB, Class A orphan/other receptors have been classified into 12 conventional categories as shown in Table 8. We here tentatively classified these receptors into 15 conventional families (that were already examined in the previous section) to extract information on their possible ligands and to test the flexibility of our program to cope with highly different types of GPCR samples.

We first tested the Class A orphan/other receptors by the family map that was obtained through the learning process using the ACC5 representation without orphan data. The result showed some appreciable tendencies in classifying these samples (Table 8). All samples in the RDC1 ($n = 8$; n indicates the number of sequences) and ORPH ($n = 4$) categories were located in the peptide-family area. All samples in the GP40-like category ($n = 7$) were also classified in the peptide family. Similarly, four samples out of five in the Mas proto-oncogene category were located in the peptide family area. LGR-like samples ($n = 19$) were all mapped in the hormone protein area. In contrast, samples in other orphan categories were mapped heterogeneously.

To examine the nature of orphan "misclassification", here, we focused on orphan receptors that were classified into olfactory receptors (Table 9). It seemed that all misclassified receptors were expressed in neural tissues. Moreover, two different groups of somatosensory chemoreceptors were revealed,[40,41] suggesting weak but detectable relations to olfactory receptors that have not been fully appreciated yet.

Second, we made the family map that was obtained through the learning process using the ACC3 and ACC5 representations including the Class A orphan/other receptors. After the learning process, each orphan sample was located in the map. Here, we found that, as in the previous map, most samples from the same orphan categories clustered very well (data not shown). An exception was the GPR category, whose samples were distributed widely. Classification accuracy among these tentative categories was very high except in the GPR and Mas-related receptors (Table 10). Thus, it seems that the result of classification using SOM supports the soundness of the conventional orphan classification with some exceptions.

3. DISCUSSION

We showed that the SOM-NN-based method is able to classify GPCR samples into 15 conventional categories with high accuracy. The program works well on both aligned and nonaligned samples. Although the clusters formed by SOM are not clearly separated from each other and several families are subdivided into subareas in the colored family map, the baseline SOM (i.e., alignment-assisted SOM) gives a very high total accuracy, as shown in Table 2. Alignment-free SOM also gives a high total accuracy, as also shown in Table 2.

Compared with several approaches to the alignment-free classification of protein sequences proposed so far, it seems that our SOM-NN-based approach is equal or superior in regard to the accuracy of the classification of protein families.[18,19] The SOM application to protein analysis was first reported in a series of studies by Ferrán et al.[23,24] to get a clustering of human proteins including GPCR sequences. The output space used in their study consists of 15×15 neurons, using an encoding scheme of dipeptide matrices, that is, 20-dimensional input vectors of amino acid bi-grams

Table 7. Nonolfactory Receptors Misclassified as Olfactory Receptors (i.e., False Positive) in the Alignment-Free SOM Using ACC5

sample ID	conventional family	identity
O93281	peptide	chiken putative CXC chemokine receptor
CCR5_RAT	peptide	rat CXC chemokine receptor type 5
GP44_MOUSE	peptide	mouse chemoattractant receptor homologue
CXC1_HUMAN	peptide	human chemokine XC receptor 1
CML1_RAT	peptide	rat chemokine receptor-like 1
R3R2_HUMAN	peptide	human relaxin 3 receptor 2
Q8NGA4	peptide	mouse C5 anaphylatoxin receptor
Q8BW93	peptide	mouse C5 anaphylatoxin receptor
5H6_MOUSE	amine	mouse serotonin receptor 6
Q63004	amine	rat serotonin receptor 6
Q8NGA4	peptide	human GPCR
Q868G4	rhodopsin	amphoxius rhodopsin
OPS2_PATYE	rhodopsin	ezo giant scallop rhodopsin
REIS_TODPA	rhodopsin	Japanese flying squid rhodopsin
RGR_HUMAN	rhodopsin	human opsin-related
CB2R_RAT	cannabinoid	rat cannabinoid receptor 2
Q869J1	gonadotropin-releasing hormone	ascidian ciona intestinalis GnRHR2
Q7TFC8	viral	rhesus cytomegalovirus
Q8BC83	viral	cercopithecine herpesvirus
Q8BC82	viral	cercopithecine herpesvirus
UL33_RCMVM	viral	rat cytomegalovirus

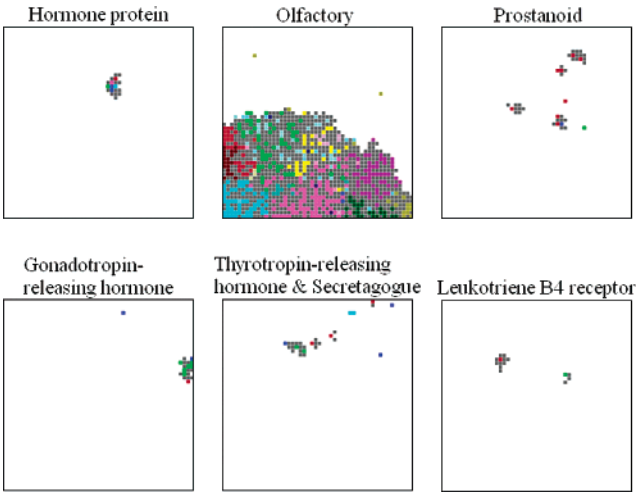

Figure 4. Subfamily maps of several families that have subfamilies of level 1. For instance, the olfactory family has 17 subfamilies, and 17 subareas are depicted with different colors within the olfactory family area. The subfamilies are distributed separately, which indicates that SOM learning preserves subfamily structures. Color codes are different from those of Figures 2 and 3 and are used only to distinguish subfamilies in a family.

in a sequence. The clustering obtained was used for comparative protein analysis and was not intended to classify detailed GPCR families. More recently, the SOM with encoding scheme of amino acid composition vectors has been successfully applied to the clustering and visualization of proteins, which was used to provide appropriate training data sets for artificial three-layered neural networks to predict subcellular localization.[25] The SOM with encoding scheme of physicochemical properties has also been applied to feature extraction of the targeting signal to make a predictive model for genome analysis.[26] The SOM output map is useful as a preliminary or complementary tool for a visual grasp of clustering.[27,28]

However, there is no assurance that SOM output could function as a classifier of proteins as described in section 2.3, so we employed the NN method in the weight vector space after SOM learning to classify sequences. The colored family map corresponds roughly to clusters learned by SOM and reflects family sizes and hierarchical family structures. For instance, the olfactory family contains 1942 sequences and 17 subfamilies occupying a broad area, the peptide family contains 1014 sequences and 30 subfamilies with a maximal depth of three dispersing to several areas, the leukotriene B4 receptor family contains only seven sequences, and so forth. Examples of subfamily clusters of olfactory receptors are shown in Figure 4. Thus, the colored family map roughly makes clusters of GPCR families preserving subfamily structures, which could give users convenience for visual inspections.

In addition, the key to success in employing machine learning methods, including SOM, is first how the representation of input data is set up. In this context, it seems that the representation using *z* scales and ACC terms contributes to the high performance of our SOM program, which reflects the physicochemical properties of amino acids, the interaction between amino acids, and the length information in the sequence including transmembrane parts.[18] The most significant feature of the modified ACC is that it concatenates multiple ACC terms to cover long-range interaction between amino acids in a sequence,[18] and it substantially improved the classification accuracy. Thus, the representation employed in our SOM program is advantageous over the ones using simple amino acid composition or a dipeptide matrix.[19,23] Actually, a PCA-based approach already yielded high classification precision: 535 Class A GPCRs were classified into 12 categories with a 97.4% precision using 929 GPCR samples with known families.[18] In contrast to this work, our GPCR samples were much more broadly collected ($n = 4116$) and classified into 15 (instead of 12) families with 97.8% precision, which shows the potential usefulness of this method as well as the PCA approach. Furthermore, SOM is highly flexible in processing a large amount of any protein sequence data in addition to GPCRs.

Although some samples were misclassified, in many cases, it seemed that misclassified samples were not the "misclassified" per se. That is, sources of misclassification will unavoidably emerge not from the inability of the program to differentiate but from incorrect conventional classification

Table 8. Orphan/Other Receptor Classification by the SOM Using ACC5 without Orphans

tentative classification	samples	SOM output[a]
platelet ADP and KI01	14	amine (9)/peptide (4)/platelet-activating factor (1)
SREB	9	prostanoid (3)/peptide (3)/gonadotropin-releasing (2)/leukotriene B4 (1)
Mas proto-oncogene	5	peptide (4)/lysosphingolipid and LPA (EDG) (1)
RDC1	8	peptide (8)
EBV-induced	1	peptide (1)/nucleotide-like (1)
ORPH	4	peptide (4)
LGH-like	19	hormone protein (19)
GPR	51	peptide (30)/amine (8)/prostanoid (4)/gonadotropin-releasing (4)/ platelet-activating factor (2)/olfactory (2)/rhodopsin (1)
GPR45-like	8	amine (3)/hormone protein (3)/peptide (2)
cysteinyl leukotriene	10	peptide (6)/nucleotide-like (3)/lysosphingolipid and LPA (EDG) (1)
Mas-related receptors (MRGs)	44	Peptide (26)/olfactory (6)/leukotriene B4 (4)/amine (3)/melatonin (3)/ nucleotide-like (2)
GP40-like	7	peptide (7)
total	181	

[a] Numbers of samples are indicated in parentheses.

Table 9. Nonolfactory Orphan/Other Receptors Misclassified as Olfactory Receptors in the Test Trials (False Positive)

sample ID	original category	description
C5L2_Human	peptide	C5a anaphylatoxin, expressed in neural tissue
OPS2_PATYE	(rhod)opsin	mollusc rhodospin
GPR7_BOVIN	orphan (GPR)	GPR7—neuropeptide B/W receptor type 1, expressed in the CNS
Q8BYC4	orphan (GPR)	GPR20—putative G-protein-coupled receptor homologue
Q8TDD8	orphan (Mas-related)	G-protein-coupled receptor SNSR4[a]
Q8TDE0	orphan (Mas-related)	G-protein-coupled receptor SNSR2[a]
Q8TDE1	orphan (Mas-related)	G-protein-coupled receptor SNSR1[a]
Q91ZB5	orphan (Mas-related)	MRGG—G-protein-coupled receptor[b]
Q96LB0	orphan (Mas-related)	MRGX3—G-protein-coupled receptor[b]
Q96LB2	orphan (Mas-related)	MRGX1—G-protein-coupled receptor[b]

[a] Sensory neuron-specific GPCRs. [b] Somatosensory nociceptive chemoreceptors.

Table 10. Recall/Precision of Orphan/Other Receptor Classification by the SOM Using ACC5 with Orphans

	training data		test data		training data		test data	
family	recall	SD	recall	SD	precision	SD	precision	SD
platelet ADP and KI01	100.0	0.00	94.76	8.78	99.17	3.17	97.95	6.42
SREB	96.00	15.22	100.0	0.00	95.73	10.25	92.14	15.71
Mas proto-oncogene	96.67	18.26	96.67	18.26	100.0	0.00	97.22	10.80
RDC1	96.67	18.26	95.83	18.26	97.67	9.71	95.43	12.09
EBV-induced	83.33	37.90	80.00	40.68	91.11	20.40	98.33	9.13
ORPH	93.33	25.37	93.33	25.37	95.56	13.79	98.89	6.09
LGR-like	100.0	0.00	96.67	7.80	97.37	5.13	89.65	12.85
GPR	95.64	4.25	76.13	11.29	94.17	4.47	86.23	10.06
GPR45 like	100.0	0.00	95.83	14.80	89.79	15.82	90.67	15.30
cysteinyl leukotriene	97.33	6.91	92.00	24.41	94.65	9.40	93.39	14.87
Mas-related (MGRs)	99.70	1.15	96.06	5.17	99.42	1.50	96.08	6.51
GP40-like	93.33	15.99	77.78	31.96	90.44	13.41	89.92	18.53

as well as from molecular features themselves. While there exist some neurons which respond to two or more GPCR sequences, each neuron is assigned a single family determined by their own family vector, as described in section 2.3. The classification of input data is performed according to the family map generated in this manner, and thus, the sequences other than "winning family" are misclassified inevitably. However, the misclassification suggests the proximity of those misclassified sequences to the family. For example, an olfactory receptor from the European lamprey[39] is one of the misclassified samples, but it should be considered to be "correct", because it is indeed an amine receptor. Similarly, many misclassified fish receptors may be amine or peptide receptors. The misclassification of chemokine receptors as olfactory is also interesting in terms of their functional analogy.

It is noteworthy that the conventional classification unavoidably reflects the history of biological sciences. For instance, the independence of the thyrotropin-releasing hormone family from the peptide family in the conventional classification may not be fully justified, because thyrotropin-releasing hormone is a peptide. While most families are named after chemical groups of their ligands, the olfactory family is a collection of receptors that recognize a wide variety of chemical compounds. The fact that our SOM program was, nonetheless, capable of differentiating most receptor samples into the conventional categories readily points out that the SOM output data are not only a reflection of ligand binding sites but also other structural features of receptors. This is also confirmed in examining the ligand−receptor specificity in olfactory receptors. We noticed that some but not all olfactory receptors that have similar ligands

tended to be located in similar places in the map (data not shown).

To be sure, our approach also needs further refinement. The location and shape of a family area in the two-dimensional output maps depend on the number of training samples in that family and do not necessarily reflect the density of accumulated samples in their neurons. Neurons located in a family boundary could belong to both families, but it was not shown in this map, because only a single color was assigned for a single neuron for simplicity and clarity. Furthermore, in some families, neurons with the same color were found at different locations despite higher classification precision. Because it has been established that similar patterns or input samples in general cluster at close locations in SOM,[20] the implication is that (1) input vectors are not appropriate, (2) the conventional classification is not appropriate, or (3) the conventional and SOM-based classification systems reveal different aspects or levels. On the basis of the high precision values in the alignment-free SOM and also the fact that the conventional one is not solely based on the sequence data, it is conceivable that our results are not unreasonable but simply an alternative mode of classification using different input information.

Because studying orphan GPCRs may lead to the discovery of a new class of natural ligands and drugs, extensive efforts have been made in rationally classifying GPCRs.[1−3] It appeared that the tentative conventional orphan receptor classification is largely consistent with our results. Intriguingly, some orphan sensory receptors were "misclassified" as olfactory receptors. This probably suggests distant but significant relations between these orphan receptors and the conventional olfactory receptors. Systematic combinations of several computational methods for protein characterization will yield a more effective classification and functional identification of proteins and ensure the inference. We believe that this method provides us with one of the important tools for the functional classification of a large number of proteins in proteomics.

ACKNOWLEDGMENT

We thank Souichi Mizuniwa for technical assistance. This work was partly supported by the High-tech Research Center Project from the Ministry of Education, Culture, Sports, Science, and Technology, and by the Grant for the Advancement of Scientific Collaborations from Kanagawa University.

REFERENCES AND NOTES

(1) Nambi, P.; Aiyar, N. G protein-coupled receptors in drug discovery. *Assay Drug Dev. Technol*. **2003**, *1*, 305−10.
(2) Shaaban, S.; Benton, B. Orphan G protein-coupled receptors: from DNA to drug targets. *Curr. Opin. Drug Discov*ery *Dev*. **2001**, *4*, 535−47.
(3) George, S. R.; O'Dowd, B. F.; Lee, S. P. G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 808−20.
(4) Schwartz, T. W. Molecular structure of G-protein-coupled receptors. In *Textbook of Receptor Pharmachology*; Foreman, J. C., Johansen, T., Eds.; CRC Press: Boca Raton, FL, 1996.
(5) Wess, J. Molecular basis of receptor/G-protein-coupling selectivity. *Pharmacol. Ther*. **1998**, *80*, 231−264.
(6) Bockaert, J.; Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J*. **1999**, *18*, 1723−1729.
(7) Lynch, K. R. G protein-coupled receptor informatics and the orphan problem. In *Identification and Expression of G Protein-coupled Receptors*; Lynch, K. R., Ed.; Wiley-Liss: New York, 1999
(8) Schöneberg, T. GPCR superfamily and its structural characterization. In *Understanding G Protein-coupled Receptors and their Role in the CNS*; Pangalos, M. N., Davies, C. H., Eds.; Oxford University Press: Oxford, U. K., 2002.
(9) Otaki, J. M.; Firestein, S. Length analyses of mammalian G-protein-coupled receptors. *J. Theor. Biol.* **2001**, *211*, 77−100.
(10) Otaki, J. M.; Yamamoto, H. Length analyses of *Drosophila* odorant receptors. *J. Theor. Biol.* **2003**, *223*, 27−37.
(11) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Trong, I. L.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **2000**, *289*, 739−745.
(12) Terakita, A.; Yamashita, T.; Shichida, Y. Highly conserved glutamic acid in the extracellular IV−V loop in rhodopsins acts as the counterion in retinochrome, a member of the rhodopsin family. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 14263−14267.
(13) Yan, E. C.; Kazmi, M. A.; De, S.; Chang, B. S.; Seibert, C.; Marin, E. P.; Mathies, R. A.; Sakmar, T. P. Function of extracellular loop 2 in rhodopsin: glutamic acid 181 modulates stability and absorption wavelength of metarhodopsin II. *Biochemistry* **2002**, *41*, 3620−3627.
(14) Yan, E. C.; Kazmi, M. A.; Ganim, Z.; Hou, J. M.; Pan, D.; Chang, B. S.; Sakmar, T. P.; Mathies, R. A. Retinal counterion switch in the photoactivation of the G protein-coupled receptor rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9262−9267.
(15) Graul, R. C.; Sadee, W. Evolutionary relationships among G protein-coupled receptors using a clustered database approach. *AAPS Pharm-Sci.* **2001**, *3*, E12.
(16) Joost, P.; Methner, A. Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biol*. **2002**, *3*, 0063.
(17) Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147−159.
(18) Lapinsh, M.; Gutcaits, A.; Prusis, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. S. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.* **2002**, *11*, 795−805.
(19) Huang, Y.; Cai, J.; Ji, L.; Li, Y. Classifying G-protein coupled receptors with bagging classification tree. *Comput. Biol. Chem.* **2004**, *28*, 275−280.
(20) Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: New York, 2001.
(21) Kasturi, J.; Acharya, R.; Ramanathan, M. An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics* **2003**, *19*, 449−458.
(22) Abe, T.; Kanaya, S.; Kinouchi, M.; Ichiba, Y.; Kozuki, T.; Ikemura, T. Informatics for unveiling hidden genome signatures. *Genome Res.* **2003**, *13*, 693−702.
(23) Ferrán, E. A.; Ferrara, P. Topological maps of protein sequences. *Biol. Cybern.* **1991**, *65*, 451−458.
(24) Ferrán, E. A.; Pflugfelder, B.; Ferrara, P. Self-organized neural maps of human protein sequences. *Protein Sci*. **1994**, *3*, 507−521.
(25) Schneider, G. How many potentially secreted proteins are contained in a bacterial genome? *Gene* **1999**, *237*, 113−121.
(26) Zuegge, J.; Ralph, S.; Schmuker, M.; McFadden, G. I.; Schneider, G. Deciphering apicoplast targeting signals − feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **2001**, *280*, 19−26.
(27) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biol. Mol. Biol*. **1998**, *70*, 175−222.
(28) Schneider, G.; Fechner, U. Advances in the prediction of protein targeting signals. *Proteomics* **2004**, *4*, 1571−1580.
(29) Ultsch, A.; Siemon, H. P. Kohonen's self-organizing feature maps for exploratory data analysis. In *Proc. Intern. Neural Networks*; Kluwer Academic Press: Paris, 1990; pp 305−308.
(30) Ultsch, A.; Guimaraes, G.; Korus, D.; Li, H. Knowledge extraction from artificial neural networks and applications. In *Proc. Transputer Anwender Treffen/World Transputer Congress TAT/WTC 93 Aachen*; Springer-Verlag: New York, 1993; pp 194−203.
(31) Chou, K. C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 319−344.
(32) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepline agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.
(33) Hanke, J.; Reich, J. G. Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. *CABIOS* **1996**, *12*, 447−454.
(34) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure−activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(35) Wold, S.; Jonsson, J.; Sjöström, M.; Sandberg, M.; Rännar, S. DNA and peptide sequences and chemical processes, multivariately modeled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta* **1993**, *277*, 239−253.
(36) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.
(37) Horn, F.; Bettler, E.; Oliveira, L.; Campagne, F.; Cohen, F. E.; Vriend, G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* **2003**, *31*, 294−297.
(38) Liu, A. H.; Zhang, X.; Stolovitzky, G. A.; Califano, A.; Firestein, S. J. Motif-based construction of a functional map for mammalian olfactory receptors. *Genomics* **2003**, *81*, 443−456.
(39) Berghard, A.; Dryer, L. A novel family of ancient vertebrate odorant receptors. *J. Neurobiol.* **1998**, *37*, 383−392.
(40) Dong, X.; Han, S.; Zylka, M. J.; Simon, M. I.; Anderson, D. J. A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. *Cell* **2001**, *106*, 619−632.
(41) Lembo, P. M.; Grazzini, E.; Groblewski, T.; O'Donnell, D.; Roy, M. O.; Zhang, J.; Hoffert, C.; Cao, J.; Schmidt, R.; Pelletier, M.; Labarre, M.; Gosselin, M.; Fortin, Y.; Banville, D.; Shen, S. H.; Strom, P.; Payza, K.; Dray, A.; Walker, P.; Ahmad, S. Proenkephalin A gene products activate a new family of sensory neuron-specific GPCRs. *Nat. Neurosci.* **2002**, *5*, 201−209.

CI050382Y