JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Homology Similarity Analysis of Sequences of Lactoferricin and Its Derivatives

SHURYO NAKAI,* JUDY C. K. CHAN, EUNICE C. Y. LI-CHAN, JINGLIE DOU, AND MASAHIRO OGAWA

Food, Nutrition, and Health Program, Faculty of Agricultural Sciences, The University of British Columbia, 6650 N.W. Marine Drive, Vancouver, British Columbia, Canada V6T 1Z4

A new method, homology similarity analysis (HSA), was developed to investigate homology pattern similarities of selected segments within sequences of peptides. This new approach facilitated elucidation of the structure−function relationships of lactoferricin derivatives. Helix propensity of positions 4−9 in the lactoferricin sequence was the most important in determining the antimicrobial activity of lactoferricin against *Escherichia coli*, followed by cationic charge pattern at positions 4−9 and 1−3. The pattern similarity of segments within sequences could be a useful tool for representing the distribution attributes of amino acid residue properties to the structure−function relationships of proteins and peptides, especially when used in conjunction with principal component similarity analysis followed by the regression version of artificial neural networks.

KEYWORDS: Lactoferricin; cationic antimicrobial peptide; quantitative structure−activity relationships; homology similarity analysis; principal component similarity; artificial neural network

INTRODUCTION

A potent peptide was first separated from pepsin hydrolysates of bovine lactoferrin (LFB) by Tomita et al. (*1*), during the investigation of antimicrobial components in cow's milk. This peptide, named bovine lactoferricin (LfcinB), was identified to be residues 17−41 of LFB, with the following sequence: ^{17}FKCRR WQWRM KKLGA PSITC VRRAF^{41}. Many functional studies of LfcinB have been published since then (*2*−*11*). However, the actual bioactive mechanisms of this peptide are still not fully understood.

Quantitative structure−activity relationships (QSAR) are an important aspect in the study of the functionality of bioactive peptides. A recent study by Siebert (*12*) reported that the three-z approach of Hellberg et al. (*13*) was useful in modeling the functional behavior of peptides as a function of amino acid composition. Using the three-z method and partial least squares (PLS) regression analysis, it was possible to predict the antimicrobial activity of 15-residue murine Lfcin derivatives on the basis of the three major principal component (PC) scores (three z's) obtained from principal component analysis (PCA) (*9*). However, the same group of researchers (*14*) emphasized the importance of the position of specific amino acid residues in a peptide sequence in the physicochemical behavior of the peptide, a factor that was not fully taken into consideration in their QSAR computation.

In an exhaustive QSAR study of protein sequences, Klein et al. (*15*) discovered the importance of using attributes relating to hydrophobicity, charge, and their distributions as represented by frequency of occurrence and periodicity of appearance, in addition to propensity of secondary structure, for the accurate classification of functionality. However, there has been no report on how to represent the distribution of property attributes of amino acid side chains in protein sequences.

Another limitation of the three-z method is the difficulty of visualizing and assessing the effects of several PC scores simultaneously through two-dimensional scattergrams. This limitation has been overcome using principal component similarity (PCS) analysis (*16*). LfcinB belongs to the class of peptides known as cationic antimicrobial peptides (CAPs), which could be further classified according to parameters such as helical propensity, specific side-chain properties, or the presence of cystine-rich groups (*17*). The PCS analysis was successfully used to classify several CAPs into those different groups according to the parameters of amino acids in their sequences (*18*). When the charge scale was used in the PCS analysis, LfcinB and indolicidin were shown to belong to the same low-charge group but also classified as having high helical content, in contrast to the charge dominant protamine group of CAP. However, in that study, PCS analysis was used as a classification method based on properties of the entire peptide sequence. Properties of segments within the sequences and distribution patterns of properties within the peptide sequences were not incorporated in that classification.

To recognize the importance of the position of amino acid residues within a peptide sequence, a new computer program was developed as shown in this study for comparing the homology "pattern similarity" of functionally similar segments

* Author to whom correspondence should be addressed [telephone (604) 822-4427; fax (604) 822-3959; e-mail nakai@interchange.ubc.ca].

10.1021/jf0206062 CCC: $25.00
Published on Web 01/28/2003

between different sequences to circumvent the drawback of the current QSAR method. It was anticipated that homology pattern similarity could be an effective way of expressing the distribution of characteristic property indices in sequences, to use as attributes in QSAR study. Definition of pattern similarity, which is different from that of dissimilarity, has already been discussed (*19*). It is worth noting that the pattern similarity used here is, therefore, different from the conventional algorithms for computing sequence alignment.

The objectives of the current study were to apply our novel computer program called "homology similarity analysis" (HSA) to elucidate the structural mechanisms for the antimicrobial activity of lactoferricin and its derivatives and to classify these peptides using the PCS approach applied to the results obtained from HSA. Artificial neural network analysis was then applied to demonstrate the accuracy of predicted antimicrobial activity and the relative importance of input variables on the antimicrobial activity of lactoferricin derivatives.

MATERIALS AND METHODS

Lactoferricin Sequences. Tables 1−5 list the sequences of lactoferricin and its derivatives, as well as their antimicrobial activities expressed as minimal inhibitory concentration (MIC) against different strains of *Escherichia coli*, namely, IID 861 (*2*), O111 (*3*), NCTC 8007 (*4*, *5*), and ATCC 25922 (*6*, *8*−*11*). Peptide 2 is the 25-residue peptide representing bovine lactoferricin, whereas peptide 13 is the corresponding 15-residue lactoferricin derivative that was used as a reference in the PCS computation in this study.

Amino Acid Property Scales. The amino acid property scales used for HSA computation are listed in **Table 6**. The properties of amino acid side chains used in this new computer program were hydrophobicity, charge, helix propensities, and bulkiness, as previously reported for optimization of site-directed mutagenesis (*20*). Helix propensity was described as the free energy required to fix the different amino acids in an α-helical region defined by two intervals of the Ramachandran plot (*21*). Helix propensity values ranged from 0.617 for alanine to 1.780 for proline, thus being opposite in relation to the propensity. The amino acid hydrophobicity scale statistically derived from a large set of experimental chromatographic retention data (*22*) was used, with values ranging from −2.24 for histidine to 4.80 for phenylalanine. The isoelectric points of the amino acids were used in a scale representing the charge of their side chains (*23*). Last, bulkiness values of amino acids derived from compressibility data (*23*) ranged from 3.40 for glycine (the least bulky) to 21.61 for tryptophan (the most bulky).

Homology Similarity Analysis. HSA was conducted on three selected segments of the peptide sequences, corresponding to residues 1−15 of bovine lactoferricin. It was reported that this 15-mer Lfcin derivative exhibited antimicrobial activity similar to that of the 25-residue native peptide (*8*); in addition, the hexamer corresponding to residues 4−9 was reported to be the antimicrobial core of Lfcin (*2*). Thus, in the present study, the sequences of lactoferricin derivatives were divided into three segments, namely, segment I (positions 1−3), segment II (positions 4−9), and segment III (positions 10−15). HSA was performed to examine the helical propensity in segment II (Hx II), charge in segments I (Ch I), II (Ch II), and III (Ch III), hydrophobicity in segment III (Hp III), and bulkiness in segment III (Bk III).

The specific property of amino acid residues in a designated segment within a sample peptide sequence was compared with that of amino acid residues belonging to the corresponding segment within the sequence of the reference peptide, that is, peptide 13 **Tables 1−5**). The property index values of the amino acid residues in the segment in sample sequence were plotted against the index values of amino acid residues in the corresponding segment in the reference sequence. This procedure was repeated for other segments of the same sample sequence. Linear regression was carried out, and the resultant coefficient of determination (r^2) was referred to as the homology similarity constant. Average values of the property index values within the segments were also calculated.

The software package for HSA used in this study can be downloaded from ftp://ftp.agsci.ubc.ca/foodsci/HSA.

Principal Component Similarity Analysis. The principle of the PCS program has been previously described (*16*). HSA similarity coefficients and average values of Hx II, Ch II, and Ch I were used during the PCS computation.

Missing values appearing as similarity coefficient values of zero occurred as a result of shorter peptides that were lacking amino acids at certain positions in segments being analyzed. These missing values were replaced by the mean values of all available data, but selecting only those individuals that have no missing values for calculation of the within-groups dispersion matrix (*24*).

The software package for PCS used in this study can be downloaded from ftp://ftp.agsci.ubc.ca/foodsci/SPCS.

Artificial Neural Networks (ANN). To focus on peptides that may have good antimicrobial activity, regression study using ANN was performed using the peptide with the second lowest MIC (peptide 25 with an MIC of 15 μg/mL) as a reference for HSA. Peptide 6 with the lowest MIC of 11 μg/mL was avoided because it lacks residues in segment I, which would introduce null values for the pattern similarity coefficients for other peptides compared to it. This assignment, which was different from that for the above HSA used for PCS, was made due to the difference in purposes of regression from that of classification. Group III peptides, which were observed to be exceptional by HSA−PCS analysis, were eliminated from the data set used for the regression analysis, thus comprising a total of 65 peptides. After PCA computation, five major PC scores, all with eigenvalues above 1.0, were selected for the subsequent ANN computation. Two output variables were used: log MIC and $1000 \times \mathrm{MIC}^{-1}$. STATISTICA Neural Networks (*25*) was used for the nonlinear regression computation. Sensitivity analysis was also performed, in which the data set was submitted to the network repeatedly for the training and verification subsets separately, to give information about the relative importance of the variables used in the neural network.

RESULTS

Homology Similarity Analysis of Lactoferricin Sequences. The results of HSA analysis of lactoferricin and its derivatives were classified into five groups (**Tables 1−5**).

Lactoferricin derivatives with high antimicrobial activity (low MIC) and HSA similarity coefficients for helical propensity in segment II that were very close to 1.0 (highly resembling to the reference peptide) were classified into group I (**Table 1**). Highly antimicrobial Lfcin derivatives, but being highly cationic in segment I ($P < 0.01$ vs group I by t test), were classified into group II (**Table 2**). Group III (**Table 3**) was the group exceptional to the general rules herein adopted, consisting of peptides with unexpectedly high or low antimicrobial activities that could not be readily correlated to structural attributes. Group IV (**Table 4**) consisted of Lfcin derivatives with low antimicrobial activity and low HSA coefficients for charge in segment II ($P < 0.01$ vs group I). Last, Lfcin derivatives in group V (**Table 5**) were also low in antimicrobial activity, but showing low or negative HSA coefficients for helical propensity in segment II ($P < 0.01$ vs group I).

Segment II (positions 4−9) in the sequences appears to play an important role in the antimicrobial activity of Lfcin derivatives against *E. coli*. Lfcin derivatives with higher similarity constant values for helical propensity in segment II demonstrated greater antimicrobial activity. When the helical pattern similarity

Table 1. Homology Similarity Coefficients and Average Property Values for Helix (Hx), Charge (Ch), Hydrophobicity (Hp), and Bulkiness (Bk) Calculated for Three Segments in the Sequences of Group I Lactoferricin Derivatives along with Minimum Inhibitory Concentration (MIC) against *E. coli* As Reported in the Literature[a]

Peptide #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
									PEW																
Segment I		FKC	FKS					FKC	FKC	FKC	KC		FKC	AKC	FAC	FKA	FKC	FKC	FKC	FKC	FKC	AKC	FKC	FKC	FK F
Segment II	RRWQWR	RRWQWR	RRWQWR	RRWQWR	KKWQWK	RRWQWR	KKWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWA	RRWQWR	RRWQWR	RRWQWR	RRWQWR	RRWQWR	LRWQWR	RRFQWF	RRWQWR	RRWQWR
Segment III		MKKLGA	MKKLGA	MKKLG	MKKLG	MRRLG	MKKLG	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MKKLGA	MRKLGG	MKKLGA	MKKLGA	MKKLGA
		PSITCVRRAF	PSITSVRRAF					PSITCVRRAF																	
MIC (μg / mL)	50	23	30	23	30	11	30	30	70	40	80	100	50	70	80	25	30	70	50	50	25	50	50	25	15
Literature #	2	3	3	3	3	3	3	6	6	6	6	6	8	8	8	8	8	8	8	8	8	9	10	10	11
Hx II																									
Coefficient	1.000	1.000	1.000	1.000	0.987	1.000	0.982	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.840	1.000	1.000	1.000	1.000	0.998	0.981	0.981	1.000
Average	0.808	0.808	0.808	0.808	0.824	0.808	0.819	0.808	0.808	0.808	0.808	0.808	0.808	0.808	0.808	0.808	0.783	0.808	0.808	0.808	0.808	0.806	0.818	0.818	0.808
Ch II																									
Coefficient	1.000	1.000	1.000	1.000	1.000	1.000	0.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.707	0.998	0.998	1.000
Average	8.400	8.400	8.400	8.400	7.850	8.400	8.033	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	8.400	7.600	8.317	8.317	8.400
Ch I																									
Coefficient	0.000	1.000	0.982	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.994	0.000	1.000	0.994	0.928	0.982	1.000	1.000	1.000	1.000	1.000	0.994	1.000	1.000	0.997
Average	6.000	6.767	7.067	6.000	6.000	6.000	6.000	6.767	6.767	6.767	6.933	6.000	6.767	6.933	5.533	7.067	6.767	6.767	6.767	6.767	6.767	6.933	6.767	6.767	6.900
Ch III																									
Coefficient	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.632	0.632	1.000	1.000	0.988	1.000	1.000	1.000
Average	6.000	7.233	7.233	7.233	7.233	7.600	7.233	7.233	7.233	7.233	7.233	7.233	7.233	7.233	7.233	7.233	7.233	6.617	6.617	7.233	7.233	7.417	7.233	7.233	7.233
Hp III																									
Coefficient	0.000	1.000	1.000	0.425	0.425	0.361	0.425	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.931	0.931	0.694	1.000	0.966	1.000	1.000	1.000
Average	-10.000	0.113	0.113	-1.563	-1.563	-1.307	-1.563	0.113	0.113	0.113	0.113	0.113	0.113	0.113	0.113	0.113	0.113	0.393	0.393	-0.460	0.098	-0.012	0.113	0.113	0.113
Bk III																									
Coefficient	0.000	1.000	1.000	0.832	0.832	0.834	0.832	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.959	0.959	0.763	0.882	0.879	1.000	1.000	1.000
Average	0.000	13.925	13.925	12.008	12.008	11.535	12.008	13.925	13.925	13.925	13.925	13.925	13.925	13.925	13.925	13.925	13.925	13.225	13.225	12.342	15.275	11.843	13.925	13.925	13.925

[a] Segment I = positions 1–3, segment II = positions 4–9, and segment III = positions 10–15 in the reference (peptide 13) with 15 amino acid sequence identical to bovine lactoferricin. Shaded zones in the sequences represent regions with high similarity to the reference peptide, whereas the shading of Hx II indicates the high similarity coefficient for helix propensity in segment II for group I derivatives.

Table 2. Homology Similarity Coefficients and Average Property Values for Helix (Hx), Charge (Ch), Hydrophobicity (Hp), and Bulkiness (Bk) Calculated for Three Segments in the Sequences of Group II Lactoferricin Derivatives along with Minimum Inhibitory Concentrations (MIC) against *E. coli* As Reported in the Literature[a]

Peptide #	26	27	28	29	30
Segment I	K	R	R	R	R
	K	K	K	K	K
	C	C	C	C	C
Segment II	R	L	L	L	L
	R	R	R	R	R
	W	W	W	W	W
	Q	Q	Q	Q	Q
	W	W	W	W	W
	R	R	E	A	R
Segment III	M	M	M	M	M
	K	R	R	R	R
	K	K	K	K	K
	L	V	Y	Y	Y
	G	G	G	G	G
	A	G	G	G	G
MIC (μg / mL)	60	20	40	25	20
Literature #	6	9	9	9	9
Hx II					
Coefficient	1.000	0.998	0.955	0.892	0.998
Average	0.808	0.806	0.793	0.783	0.806
Ch II					
Coefficient	1.000	0.707	0.149	0.447	0.707
Average	8.400	7.600	6.333	6.800	7.600
Ch I					
Coefficient	0.566	0.407	0.407	0.407	0.407
Average	8.167	8.533	8.533	8.533	8.533
Ch III					
Coefficient	1.000	0.988	0.988	0.988	0.988
Average	7.233	7.417	7.417	7.417	7.417
Hp III					
Coefficient	1.000	0.950	0.966	0.966	0.966
Average	0.113	-0.062	-0.012	-0.012	-0.012
Bk III					
Coefficient	1.000	0.900	0.879	0.879	0.879
Average	13.925	12.422	11.843	11.843	11.843

[a] Segment I = positions 1–3, segment II = positions 4–9, and segment III = positions 10–15 in the reference (peptide 13) with 15 amino acid sequence identical to bovine lactoferricin. Shaded zones in the sequences represent regions with high similarity to the reference peptide, whereas the shading of Ch I indicates the high average values for charge in segment I characteristic of group II derivatives.

Table 3. Homology Similarity Coefficients and Average Property Values for Helix (Hx), Charge (Ch), Hydrophobicity (Hp), and Bulkiness (Bk) Calculated for Three Segments in the Sequences of Group III Lactoferricin Derivatives along with Minimum Inhibitory Concentrations (MIC) against *E. coli* As Reported in the Literature[a]

Peptide #	31	32	33	34	35
Segment I		F		F	F
		K		K	K
		C	C	C	C
Segment II	R	F	R	R	R
	R	R	R	R	R
	W	W	W	W	F
	Q	Q	Q	Q	Q
	R	W	W	W	F
	W	R	R	R	R
Segment III	M	M	M	A	M
	K	K	K	K	K
	K	K	K	K	K
	L	L	L	L	L
	G	G	G	G	G
		A	A	A	A
MIC (μg / mL)	23	20	200	140	300
Literature #	3	6	6	8	10
Hx II					
Coefficient	0.213	0.512	1.000	1.000	1.000
Average	0.808	0.844	0.808	0.808	0.828
Ch II					
Coefficient	0.333	0.651	1.000	1.000	0.998
Average	8.400	7.517	8.400	8.400	8.233
Ch I					
Coefficient	0.000	1.000	0.566	1.000	1.000
Average	6.000	6.767	5.700	6.767	6.767
Ch III					
Coefficient	1.000	1.000	1.000	1.000	1.000
Average	7.233	7.233	7.233	7.233	7.233
Hp III					
Coefficient	0.425	1.000	1.000	0.999	1.000
Average	-1.563	0.113	0.113	0.088	0.113
Bk III					
Coefficient	0.832	1.000	1.000	0.947	1.000
Average	12.008	13.925	13.925	13.133	13.925

[a] Segment I = positions 1–3, segment II = positions 4–9, and segment III = positions 10–15 in the reference (peptide 13) with15 amino acid sequence identical to bovine lactoferricin.

was low in segment II (group V), Lfcin derivatives exhibited lower inhibitory effects against *E. coli* ($P < 0.01$ vs group I). Conversely, lactoferricin derivatives with high similarity in helical propensity in segment II (group I) exhibited higher antimicrobial effect. However, it is important to remember that high similarity in helical propensity of derivatives compared to the reference peptide is not necessarily correlated with high average helical propensity values per se.

In addition to the helical pattern, the pattern of charge distribution in segment II is of significance from the aspect of antimicrobial ability of Lfcin derivatives. Although peptides in group IV exhibited a similar helical pattern in segment II to the reference, they exhibited lower antimicrobial activity ($P < 0.01$ vs group I). The low pattern similarity in charge distribution and lower average cationic values in segment II of these peptides in group IV might have been responsible for their higher MIC against *E. coli.* In contrast, high cationic values in segment I could compensate for low cationic values in segment II, thereby leading to improved antimicrobial activity, as illustrated by the peptides in group II.

With a few exceptions, homology similarities of the charge, hydrophobicity, and bulkiness patterns in segments III did not appear to play an important role in MIC of Lfcin and its derivatives.

Principal Component Similarity Analysis. On the basis of the information recovered from HSA computation, the major structural factors that influenced the MIC values of Lfcin derivatives were the similarity constants and average values for helical propensity in segment II, charge in segment II, and charge in segment I. Thus, these six variables were used for PCA followed by PCS analysis. The purpose of this analysis was to find properties of amino acid side chains, which were playing an important role in grouping of lactoferricin derivatives based on their sequences. Information on this grouping is important in helping to explain the underlying principle of antimicrobial activity of lactoferricins.

As shown in **Figure 1**, peptides were successfully classified in groups I, II, IV, and V. These results confirm the importance of helicity and charge in the antimicrobial activity of lactoferricin. Group III peptides having exceptionally high or low antimicrobial activity were found to overlap with group I (peptides 33−35) and group IV (peptides 31 and 32), respectively, although the structural basis for the irregular activity of these group III peptides is unknown.

Regression Analysis Using Artificial Neural Networks. ANN were used successfully to predict the output variables, that is, logarithmic or log MIC and 1000/MIC or reciprocal MIC, with correlation coefficients of 0.95−0.97 and 0.88−0.90, respectively. Predicted values versus observed values are

Table 4. Homology Similarity Coefficients and Average Propensity Values for Helix (Hx), Charge (Ch), Hydrophobicity (Hp), and Bulkiness (Bk) Calculated for Three Segments in the Sequences of Group IV Lactoferricin Derivatives along with Minimum Inhibitory Concentrations (MIC) against *E. coli* As Reported in the Literature[a]

Peptide #	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
Segment I		E	E	S	F	F	F	E	S	E	A	E	A	R	E	A	R	A	E	A	E
		K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Segment II	E	L	L	R	A	R	R	L	R	L	L	L	L	L	L	L	L	L	L	L	L
	E	R	R	Q	R	A	R	R	Q	R	R	R	R	R	R	R	R	R	R	R	R
	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
	W	N	N	S	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
	E	E	E	K	R	R	A	E	K	E	E	A	A	E	R	R	A	E	A	A	R
Segment III	M	M	M	I	M	M	M	M	I	M	M	M	M	M	M	M	M	M	M	M	M
	E	R	R	R	K	K	K	R	R	R	R	R	R	R	R	R	R	R	R	R	R
	E	K	K	R	K	K	K	K	R	K	K	K	K	K	K	K	K	K	K	K	K
	L	V	V	T	L	L	L	V	T	V	V	V	V	V	V	V	V	Y	Y	Y	Y
	G	G	G	N	G	G	G	G	N	G	G	G	G	G	G	G	G	G	G	G	G
		G	G	P	A	A	A	G	P	G	G	G	G	G	G	G	G	G	G	G	G
		P																			
		P																			
		L																			
		S																			
		C																			
		V																			
		K																			
		K																			
		S																			
		S																			
MIC (μg / mL)	120	200	1000	1000	70	120	55	1000	440	1000	800	1000	270	75	200	62	62	300	1000	175	125
Literature #	3	6	8	8	8	8	8	8	8	9	9	10	10	10	10	10	10	10	10	10	10
Hx II																					
Coefficient	0.971	0.910	0.910	0.963	0.883	0.883	0.883	0.955	0.988	0.955	0.955	0.892	0.892	0.955	0.998	0.998	0.892	0.955	0.892	0.892	0.998
Average	0.769	0.823	0.823	0.828	0.786	0.786	0.786	0.793	0.816	0.793	0.793	0.783	0.783	0.793	0.806	0.806	0.783	0.793	0.783	0.783	0.806
Ch II																					
Coefficient	-1.000	0.191	0.191	0.698	0.707	0.707	0.707	0.149	0.698	0.149	0.149	0.447	0.447	0.149	0.707	0.707	0.447	0.149	0.447	0.447	0.707
Average	4.600	6.233	6.233	7.417	7.600	7.600	7.600	6.333	7.417	6.333	6.333	6.800	6.800	6.333	7.600	7.600	6.800	6.333	6.800	6.800	7.600
Ch I																					
Coefficient	0.000	0.933	0.933	0.994	1.000	1.000	1.000	0.933	0.994	0.933	0.994	0.933	0.994	0.407	0.933	0.994	0.407	0.994	0.933	0.994	0.933
Average	6.000	6.000	6.000	6.933	6.767	6.767	6.767	6.000	6.933	6.000	6.933	6.000	6.933	8.533	6.000	6.933	8.533	6.933	6.000	6.933	6.000
Ch III																					
Coefficient	-1.000	0.988	0.988	0.996	1.000	1.000	1.000	0.988	0.996	0.988	0.988	0.988	0.988	0.988	0.988	0.988	0.988	0.988	0.988	0.988	0.988
Average	5.067	7.417	7.417	7.500	7.233	7.233	7.233	7.417	7.500	7.417	7.417	7.417	7.417	7.417	7.417	7.417	7.417	7.417	7.417	7.417	7.417
Hp III																					
Coefficient	0.297	0.950	0.950	0.409	1.000	1.000	1.000	0.950	0.409	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.966	0.966	0.966	0.966
Average	-1.057	-0.062	-0.062	0.485	0.113	0.113	0.113	-0.062	0.485	-0.062	-0.062	-0.062	-0.062	-0.062	-0.062	-0.062	-0.062	-0.012	-0.012	-0.012	-0.012
Bk III																					
Coefficient	0.833	0.900	0.900	0.381	1.000	1.000	1.000	0.900	0.381	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.879	0.879	0.879	0.879
Average	11.298	12.422	12.422	16.002	13.925	13.925	13.925	12.422	16.002	12.422	12.422	12.422	12.422	12.422	12.422	12.422	12.422	11.843	11.843	11.843	11.843

[a] Segment I = positions 1–3, segment II = positions 4–9, and segment III = positions 10–15 in the reference (peptide 13) with 15 amino acid sequence identical to bovine lactoferricin. Shaded zones in the sequences represent regions with high similarity to the reference peptide, whereas the shading of Ch II indicates low similarity coefficient and average values of charge in segment II of group IV derivatives.

Table 5. Homology Similarity Coefficients and Average Property Values for Helix (Hx), Charge (Ch), Hydrophobicity (Hp), and Bulkiness (Bk) Calculated for Three Segments in the Sequences of Group V Lactoferricin Derivatives along with Minimum Inhibitory Concentrations (MIC) against *E. coli* As Reported in the Literature[a]

Peptide #	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Segment I	S	K	S	F	F	T	S							T	T
	K	K	K	K	K	K	K							K	K
	C	C	C	C	C	C	C							C	C
Segment II	Y	F	Y	R	R	F	Y	R	F	F	F	F	F	F	F
	Q	R	Q	R	R	Q	Q	R	Q	Q	Q	Q	Q	Q	Q
	W	W	W	A	W	W	W	A	W	W	W	W	W	W	W
	Q	Q	Q	Q	Q	Q	Q	A	Q	Q	Q	Q	Q	Q	Q
	R	W	R	W	A	W	W	A	R	R	R	R	R	R	R
	R	R	R	R	R	N	R	R	N	N	N	N	N	N	N
Segment III	M	M	M	M	M	M	M	A	M	M	M	M	P	M	M
	R	K	R	K	K	R	R	K	R	R	R	R	R	R	R
	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
	L	L	L	L	L	V	L	A	V	V	V	V	V	V	V
	G	G	G	G	G	R	G	G	R	R	R	R	R	R	R
	A	A	A	A	A	G	A		G		G			G	G
	P								P		P			P	
	S								P		S			P	
	I								V					S	
	T								S					C	
	C													I	
	V													K	
	R													R	
	R													D	
	T													S	
	S														
MIC (μg / mL)	750	200	500	200	200	150	350	120	1000	1000	1000	580	2000	200	1000
Literature #	6	6	8	8	8	8	8	3	4	4	5	5	5	6	8
Hx II															
Coefficient	0.007	0.512	0.007	0.040	0.041	0.000	0.393	-0.764	-0.297	-0.297	-0.297	-0.297	-0.297	-0.297	-0.297
Average	0.828	0.844	0.828	0.759	0.759	0.903	0.854	0.685	0.877	0.877	0.877	0.877	0.877	0.877	0.877
Ch II															
Coefficient	0.000	0.651	0.000	1.000	1.000	-0.703	0.447	1.000	-0.521	-0.521	-0.521	-0.521	-0.521	-0.521	-0.521
Average	7.600	7.517	7.600	8.400	8.400	5.817	6.800	8.400	6.617	6.617	6.617	6.617	6.617	6.617	6.617
Ch I															
Coefficient	0.994	0.566	0.994	1.000	1.000	0.994	0.994	0.000	0.000	0.000	0.000	0.000	0.000	0.994	0.994
Average	6.933	8.167	6.933	6.767	6.767	6.933	6.933	6.000	6.000	6.000	6.000	6.000	6.000	6.933	6.933
Ch III															
Coefficient	0.988	1.000	0.988	1.000	1.000	0.640	0.988	1.000	0.640	0.640	0.640	0.640	0.640	0.640	0.640
Average	7.417	7.233	7.417	7.233	7.233	8.217	7.417	7.233	8.217	8.217	8.217	8.217	8.217	8.217	8.217
Hp III															
Coefficient	0.987	1.000	0.987	1.000	1.000	0.915	0.987	0.173	0.915	0.265	0.915	0.265	0.263	0.915	0.915
Average	0.242	0.113	0.242	0.113	0.113	-0.228	0.242	-2.162	-0.228	-1.920	-0.228	-1.920	-1.837	-0.228	-0.228
Bk III															
Coefficient	0.995	1.000	0.995	1.000	1.000	0.477	0.995	0.658	0.477	0.431	0.477	0.431	0.438	0.477	0.477
Average	13.688	13.925	13.688	13.925	13.925	14.235	13.688	9.633	14.235	13.668	14.235	13.668	13.865	14.235	14.235

[a] Segment I = positions 1–3, segment II = positions 4–9, and segment III = positions 10–15 in the reference (peptide 13) with 15 amino acid sequence identical to bovine lactoferricin. Shaded zones in the sequences represent regions with high similarity to the reference peptide, whereas the shading of Hx II indicates low similarity coefficient values for helix propensity in segment II of group V derivatives.

illustrated in **Figure 2** with correlation coefficients of 0.950 and 0.900 for the output variables of log MIC and reciprocal MIC, respectively.

Log MIC values tended to neglect differences in MIC of the most potent antimicrobial peptides in group I. The smooth regression line near the origin of the reciprocal MIC regression as shown in **Figure 2b** may reflect reasonable prediction of high MIC values. The sensitivity ranks of the five input variables (PC scores 1−5) were 1:4:2:5:3 and 1:4:5:3:2 for log MIC and reciprocal MIC, respectively. This implies that the importance of the variables in the correlation is in this order. The sensitivity ratio for each of the input variables was >1.0, thereby indicating that none of them should be pruned from the network (*25*).

Loadings of PC scores 1−5 are shown in **Table 7**. The importance of helix structure is explicitly shown in the higher loading (greater size of the absolute values) of similarity constants as well as greater loading of average helix values. Charge distribution at segment II (positions 4−9) is also important, especially the similarity coefficient for PC4, whereas that at segment I (positions 1−3) is supplemental (**Table 7**) with high loadings for PC1, PC2, and PC5. In contrast, high loading of Ch III appears for only PC1. The sensitivity ranks obtained by ANN computation demonstrate the importance of helix propensity value with a supplemental contribution of charge of segments I−III, which coincides with what has been observed during the above HSA study.

DISCUSSION

In the pioneering work of Hellberg et al. (*13*) on peptide QSAR, 29 variables were employed to characterize amino acid side chains, of which dimension was reduced to three descriptors, z1, z2, and z3, using PCA to avoid the collinearity problem during the subsequent regression analysis. However, a new risk was introduced when potentially important conformation effects of peptides on the biological activity were ignored. To circumvent this shortcoming, Ström et al. (*9*) employed two parameters for α-helices measured for micelle affinity, three parameters describing α-helical propensities, two parameters related to charge, four hydrophobicity parameters, and one parameter

Table 6. Amino Acid Property Values Used for Homology Similarity Analysis (HSA)

	helix	charge	hydrophobicity	bulkiness
Ala	0.617	6.0	0.06	11.50
Arg	0.753	10.8	–0.85	14.28
Asn	1.089	5.4	0.25	12.82
Asp	0.932	2.8	–0.20	11.68
Cys	1.107	5.1	0.49	13.46
Gln	0.770	6.0	0.31	14.45
Glu	0.675	3.2	–0.10	13.57
Gly	1.361	6.0	0.15	3.40
His	1.034	7.6	–2.24	13.67
Ile	0.876	6.0	3.00	21.40
Leu	0.740	6.0	3.50	21.00
Lys	0.784	9.7	–1.62	15.70
Met	0.730	6.0	0.21	16.25
Phe	0.968	5.5	4.80	19.80
Pro	1.780	6.0	0.71	17.43
Ser	0.980	6.0	–0.62	9.47
Thr	1.053	6.0	0.65	15.80
Trp	0.910	6.0	2.29	21.61
Tyr	1.009	6.0	1.89	18.03
Val	0.940	6.0	1.59	21.50
lit. ref	21	*a*	22	23

[a] Isoelectric point.

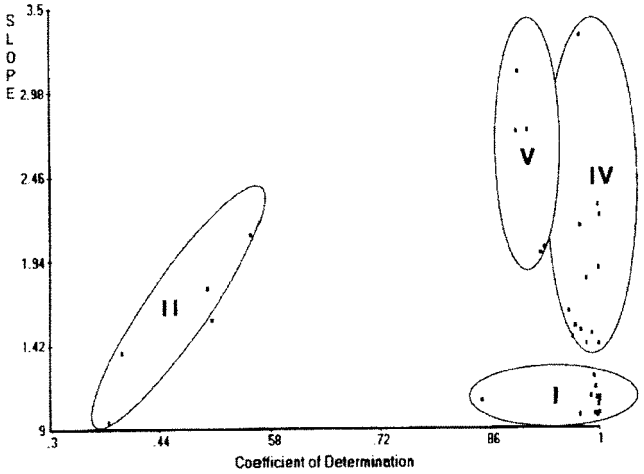

Figure 1. PCS scattergram of lactoferricin derivatives. Peptides classified in groups I–V are identified in **Tables 1**–**5**. Peptides in group III (not shown) included exceptional cases with unexplainable low and high MIC values that distributed between groups I and IV, respectively, near the reference point (coefficient of determination = 1.0 and slope = 1.0).

Table 7. Loadings of Input Variables[a]

	helix 4–9		charge 4–9		charge 1–3		charge 10–15	
	SC	av	SC	av	SC	av	SC	av
PC1	–0.55	0.51	–0.24	–0.01	0.35	0.15	0.47	0.09
PC2	0.42	–0.62	0.19	0.06	0.47	0.34	0.06	–0.21
PC3	–0.66	0.60	0.42	–0.02	0.04	0.10	–0.01	0.05
PC4	–0.49	0.44	0.64	0.06	0.03	0.25	–0.26	–0.08
PC5	–0.67	–0.20	–0.14	0.36	0.39	–0.27	–0.27	0.24

[a] SC, pattern similarity constant; av, average.

relating to surface. Although PCA can probably extract important information to be used for regression analysis for QSAR, it is difficult to identify and isolate the truly influencing factors in the underlying mechanism for antimicrobial activity of lactoferricin and its derivatives as different indices used for one property may simultaneously include components of other properties.

Unlike the previous studies of Hellberg et al. (*13*) and Strom et al. (*9*), the present study used only one scale for each side-chain property. Selection of the single index to describe each property was made with great care by choosing the most recently published index values that were also considered to be reliable with respect to their analytical principle. Furthermore, comparison of segments rather than entire sequences of the peptides allowed more detail in the underlying mechanisms of functions to be discovered because the distribution of property patterns within sequences is taken into consideration. As stated by Lejon et al. (*14*), the position of a specific amino acid in a sequence is important to peptide functions. Data of pattern similarity of segments within the sequence, rather than the sequence as a whole, could be an answer to the question they submitted.

When the reciprocal MIC was assigned as the output variable in ANN computation in contrast to log MIC, it was possible that better comparison of higher antimicrobial activities, for example, MIC < 100 μg/mL despite greater prediction error, could be made along with better comparisons between peptides of weaker activities. This could be useful for searching for greater antimicrobial activity of CAP. We did not intend to repeat the regression study made by Strøm et al. (*9*) to compare with our new approach used in this study, mainly because of the possible unreliability of literature values reported from different researchers. In addition, we were not excessively confident in the regression of all possible changes when chances of dramatic alteration in activity may not be excluded by simply replacing the amino acid residue at a single position with other amino acid residues. The low antimicrobial activity of peptide 35 (group III) may be attributed to the simultaneous replacement of two tryptophan (W) residues at positions 6 and 8 in segment II. In contrast, the corresponding single replacements of W made separately in peptides 23 and 24 in group I did not affect the antimicrobial activity. The specific function of tryptophan residues can be seen in indolicidin (*17*). Indolicidin activity was elucidated on the basis of the overall charge and amphipathic character of the extended helix (*26*). Unlike lactoferricin, charge is playing a more important role in indolicidin, but probably less important than for protamines (*18*).

The new approach employed in this study, namely, HSA, can incorporate potential distribution effects of side-chain properties into peptide sequences. It may be used to address the question posed by Lejon et al. (*14*) regarding the need for information about peptide sequence in the three-z approach. Nevertheless, it is extremely difficult to obtain complete understanding of the underlying mechanism of a function such as MIC, as observed in the exceptional cases that were classified as group III derivatives in this study (**Table 3**).

It is likely that unknown factors are still influential in the MIC data. At any rate, the accuracy in prediction of output variables could be improved, despite the potential errors arising from the output data of MIC values reported from different sources. Direct comparison of the correlation coefficient between log MIC and rcp MIC in **Figure 2** is not therefore warranted, as they use different scales. Furthermore, to enhance the validity of the conclusions, many more data are required for the reciprocal computation in the future.

Uniformity in group I, particularly in segment II, is demonstrated in the homology similarity coefficient values, which are all close to 1.0 as seen in **Table 1**. The lesser importance of segment III (positions 10−15) appears as a departure of the similarity values from 1.0 in some derivatives in this segment, as well as lesser contributions to PC scores 1−5 as represented

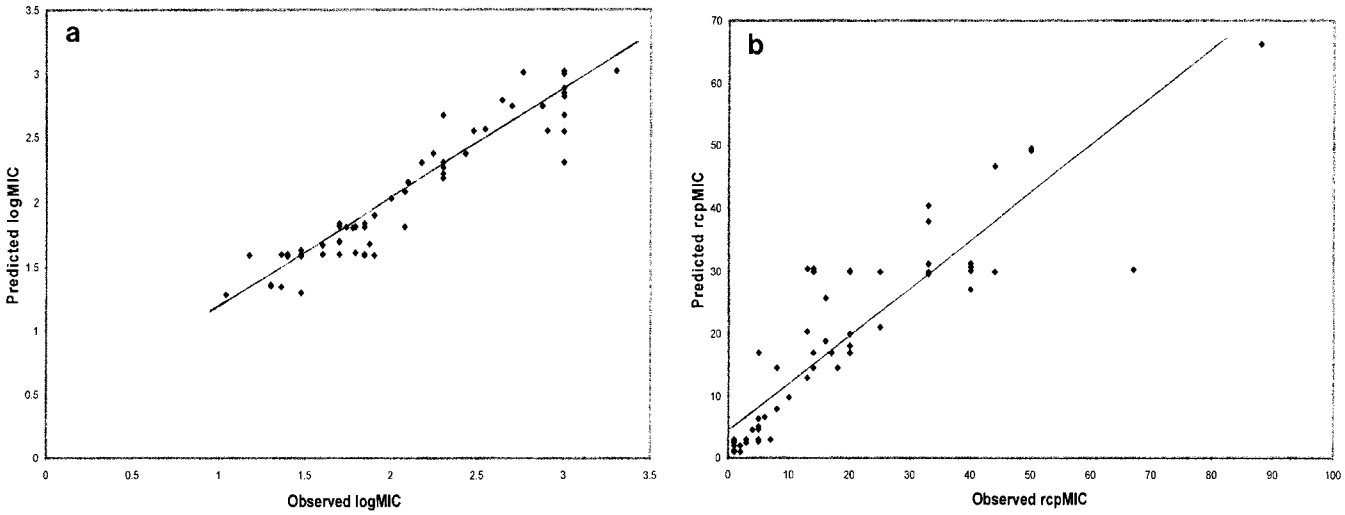

Figure 2. Predicted versus observed MIC values against *E. coli* of 65 different lactoferricin derivatives with (a) logarithmic (log MIC) or (b) reciprocal ($1000 \times MIC^{-1}$ or rcp MIC) transformation, as determined by ANN.

in the loadings (**Table 7**). This does not imply the absence of a role of segments I and III in MIC. Examples of the importance of these segments in particular derivatives are the smaller charge value of segment I in peptide 33 (classified into group III), the lower similarity in hydrophobicity of segment III in peptides 36 and 39 (classified into group IV), and the lower bulkiness similarity of segment III in peptides 34 and 44 (in groups III and IV, respectively). Peptide 31 (in group III) with multiple lower similarities may also be another example of these deviations.

To avoid the most crucial problem of ANN, namely, overlearning, the number of input variables should be restrained to the minimum. The most popular technique used for dimension reduction has been PCA, as shown by Strøm et al. (*9*). However, as discussed by Lejon (*14*), there is a further problem in that the selected major PC scores may not be always closely related to the objective function (MIC in this case). The same is true in the case of the PCS result shown in **Figure 1**, which is a classification result without considering the relation with MIC. The result of no appreciable difference between the SD ratios for training and verification subsets obtained in this study (0.26−0.33 vs 0.38−0.46 and 0.46−0.55 vs 0.26−0.4, for log MIC and rcp MIC, respectively) may provide evidence for the absence of overlearning in our ANN computation. The SD ratio is (prediction error SD)/(data SD). An SD ratio of significantly <1.0 is an index to demonstrate the effectiveness of ANN (*25*).

Because the literature values of MIC reported by different researchers from different laboratories were used in this study, the reliability of the observed trends is dependent on the consistency of these reported MIC values. However, this problem may be common as far as large databases are the source of data used for data mining. Hence, the regression results shown in this study should be regarded as an approximation of the trend, and a more detailed interpretation may have to wait until ample reliable data are collected based on collaborative efforts among researchers in the future. In a previous paper (*18*), we proposed a system of data mining (DM) for QSAR purposes as PCS → (RCO → PCS) → ANN, where RCO was random-centroid optimization (*27*). By adding HSA to the end of the first PCS in the above DM system, it could be more effective in elucidating the functional mechanisms as it provides more detail in the contribution of major PC scores to function as shown in this study.

In conclusion, the new DM system using pattern similarity data based on properties of segment amino acid side chains in homology profiles can enhance the reliability in predictability of QSAR compared with QSAR computed using the properties of the whole sequences.

ABBREVIATIONS USED

ANN, artificial neural networks; CAP, cationic antimicrobial peptides; Ch, charge; DM, data mining; Hp, hydrophobicity; HSA, homology similarity analysis; Hx, helix; LFB, bovine lactoferrin; Lfcin, lactoferricin; MIC, minimal inhibitory concentration; PCA, principal component analysis; PCS, principal component similarity; PLS, partial least squares; QSAR, quantitative structure−activity relationship; RCO, random-centroid optimization.

LITERATURE CITED

(1) Tomita, M.; Bellamy, W.; Takase, M.; Yamauchi, K.; Wakabayashi, H.; Kawase, K. Potent antibacterial peptides generated by pepsin digestion of bovine lactoferrin. *J. Dairy Sci.* **1991**, *74*, 4137−4142.

(2) Tomita, M.; Takase, M.; Bellamy, W.; Shimamura, S. A review: The active peptide of lactoferrin. *Acta Paedeatr. Jpn.* **1994**, *36*, 585−591.

(3) Kang, J. H.; Lee, M. K.; Kim, K. L.; Hahm, K.-S. Structure−biological activity relationships of 11-residue highly basic peptide segment of bovine lactoferrin. *Int. J. Pept. Protein Res.* **1996**, *48*, 357−363.

(4) Odell, E. W.; Sarra, R.; Foxworthy, M.; Chappele, D. S.; Evanse, R. W. Antibacterial activity of peptides homologous to a loop region in human lactoferrin. *FEBS Lett.* **1996**, *382*, 175−178.

(5) Chapple, D. S.; Mason, D. J.; Joannou, C. L.; Odell, E. W.; Gant, V.; Evans, R. W. Structure−function relationship of antibacterial synthetic peptides homologous to a helical surface region on human lactoferrin against *E. coli* serotype O111. *Infect. Immunol.* **1998**, *66*, 2434−2440.

(6) Rekdal, Ø.; Andersen, J.; Vorland, L. H.; Svendsen, J. S. Construction and synthesis of lactoferricin derivatives with enhanced antibacterial activity. *J. Pept. Sci.* **1999**, *5*, 32−45.

(7) Wakabayashi, H.; Matsumoto, H.; Hashimoto, K.; Teraguchi, S.; Takase, M.; Hayakawa, H. N-Acylated and D enantiomer derivatives of a nonamer core peptide of lactoferricin B showing improved antimicrobial activity. *Antimicrob. Agents Chemother.* **1999**, *43*, 1267−1269.

(8) Strøm, M. B.; Rekdal, Ø.; Svendsen, J. S. Antibacterial activity of 15-residue lactoferricin derivatives. *J. Pept. Res.* **2000**, *56*, 265−274.
(9) Strøm, M. B.; Rekdal, Ø.; Stensen, W.; Svendsen, J. S. Increased antibacterial activity of 15-residue murine lactoferricin derivatives. *J. Pept. Res.* **2001**, *57*, 127−139.
(10) Haug, B. E.; Svendsen, J. S. The role of tryptophan in the antibacterial activity of a 15-residue bovine lactoferricin peptide. *J. Pept. Sci.* **2001**, *7*, 190−196.
(11) Haug, B. E.; Skar, M. L.; Svendsen, J. S. Bulky aromatic amino acids increase the antibacterial activity of 15-residue bovine lactoferricin derivatives. *J. Pept. Sci.* **2001**, *7*, 425−432.
(12) Siebert, K. Quantitative structure−activity relationship modeling of peptide and protein behavior as a function of amino acid composition. *J. Agric. Food Chem.* **2001**, *49*, 851−858.
(13) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure−activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.
(14) Lejon, T.; Strøm, M. B.; Svendsen, J. S. Is information about peptide sequence necessary in multivariate analysis? *Chemom. Intell. Lab. Syst.* **2001**, *57*, 93−95.
(15) Klein, P.; Jacquez, I. A.; Delis, C. Prediction of protein function by discriminant analysis. *Math. Biosci.* **1986**, *81*, 177−189.
(16) Vodovotz, Y.; Arteaga, G. E.; Nakai, S. Principal component similarity analysis for classification and its application to GC data of mango. *Food Res. Int.* **1993**, *26*, 355−363.
(17) Hwang, P. M.; Vogel, H. J. Structure−function relationships of antimicrobial peptides. *Biochem. Cell Biol.* **1998**, *76*, 235−246.
(18) Nakai, S.; Ogawa, M.; Nakamura, S.; Dou, J.; Funane, K. A computer-aided strategy for structure−function relation study of food proteins using unsupervised data mining. *Int. J. Food Prop.* **2003**, *6*, 25−47.
(19) Nakai, S.; Amantea, G.; Nakai, H.; Ogawa, M.; Kanagawa, S. Definition of outliers using unsupervised principal component similarity analysis for sensory evaluation of foods. *Int. J. Food Prop.* **2002**, *5*, 289−306.
(20) Nakai, S.; Nakamura, S.; Scaman, C. H. Optimization of site-directed mutagenesis. Application of random-centroid optimization to one-site mutation of *B. stearothermophilus* neutral protease for improving thermostability. *J. Agric. Food Chem.* **1998**, *46*, 1655−1661.
(21) Muñoz, V.; Serrano, L. Intrinsic secondary structure propensities of the amino acids, using statistical $\phi-\varphi$ matrices: Comparison with experimental scales. *Proteins* **1994**, *20*, 301−311.
(22) Wilce, M. C. J.; Aguilar, M.-I.; Heam, M. T. Physicochemical basis of amino acid hydrophobicity scales: Evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides. *Anal. Chem.* **1995**, *67*, 1210−1219.
(23) Gromiha, M. M.; Ponnuswamy, P. K. Relationship between amino acid properties and protein compressibility. *J. Theor. Biol.* **1993**, *165*, 87−100.
(24) Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Clarendon Press: Oxford, U.K., 1988; pp 24−32.
(25) Statsoft. Addendum for version 4. In *Statistica Neural Networks*; Statsoft, Tulsa, OK, 1999; www.statsoft.com.
(26) Falla, R. J.; Hancock, R. E. W. Improved activity of a synthetic indolicidin analog. *Antimicrob. Agents Chemother.* **1997**, *41*, 771−775.
(27) Nakai, S.; Dou, J.; Lo, K. V.; Scaman, C. H. Optimization of site-directed mutagenesis. 1. New random-centroid optimization program for Windows used in research and development. *J. Agric. Food Chem.* **1998**, *46*, 1642−1654.

Received for review May 29, 2002. Revised manuscript received December 3, 2002. Accepted December 16, 2002. This work was financially supported by a Multidisciplinary Network Group Research Program, "Structure−Function of Food Macromolecules" (Dr. Rick Yada of the University of Guelph is the principal investigator), of the Natural Sciences and Engineering Council of Canada.

JF0206062