JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Modeling Physicochemical Properties and Activity of Aspartyl Proteinases Based on Amino Acid Composition

LEI NIE[†] AND KARL J. SIEBERT*

Department of Food Science and Technology, Cornell University, Geneva, New York 14456

A data set containing physicochemical properties and enzymatic activity measurements of aspartyl proteinases was employed for quantitative structure−property relationship (QSPR) and quantitative structure−activity relationship (QSAR) modeling based on either three or five amino acid principal property sums. All but one of the models based on five principal properties were stronger than those based on three properties. Models of zeta potential ($R^2 = 0.846$), circular dichroism ($R^2 = 0.638$), Bigelow average hydrophobicity ($R^2 = 0.692$), accessible surface area ($R^2 = 0.897$), and two dye-based assessments of hydrophobicity ($R^2 = 0.581$ and 0.595) were constructed. Model quality was evaluated by cross-validation and permutation. The amino acids most influential for each modeled property were identified. It is clearly possible to model physicochemical properties of proteins as a function of amino acid principal property sums. Surprisingly, it was also possible to model an enzyme activity ratio (milk clotting/proteolytic activity) in the same manner ($R^2 = 0.699$).

KEYWORDS: Principal properties; proteolytic activity; QSAR

INTRODUCTION

Proteins and peptides are responsible for a huge range of biological functions. They act as enzymes, hormones, permeases, transport systems, antibodies, and bacteriocins and are an integral part of cell membrane and cell wall structures, among others. Proteins and peptides contribute functional properties in foods such as solubility, wettability, viscosity, gelation, fat binding, water binding, emulsification, and foam, glass, and film formation (*1*−*3*).

Many of the biological functions of proteins are due to a small region of the molecule (e.g., the active site in an enzyme or the binding site in an antibody or hormone). In contrast, the physicochemical attributes are bulk properties that depend on the overall nature of a protein. For example, proteins with a high proportion of nonpolar amino acids are more hydrophobic. The physicochemical properties, singly or in combination, are responsible for protein functional properties in foods (*4*−*7*). Many relationships or models between protein physicochemical and functional properties have been described. Because the physicochemical properties of a peptide are due to its amino acid composition and the functional properties are a result of the physicochemical properties, it should in principle be possible to directly model functional properties from amino acid composition (*8*).

To construct models of any sort from amino acid properties it is necessary to somehow parametrize the latter. In other words, the dissimilar properties of the different amino acids need to be expressed on a common set of scales and, preferably, on a relatively small number of scales.

It has long been desired to model biological properties of peptides from their amino acid constituents. A number of approaches to modeling biological activity from amino acid composition have been developed. Sneath used principal component analysis (PCA) of amino acid physicochemical properties to arrive at four scales that were useful for qualitatively selecting amino acid alterations that would greatly alter the biological activity of oxcytocin−vasopressin and hypertensin (*9*).

Hellberg and co-workers used PCA to construct three principal property scales for amino acids based on 29 physicochemical variables (*10*). These were designated z_1, z_2, and z_3 and essentially represented lipophilicity, molecular size, and electronic properties, respectively. They were successfully used to model several biological activities as a function of the principal property values of amino acids in two or three positions that were varied in oxcytocin analogues, pepstatin analogues, and bradykinin potentiating pentapeptides. This modeling approach enables predictions of behavior from the model for amino acid substitutions other than those used to construct the model.

Jonsson et al. used the principal properties of Hellberg et al. to successfully model two very different biological activities, the bitterness of dipeptides and the bradykinin potentiating activity of pentapeptides, both as a function of amino acid sequence. In each case there was a term for each of the three principal properties at each position in the peptide, so for the dipeptides there were 6 terms and for the pentapeptides 15 terms (*11*).

* Author to whom correspondence should be addressed [telephone (315) 787-2299; fax (315) 787-2284; e-mail kjs3@cornell.edu].

[†] Permanent address: School of Pharmaceutical Sciences, Shandong University, Jinan, P. R. China 250012.

10.1021/jf802971n CCC: $40.75 © 2009 American Chemical Society
Published on Web 02/16/2009

Collantes and Dunn used two different amino acid properties, the isotropic surface area and electronic charge index, to model the bradykinin potentiating activity of pentapeptides, the bitterness of dipeptides, and inhibition of angiotensin by dipeptides (*12*).

Sandberg et al. (*13*) carried out PCA with a larger set of amino acids and properties than used by Hellberg et al. They computed five rather than three principal properties, which they called the *zz*-scales or extended *z*-scales. The first three of these corresponded largely to the original *z*-scales of Hellberg. The fourth property was positively related to heat of formation and negatively to electronegativity. The fifth property was positively related to electrophilicity and negatively to polarizability. These five principal properties were used to model analogues of elastase substrates with two amino acid positions varied and of neurotensin with three positions varied.

Almost all of the aforementioned work was either based on sequence modeling of short peptides or modeling substitutions of amino acids in short or modest length peptides at only two or three positions. This is because the number of terms required with this type of modeling increases by a factor of 2−5 for each additional amino acid that is varied. Sequence modeling of longer peptide chains is impractical because of the large number of variants needed to fit models with so many terms.

Siebert showed that in some special cases in which the proportion of one or a few amino acids in a peptide can explain a property it was possible to model the property as a function of the contribution of each relevant amino acid to each of the principal properties (*14*). These were computed by multiplying the number of moles of each relevant amino acid in a peptide by one of its *z*-scores and algebraically summing together to arrive at a *z*-score sum, or *z*-sum (Σz_i):

$$\sum z_i = \sum_{a=1}^{k} n_a z_{ia} \tag{1}$$

a indicates the identity of an amino acid, n_a refers to the number of moles of the amino acid in the peptide, and z_{ia} is the *i*th *z*-score for amino acid *a*. For example, the z_1-sum of the amino acids contributing to light absorbance at 280 nm is computed by algebraically summing the products of the number of moles of tyrosine, tryptophan, and cysteine ($k = 3$) each multiplied by the z_1 score for that amino acid. The sums for z_2 and z_3 were computed similarly. The property UV molar absorptivity at 280 nm was then modeled as a function of the *z*-sums:

$$A = b_0 + b_1 \sum z_1 + b_2 \sum z_2 + b_3 \sum z_3 \tag{2}$$

or

$$A = b_0 + \sum_{i=1}^{3} b_i \sum z_i \tag{3}$$

This approach was also successful in modeling the Coomassie brilliant blue dye binding response of proteins as a function of the *z*-sums of the basic and aromatic amino acids.

Siebert extended the approach to computing *z*-sums based on the contributions of all 20 coded amino acids using both the original three *z*-scores and the five extended *z*-scores (*8*), called here *z*′-scores. These were used successfully to model both physicochemical properties (hydrophobicity and viscosity) and a functional property (foaming) of a set of proteins.

It was of interest to see if the *z*-sum approach would be suitable for modeling additional physicochemical or functional properties or for other protein data sets. The objective of the current study was to see if successful models of proteinase

Table 1. Physicochemical and Proteolytic Activity Properties of Proteinases at pH 6.3 from Yada and Nakai (*15*)

						$[\theta]_{MRW\lambda}$(×10^{-3})[f]										
protein	*H*(av)[a]	ASA[b]	ANS[c]	CPA[d]	ZP[e]	190	193	198	200	202	210	213	222	224	225[h]	MC/PA[h]
M. miehei proteinase	1109	13000	2	21	−32.9	313	−6	62	−356	−1286	−2811	−2820	−1575	−1168	−952	
E. parasitica proteinase	923	12900	7	113	−26.1	1141	1414	385	26	−836	−2492	−2377	−1462	−1193	−1050	71.05
chymosin	1120	11300	48	96	−14	8912	8913	2788	−1125	−4610	−10460	−9577	−8393	−7952	−7656	60.36
pepsin	1063	12300	1	6	−34.6	2684	4094	2051	−47	−2398	−7144	−7059	−4953	−3893	−3402	95.75
M. pusillus proteinase	1041	30600	7	3	−16.9	17	−496	−680	−1091	−1888	−3028	−2948	−1532	−990	−742	47.43
A. saitoi proteinase	973	12200	6	73	−5.79	145	247	−957	−1378	−1821	−2085	−2019	−1244	−972	−821	63.37
penicillopepsin	933	11600	3	23	−39.5	−494	214	238	−704	−1352	−2419	−2114	−295	97	211	0
trypsin	1034	9600	12	6	15.77	−1793	−3050	−3185	−3365	−3560	−2945	−2568	−1412	−1279	−1170	0.75
chymotrypsin	1030	9800	5	9	16.27	4110	870	−6204	−8830	−9730	−7270	−6200	−4260	−4110	−4110	0.02
papain	1159	8700	12	19	19.44	14820	7590	−6990	−8880	−8165	−12580	−11590	−11700	−11700	−11640	3.07

[a] $H_{(av)}$, Bigelow average hydrophobicity. [b] ASA, accessible surface area. [c] ANS, apparent surface hydrophobicity using 1-anilino-8-naphathalenesulfonate. [d] CPA, apparent surface hydrophobicity using *cis*-parinaric acid. [e] ZP, ζ potential. [g] Wavelengths in nanometers. [f] $[\theta]_{MRW\lambda}$ refers to the molar ellipticity per residue, and its unit is deg cm^2 dmol^{-1}. [h] MC/PA, milk clotting to proteolytic activity ratio.

Table 2. Three-*z*-Scale Sums and Five-*z*′-Scale Sums of Proteinases in **Table 1**

	three-*z*-scale sums			five-*z*′-scale sums				
protein	Σz_1	Σz_2	Σz_3	$\Sigma z_1'$	$\Sigma z_2'$	$\Sigma z_3'$	$\Sigma z_4'$	$\Sigma z_5'$
M. miehei proteinase	6.09	−284.9	−11.03	9.97	−252.6	14.48	−231.8	88.27
E. parasitica proteinase	17.3	−397.12	−32.94	31.55	−353.33	3.99	−245.86	72.14
chymosin	−6.8	−212.08	−21.28	−11.33	−185.38	−14.53	−192.53	59.49
pepsin	−1.85	−314.14	43.03	12.01	−273.31	44.83	−277.6	72.58
M. pusillus proteinase	58.39	−292.46	−3.44	59.61	−251.98	21.33	−247.1	81.87
A. saitoi proteinase	98.17	−307.25	−8.08	99.73	−268.07	17.37	−261.39	77.00
penicillopepsin	76.00	−342.39	7.18	81.24	−289.00	29.53	−259.39	95.05
trypsin	16.1	−192.75	19.67	12.46	−181.31	11.79	−119.75	38.57
chymotrypsin	23.21	−234.88	−26.68	27.15	−226.19	−14.44	−121.38	51.51
papain	55.48	−137.05	−40.52	42.00	−118.22	−37.21	−86.52	18.93

activity and physicochemical properties could be constructed using either three or five *z*-sums for all of the amino acids in a protein.

MATERIALS AND METHODS

The data used were reported by Yada and Nakai (*15*). This data set contains physicochemical properties and proteinase activity measurements of 10 aspartyl proteinases. The authors calculated Bigelow average hydrophobicity (*H*(av)) (*16*) and accessible surface area (ASA). Observations of apparent surface hydrophobicity using 1-anilino-8-naphthalenesulfonate (ANS), apparent surface hydrophobicity using *cis*-parinaric acid (CPA), zeta potential (ZP), and the molar ellipticity per residue ($[\theta]_{\mathrm{MRW}\lambda}$) at 10 wavelengths were made at six different pH values, as was the milk clotting to proteolytic activity ratio (MC/PA). In $[\theta]_{\mathrm{MRW}\lambda}$ MRW is the mean residue weight and λ is the wavelength of the observation. The data from a single pH (6.3) were used in the current study (see **Table 1**). Some of the observations (*H*(av), ASA, ANS and CPA) were made only at pH 6.3.

Amino acid sequences for the proteinases were obtained from the Swiss-Prot Protein Knowledgebase (Swiss Institute for Bioinformatics (SIB) and European Bioinformatics Institute) (http://www.ebi.ac.uk/swissprot/)). The sequence was analyzed using the SIB ExPASy (Expert Protein Analysis System) proteomics server (http://www.expasy.ch/) to obtain the amino acid composition, that is, the number of moles of each amino acid in the protein sequence. The three-*z*-scale and five-*z*′-scale values for each amino acid were obtained from refs *10* and *13*, respectively. The term $\sum z_i$ (or $\sum z_i'$), which is the algebraic sum of the three-*z*-scale (or five-*z*′-scale) value for each amino acid multiplied by the corresponding number of moles of that amino acid in the protein sequence and then summed, is represented as eq 4

$$\sum z_i = \sum_{a=1}^{20} n_a z_{ia} \text{ or } \sum z_\mathrm{i}' = \sum_{a=1}^{20} n_a z_{ia}' \qquad (4)$$

where n_a refers to the number of moles of an amino acid in the proteinase and *a* indicates the identity of the amino acid. The index *i* represents which of the three-*z*-scores ($i = 1-3$) or five-*z*′-scores ($i = 1-5$) is represented. The variables obtained are referred to as *z*-sums or *z*′-sums and were used for modeling.

The physicochemical properties or enzymatic activities of proteinases were modeled as a function of amino acid composition expressed as *z*-sums by partial least-squares regression (PLSR) using the SIMCA-S 6.01 computer program (Umetrics Inc., Kinnelon, NJ), see eq 2. PCA was also carried out with the SIMCA-S program.

RESULTS AND DISCUSSION

In the study reported by Yada and Nakai (*15*), the Bigelow hydrophobicity and accessible surface area were calculated for 10 aspartyl proteases by the approaches of Bigelow (*16*) and Janin (*17*), respectively. Circular dichroism, zeta potential, hydrophobicity using two different fluorescent probes, and two observations of enzyme activity (milk clotting and proteolysis) were measured for the same proteases. The physicochemical and proteolytic activity properties of the proteinases determined by Yada and Nakai are listed in **Table 1**. They carried out PCA using all of the data and mapped the relationships of the proteases in the principal component space. No modeling of properties was carried out.

The amino acid compositions of the proteins used by Yada and Nakai were obtained. From these and the *z*-score values for each amino acid, the three-*z*-sums and five-*z*′-sums for each proteinase were calculated (see **Table 2**). PLSR was used to model the physicochemical properties or enzymatic activity ratio of proteinases as a function of the three *z*-sums or five *z*′-sums.

The *z*-sums represent the overall character of a peptide. As such, they are likely to be suitable for modeling bulk characteristics such as physicochemical or functional properties. It seems unlikely that such a representation would be useful for modeling properties that depend only on a small portion of a protein (such as the active site of an enzyme). However, modeling of a protein activity ratio (MC/PA) was attempted with this set of aspartyl proteases.

Table 3. Results of Optimum Models Calculated by PLSR Relating Three-*z*-Sums or Five-*z*′-Sums to Properties of Proteinases

	three-*z*-scale sum models			five-*z*′-scale sum models		
property[a]	comps[b]	R^2	Q^2	comps	R^2	Q^2
H(av)	1	0.692	0.353	1	0.527	0.395
ASA	1	0.118	0.0	1	0.178	0.054
ANS	1	0.327	0.058	1	0.263	0.116
CPA	1	0.445	0.0	2	0.595	0.215
ZP	1	0.660	0.433	2	0.846	0.709
$[\theta]_{\mathrm{MRW}\lambda}$	1	0.532	0.219	2	0.638	0.311
MC/PA	1	0.474	0.298	2	0.699	0.397

[a] The protein property abbreviations are as explained in the footnote to **Table 1**. [b] Number of significant PLS components. R^2 is the squared multiple correlation coefficient. Q^2 is the cross-validated squared multiple correlation coefficient.

When PLSR is carried out with SIMCA-S, the statistics R^2 and Q^2 are calculated. R^2 is the squared multiple correlation coefficient; this expresses the proportion of the variation of a dependent variable that is explained by a model based on the particular data used to derive it and tends to be an optimistic estimate of model fit. The Q^2 parameter is the cross-validated squared multiple correlation coefficient; this is more conservative than R^2 and is considered to provide an estimate of how well a model will predict with new data. It is considered to be a somewhat pessimistic estimate of the prediction ability of a model. The SIMCA-S program uses cross-validation to determine the number of significant PLS components and the optimum model (the one with the highest Q^2). This balances increased predictive ability (higher R^2 values with more predictive terms) against overfitting (increased error as more coefficients are fit). **Table 3** lists the R^2 and Q^2 values for the optimal models of each property using three-*z*-scale sums and five-*z*′-scale sums, and **Tables 4** and **5** list the corresponding coefficients.

The directional contribution of each three-*z*-sum or five-*z*′-sum term to modeling the properties is indicated by the arithmetic sign of the corresponding PLSR coefficient. A better indication of the relative magnitude influences of different *x* variables on a *y* variable (but not the direction) is given by a SIMCA-S statistic called the variable importance in the projection (VIP). The VIP values for the models of proteinase properties are shown in **Table 6**.

Most, but not all, of the properties and activities modeled based on five-*z*′-sums had better fits than those based on three-*z*-sums (see **Table 3**). It is apparent that the five-*z*′-sums contained more information useful for developing models of CPA, ZP, $[\theta]_{\mathrm{MRW}\lambda}$, and MC/PA.

The fit of the Bigelow hydrophobicity, *H*(av), model based on three-*z*-sums (R^2) was better than that produced with five-*z*′-sums, although the prediction ability (Q^2) was slightly weaker (see **Table 3**). This result is not surprising because *H*(av) is obtained by a one-dimensional calculation of the contribution to hydrophobicity made by each amino acid multiplied by the number of moles of that amino acid, summed and divided by the total number of amino acids (*16*). As a result, a three-term model should be more than sufficient to model this property. In fact, both the three-*z*-sum and the five-*z*′-sum models were based on only one PLS component. For the *H*(av) model based

Table 4. PLS Regression Coefficients for the Models of Proteinase Properties and Activity

	three-*z*-sum model coefficients				five-*z*′-sum model coefficients					
property	b_0	b_1	b_2	b_3	b_0'	b_1'	b_2'	b_3'	b_4'	b_5'
H(av)	1250.7	−0.772	0.695	−0.395	1198.50	−0.394	0.297	−0.485	0.174	−0.538
ASA[a]	7508.5	−4.248	−14.810	12.012	7086.10	−12.287	−6.732	2.795	−8.261	20.781
ANS[a]	12.558	0.018	0.027	−0.048	13.959	0.0002	0.011	−0.032	0.011	−0.037
CPA	−33.915	−0.087	−0.245	−0.980	−104.82	−0.236	−0.452	−1.712	−0.265	0.015
ZP	37.541	0.054	0.196	−0.287	41.969	0.188	0.060	−0.156	0.114	−0.330
MC/PA	24.053	−0.712	−0.125	−0.118	−8.052	−0.888	−0.043	−0.208	−0.184	0.432

[a] Single outliers were excluded from the ASA and ANS models.

Table 5. PLS Regression Coefficients for the $[\theta]_{MRW\lambda}$ Model

	wavelength									
	190 nm	193 nm	198 nm	200 nm	202 nm	210 nm	213 nm	222 nm	224 nm	225 nm
					Three-*z*-Sum Model Coefficients					
b_0	13653	6980.8	−6678.2	−9606.7	−10093	−12894	−11572	−11319	−11448	−11428
b_1	−17.526	−8.218	8.920	11.553	10.726	12.438	10.917	12.547	13.360	13.628
b_2	39.089	18.329	−19.894	−25.767	−23.922	−27.741	−24.348	−27.983	−29.798	−30.395
b_3	−73.568	−34.496	37.442	48.496	45.023	52.211	45.826	52.667	56.082	57.206
					Five-*z*′-Sum Model Coefficients					
b_0'	12773	5826.1	−7772.1	−10370	−10372	−12442	−11121	−10838	−10995	−11002
b_1'	−26.272	−46.515	−49.409	−31.619	−5.514	36.955	36.519	32.007	27.917	25.326
b_2'	16.994	11.921	−1.634	−5.925	−8.375	−14.945	−13.636	−14.278	−14.392	−14.272
b_3'	−46.338	−27.726	13.295	23.097	25.979	38.396	34.556	37.243	38.278	38.362
b_4'	8.212	−8.713	−27.554	−23.879	−13.562	−0.085	1.366	−1.780	−4.033	−5.216
b_5'	−41.452	−13.095	33.544	37.662	30.937	28.574	24.477	29.175	31.879	32.958

Table 6. VIP Values for Models of Proteinase Properties and Activity

	VIP values of three-*z*-sum model			VIP values of five-*z*′-sum model				
property	$\sum z_1$	$\sum z_2$	$\sum z_3$	$\sum z_1'$	$\sum z_2'$	$\sum z_3'$	$\sum z_4'$	$\sum z_5'$
H(av)	0.778	1.521	0.285	0.961	1.370	0.821	0.853	0.893
ASA[a]	0.214	1.660	0.445	0.608	1.064	0.894	1.196	1.126
ANS[a]	0.416	1.439	0.870	0.008	0.999	1.037	1.157	1.261
CPA	0.171	1.042	1.373	0.455	0.995	1.522	0.835	0.888
ZP	0.196	1.550	0.747	0.666	0.964	0.940	1.184	1.158
$[\theta]_{MRW\lambda}$	0.300	1.450	0.899	0.777	1.020	1.060	1.066	1.047
MC/PA	1.609	0.612	0.191	1.377	0.719	0.949	0.992	0.838

[a] Single outliers were excluded for the ASA and ANS models.

on three-*z*-sums, $\sum z_1$ and $\sum z_3$ had negative signed coefficients, whereas $\sum z_2$ was positive (see **Table 4**); this indicates that lower values of $\sum z_1$ (hydrophobicity) or $\sum z_3$ (electronic properties) or larger values of $\sum z_2$ (molecular size) increase *H*(av). According to the VIP values shown in **Table 6**, the most influential variable was $\sum z_2$, followed by $\sum z_1$. These results indicate that amino acid molecular size is an important factor for the *H*(av) model. When using five-*z*′-sums to model *H*(av), the signs of the $\sum z_2'$ and $\sum z_4'$ coefficients were positive; however, those for $\sum z_1'$, $\sum z_3'$, and $\sum z_5'$ were negative (see **Table 4**). Similarly to the three-*z*-sum model, the two most influential variables were $\sum z_2'$ and $\sum z_1'$ (see **Table 6**).

For the ASA and ANS models, the R^2 and Q^2 values with both the three-*z*-sum and five-*z*′-sum models were low. This indicates that neither model was significant. This could have resulted from an outlier sample in this data set. Outliers are observations that, for some reason, do not conform to the general pattern present in a data set (*18*). The residual standard deviation (RSD) of an observation in the *Y* space is proportional to the distance from the PLS model (DModY). This is calculated by SIMCA-S after all components are extracted and can be employed to identify outliers. **Figure 1** shows the DModY values for each enzyme for ASA. It is quite apparent that *M. pusillus* proteinase had a higher DModY value than any other observation with both the three-*z*-sum and five-*z*′-sum ASA models. This indicates that this sample did not follow the general pattern and might be an outlier. This sample was removed from the data set, and a second round of modeling was carried out. The relationships between the observed and predicted ASA values are shown in **Figure 2**. The R^2 and Q^2 values of the optimum ASA models were 0.718 and 0.523 using three-*z*-sums and extracting one PLS component, and 0.897 and 0.760 using five-*z*′-sums and extracting two PLS components. These are substantially improved compared to the results shown in **Table 3**. For the ASA outlier-removed model, the term $\sum z_2$ (amino acid molecular size) was the most influential factor (see **Table 6**); it was negatively related to ASA values (see **Table 4**) in the three-*z*-sum model. The Janin calculation of ASA (*17*) used by Yada and Nakai is a function of protein molecular weight; this is also, of course, related to size, but of the entire protein rather than of the amino acids. In the five-*z*-scale model of ASA, the highest two VIP values were those for $\sum z_4'$ (electronegativity) and $\sum z_5'$ (electrophilicity), indicating that $\sum z_4'$ and $\sum z_5'$ were the two most important variables. The arithmetic signs of the $\sum z_4'$ and $\sum z_5'$ coefficients were negative and positive, respectively (shown in **Table 4**), showing that they, not

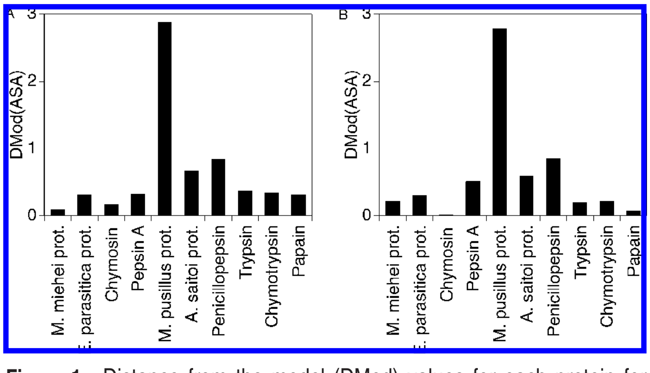

Figure 1. Distance from the model (DMod) values for each protein for the models of accessible surface area (ASA) based on (**A**) three-*z*-sums and (**B**) five-*z*′-sums.

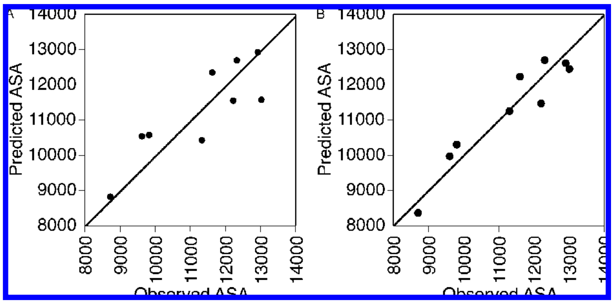

Figure 2. Relationship between observed and predicted ASA properties from the models constructed excluding *M. pusillus* proteinase: fits for ASA based on (**A**) three-*z*-sum ($R^2 = 0.718$, $Q^2 = 0.523$) and (**B**) five-*z*′-sum ($R^2 = 0.897$, $Q^2 = 0.760$) models.

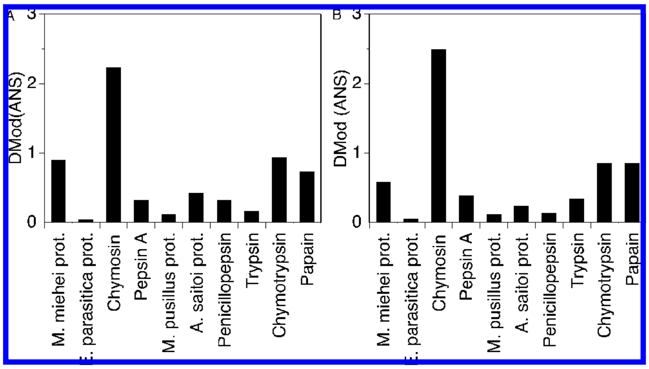

Figure 3. Distance from the model (DMod) values for each protein for the models of apparent surface hydrophobicity determined with the 1-anilino-8-naphthalenesulfonate (ANS) method based on (**A**) three-*z*-sums and (**B**) five-*z*′-sums.

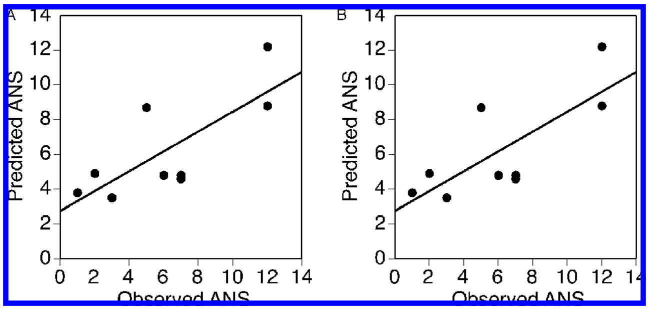

Figure 4. Relationship between observed and predicted ANS from the models constructed excluding chymosin: fits of ANS based on (**A**) three-*z*-sum ($R^2 = 0.489$, $Q^2 = 0$) and (**B**) five-*z*′-sum ($R^2 = 0.581$, $Q^2 = 0.460$) models.

surprisingly, had opposite effects on ASA. In both the three- and five-*z*-scale models a single PLS component was used; this indicates that a single fundamental property was sufficient to explain the variation in ASA.

For the ANS model chymosin had a high DModY value (see **Figure 3**), suggesting that this sample may be an outlier. A second round of ANS modeling was carried out excluding chymosin. The results obtained (see **Figure 4**) were somewhat improved. The R^2 produced with three-*z*-sums and one PLS component increased from 0.327 to 0.489, whereas Q^2 decreased from 0.058 to 0. The five-*z*′-sum model R^2, however, increased from 0.263 to 0.581 and Q^2 increased from 0.116 to 0.460. Therefore, the five-*z*′-sum models of ASA and ANS were better than the respective three-*z*-sum models.

The three-*z*-sum model of hydrophobicity measured with CPA had R^2 and Q^2 values of 0.445 and 0, respectively, whereas the corresponding five-*z*′-sum model values were 0.595 and 0.215. These were weak models. There are many methods to measure protein hydrophobicity (*19*). Among these, the ANS and CPA methods both employ fluorescence probes, and CPA has some advantages compared to ANS (*20*). The VIP value ranks for the ANS and CPA models were quite different. $\sum z_2$ was the most important variable for the ANS model, whereas the term $\sum z_3$ was most influential for the CPA model, followed by $\sum z_2$ (see **Table 6**). The signs of the $\sum z_1$ and $\sum z_2$ coefficients were opposite for the ANS and CPA models. The five-*z*-sum ANS model VIP values for $\sum z_4'$ and $\sum z_5'$ were similar and higher than for the other terms; the coefficients for these two terms were positive and negative, respectively (see **Table 4**). Unlike the ANS model, $\sum z_3'$ was the most influential term for the CPA model based on five-*z*′-sums. The regression coefficients in **Table 4** indicate that CPA increases with decreasing $\sum z_3'$ (electronic properties).

The three-*z*-sum model of zeta potential (ZP) had R^2 and Q^2 values of 0.660 and 0.433, respectively, whereas the corresponding five-*z*′-sum model values were 0.846 and 0.709. This was the strongest model produced with this data set, excepting the outlier-removed ASA model. For the ZP model, $\sum z_2$ had the largest VIP value in the three-*z*-sum model (see **Table 6**) and was positively related to ZP (shown in **Table 4**). The terms $\sum z_4'$ and $\sum z_5'$ had similar VIP values and greater influence on ZP compared with the other three terms in the five-*z*-sum model (shown in **Table 6**). As shown in **Table 4**, b_4' (positive) and b_5' (negative) had different signs denoting opposite associations with ZP.

Circular dichroism (CD) is useful for structural characterization of proteins and particularly for secondary structure determination (*21*, *22*). In the data set used in this paper CD observations were made at 10 wavelengths, so there were 10 dependent (*y*) variables. PLSR has the ability to simultaneously model multiple *y* variables as a function of the same set of *x* variables. If *y* variables are correlated, it is preferable to analyze them with a single model (*23*), because PLSR in this way can yield simpler overall results than using one separate model for each *y* variable. Before PLSR was performed, PCA was employed to assess the correlation of the *y* variables. Cross-validation indicated that four components were significant and explained 99.8% of the variance of the *y* variables. Because 4 components are considerably fewer than the 10 variables, this indicates that the *y* variables were indeed correlated. The CD observations at the 10 wavelengths were analyzed together by PLSR using three-*z*-sums or five-*z*′-sums; the results are shown in **Table 5**. The most important three-*z*-sum model term was $\sum z_2$ (see **Table 6**). As listed in **Table 5**, the response coefficients at all wavelengths except 190 and 193 nm had negative signs. For the $[\theta]_{\mathrm{MRW}\lambda}$ model based on five-*z*′-sums, the VIP values of four terms, $\sum z_2'{-}\sum z_5'$, had similar values, with $\sum z_4'$ slightly higher than the others (see **Table 6**). The corresponding coefficients tended to have opposite signs for the 190 and 193 nm responses from the higher wavelengths (shown in **Table 5**).

Enzymes used for cheesemaking need to not only clot milk but also possess general proteolytic activity (*24*). The MC/PA (milk clotting to proteolytic activity ratio) characterizes the enzymatic activity. As previously mentioned, it seems unlikely that the *z*-sum modeling approach, which characterizes bulk protein properties, would successfully represent enzyme activity. However, the three-*z*-sum model of MC/PA had R^2 and Q^2 values of 0.474 and 0.298, respectively, whereas the corresponding five-*z*′-sum model values were 0.699 and 0.397. Therefore, this model was actually stronger than the CPA and

Table 7. Cumulative R^2 and Q^2 versus Correlation Intercepts after 30 Permutations of Each Model

	three-*z*-sum model intercepts		five-*z*′-sum model intercepts	
property	R^2	Q^2	R^2	Q^2
H(av)	0.18	−0.03	0.21	−0.02
ASA[a]	0.24	−0.04	0.29	−0.17
ANS[a]	0.25	0.05	0.18	−0.06
CPA	0.29	0.06	0.36	−0.01
ZP	0.22	−0.02	0.28	−0.12
MC/PA	0.25	−0.02	0.29	−0.06

[a] Single outliers were excluded for the ASA and ANS models.

$[\theta]_{\mathrm{MRW}\lambda}$ models and similar in strength to the *H*(av) model. For the models of MC/PA, $\sum z_1$ and $\sum z_1'$ were the most influential variables (see **Table 6**) for the three-*z*-sum and five-*z*′-sum models, respectively. The coefficients of all three terms used in the three-*z*-sum model had negative signs (see **Table 4**), indicating that lower hydrophobicity, lower molecular size and lower electronic properties were related to higher MC/PA. In the five-*z*′-sum model of MC/PA, all of the coefficients except b_5' had negative signs (see **Table 4**).

In the five-*z*′-sum models for ASA (outlier excluded), ANS (outlier excluded), and ZP, the two highest VIP values were in all cases $\sum z_4'$ and $\sum z_5'$. For the $[\theta]_{\mathrm{MRW}\lambda}$ model, the VIP values of $\sum z_4'$ and $\sum z_5'$, which were similar in magnitude, were first and third highest. This indicates that the fourth and fifth scales contained useful information for developing models and presumably explains why using five-*z*′-sums most often produced stronger models than using three-*z*-sums (see **Table 3**). For the MC/PA model, the VIP value of $\sum z_4'$ ranked second of the five VIP values, which indicates that $\sum z_4'$ provides information that strengthened the model. In the CPA and *H*(av) models, although the VIP values of $\sum z_5'$ were third in the ranking, this variable had influence on the models.

The models developed were validated by response permutation (*24*), which estimates the chance (probability) of getting a good fit with randomly reordered response data (*25*). This permutation leaves the predictor (*x*) variables intact but randomly shuffles the order of the *y* variables multiple times (30 times in this study). After each reordering, the model is refit and R^2 and Q^2 are calculated and compared with the original model. The cumulative R^2 and Q^2 are plotted versus the correlation coefficient between the original *y* and the permuted *y*. The intercepts measure the extent of overfitting of a model. For a valid model the intercepts of R^2 and Q^2 should be less than 0.3 and 0.05, respectively (*26*). This was the case with all of the models except those for CPA and for $[\theta]_{\mathrm{MRW}\lambda}$ measured at 193 nm (see **Tables 7** and **8**). For the CPA model, the Q^2 intercept for the three-*z*-sum model and the R^2 intercept of the five-*z*′-sum model are slightly higher than the suggested limit. According to Wold and Eriksson (*24*), high R^2 values from permutations could be obtained with a random *y*-vector. However, a high Q^2 value is not likely to be obtained in this manner, and so the CPA model based on five-*z*′-sums can be regarded as valid. The models of $[\theta]_{\mathrm{MRW}\lambda}$ measured at 193 nm, using either three-*z*-sums or five-*z*′-sums, were invalid because the intercepts of both R^2 and Q^2 based on 30 permutations exceeded the recommended values for a valid model.

It has been stated that if the structure of a protein is known, its function can be understood (*27*). Hydrophobic, electrostatic, and steric parameters are important characteristics that describe molecules. Models have been made that predict the functional mechanisms of many chemical compounds, including drugs and olfactory stimulants, based on these three parameters (*28*). It is interesting that the physicochemical meaning of these parameters is similar to those represented by the three-*z*-sums or the first three terms of the five-*z*′-sums. According to eq 1, the terms $\sum z_i$ (or $\sum z_i'$) can be regarded as linear combinations of the contributions of each amino acid in a protein to each fundamental property. Because both the number of moles of each amino acid (n_a in eq 1) and the *z*-scale values of each amino acid are different, each proteinase has unique values for the property sums. By noting how much each amino acid contributes to a $\sum z_{\mathrm{i}}$ or $\sum z_i'$, it should be possible to discover which types of amino acids are responsible for strong positive or negative effects on the properties or activity of proteinases. The terms that had the largest VIP values in a model (see **Table 6**) were first noted, and the positive or negative relationship to a property was found from the arithmetic sign of the corresponding coefficient (see **Table 4**). The amino acids with greatest contributions to the *z*-sum, based on the magnitudes of the $n_a z_{ia}$ ($n_a z_{ia}'$) values, were the most influential for a property. **Table 9** shows the three most positively and negatively influential amino acids in various proteinases for each model.

For example, because the MC/PA model based on five-*z*′-sums was much stronger than the one obtained with three-*z*-sums, the most influential amino acids were determined on the basis of the five-*z*′-sum model. According to **Table 1**, chymosin has a high milk clotting to proteolytic activity ratio, whereas *A. saitoi* proteinase has a low value of MC/PA, so the differences between the two proteinases in their contents of amino acids were compared. As shown in **Table 6**, the most important term in the MC/PA model was $\sum z_1'$, and the sign of the corresponding coefficient in **Table 4** was negative, so the strongest positive influence on the MC/PA value was from amino acids that contributed highly negative values of $n_a z_{1a}'$. This can be seen in **Figure 5**, which shows the values of $n_a z_{1a}'$ for chymosin and the *A. saitoi* proteinase, which had, respectively, the highest and lowest values of MC/PA of the proteases studied. For chymosin the three amino acids contributing most strongly to negative $n_a z_{1a}'$ values were leucine (L), isoleucine (I), and phenylalanine (F). Aspartic acid (D), serine (S), and glycine (G) had the three largest positive contributions to $n_a z_{1a}'$, contributing to low values of MC/PA. Similar analyses were carried out for each enzyme for each modeled property, and the results are shown in **Table 9**.

Some general trends could be observed. In the *H*(av) model at least one, and often two, aromatic amino acids (phenylalanine (F), tyrosine (Y), or tryptophan (W)) contributed positively to *H*(av) for all of the proteinases. Aspartic acid (D) contributed positively in 6 of the 10 proteinases. One of the hydroxylated amino acids (threonine (T) or serine (S)) contributed to low *H*(av) values in 8 of the 10 proteinases.

In the outlier-removed ASA model the hydroxylated (S and T) and acidic (D and glutamic acid (E)) amino acids contributed positively to ASA in all nine of the modeled enzymes. Negative contributors to ASA were frequently basic (H, K, and R) and aromatic (F and W) amino acids.

The outlier-removed model of hydrophobicity measured with ANS in every case had positive contributions from aromatic (F, Y, and W) amino acids and often negative contributions from proline (P) (seven of nine), alanine (A) (eight of nine), and asparagine (N) (six of nine). This has similarities to the calculated hydrophobicity model (*H*(av)) in having aromatic amino acids contributing positively (nine of nine) and hydroxylated amino acids negatively (five of nine).

The CPA model in every case had positive contributions from one or more nonpolar amino acids (isoleucine (I), leucine (L),

Table 8. Cumulative R^2 and Q^2 versus Correlation Intercepts for Permutations of the $[\theta]_{MRW\lambda}$ Model

	wavelength									
model	190 nm	193 nm	198 nm	200 nm	202 nm	210 nm	213 nm	222 nm	224 nm	225 nm
					Three-*z*-Sum Model Intercepts					
R^2	0.28	0.30	0.10	0.14	0.19	0.18	0.19	0.16	0.13	0.13
Q^2	0.04	0.07	−0.05	−0.09	−0.06	−0.01	0.01	−0.02	−0.05	−0.06
					Three-*z*′-Sum Model Intercepts					
R^2	0.30	0.39	0.12	0.07	0.14	0.25	0.27	0.26	0.23	0.22
Q^2	0.02	0.15	−0.06	−0.22	−0.16	−0.03	0.00	−0.03	−0.07	−0.09

Table 9. Three Amino Acids Contributing Most Positively (+) or Negatively (−) to the Modeled Property

	modeled attribute						
protein	*H*(av)	ASA[a]	ANS[a]	CPA	ZP	$[\theta]_{MRW\lambda}$	MC/PA
M. miehei proteinase +	Y, N, F[b]	D, S, T	Y, G, F	V, I, T	D, S, T	D, S, T[e]	F, L, I
M. miehei proteinase −	G, V, T[c]	K, R, M	A, P, N	D, S, P	K, R, M	K, R, M[f]	D, S, G
E. parasitica proteinase +	Y, F, D	T, S, D	Y, T, G	T, V, I	T, S, D	T, S, D	F, L, I
E. parasitica proteinase −	G, T, S	W, K, F	A, S, P	S, D, P	W, K, F	W, K, F	S, D, G
chymosin +	Y, D, F	D, S, T		V, L, I	D, S, T	D, S, T	L, I, F
chymosin −	G, V, Q	H, M, G		D, S, P	H, M, G	H, M, G	D, S, G
pepsin +	Y, D, N	D, S, E	Y, C, G	I, L, V	D, S, E	D, S, E	L, I, F
pepsin −	G, S, V	W, P, M	P, S, L	D, S, P	W, P, M	W, P, M	D, S, G
M. pusillus proteinase +	Y, D, F		Y, F, G	V, K, L	D, S, T	D, S, T	F, L, V
M. pusillus proteinase −	G, V, T		P, N, A	D, S, P	K, F, P	K, F, P	D, G, S
A. saitoi proteinase +	Y, D, F	D, S, T	Y, G, T	V, L, T	D, S, T	D, S, T	L, V, F
A. saitoi proteinase −	G, S, V	K, H, W	S, A, P	D, S, P	K, H, W	K, H, W	D, S, G
penicillopepsin +	Y, F, D	S, D, T	Y, G, F	Q, V, T	S, D, T	S, D, T	L, F, V
penicillopepsin −	G, S, T	H, F, W	A, S, N	S, D, P	H, F, W	H, F, W	S, D, G
trypsin +	Y, N, W	S, D, E	C, Y, G	L, I, V	S, D, E	S, D, E	L, I, V
trypsin −	G, V, S	H, W, K	A, N, P	C, S, D	H, W, K	H, W, K	G, S, D
chymotrypsin +	W, K, N	S, T, D	C, W, T	V, K, L	S, T, D	S, T, D	L, V, I
chymotrypsin −	G, V, T	W, K, H	A, N, S	C, S, D	W, K, H	W, K, H	S, G, N
papain +	Y, R, N	E, G, S	Y, C, G	R, V, K	E, G, S	E, G, S	Y, L, I
papain −	G, V, A	R, W, K	N, P, A	C, P, S	R, W, K	R, W, K	G, R, N

[a] Outlier was excluded for the ASA or ANS model. [b] The three amino acids most strongly associated with high property values. [c] The three amino acids most strongly associated with low property values. [e] The three amino acids most strongly associated with low $[\theta]_{MRW\lambda}$ values at 190 and 213 nm, but most strongly associated with high $[\theta]_{MRW\lambda}$ at other wavelengths. [f] The three amino acids most strongly associated with high $[\theta]_{MRW\lambda}$ values at 190 and 213 nm, but most strongly associated with low $[\theta]_{MRW\lambda}$ values at other wavelengths.

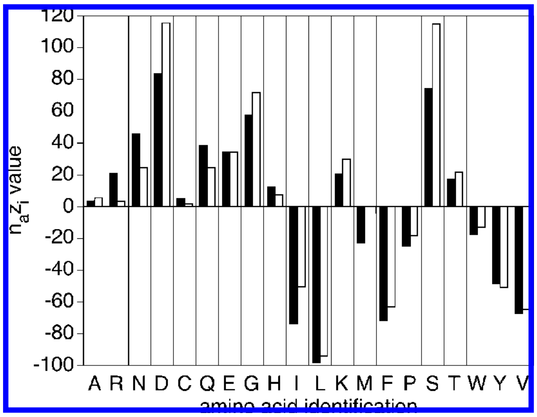

Figure 5. Contributions of amino acids to $\sum z_1'$, which is strongly related to the property MC/PA: data for chymosin (black) and *A. saitoi* proteinase (white).

or valine (V)). Negative contributions to CPA were in all cases from hydroxylated (S or T) and often (9 of 10 times) from aspartic acid (D).

MC/PA had strong positive contributions from nonpolar (I, L, and V) (10 of 10) and aromatic (F and Y) (8 of 10) amino acids. Negative contributions came from hydroxylated (S) (9 of 10) and acidic (D) (8 of 10) amino acids.

ZP had strong positive contributions from hydroxylated (S or T) and acidic (D or E) amino acids in all 10 cases. Strong negative influence came from basic amino acids (H, R, or K) in 9 of 10 cases and from aromatic amino acids (F or W) in 8 of 10 cases.

The CD model had strong positive contributions from hydroxylated (S and T) and acidic amino acids (D and E) in all 10 cases. Strong negative influence came from basic amino acids (H, K, or R) (9 of 10). Aromatic amino acids (F or W) frequently exerted negative influence as well (8 of 10).

Other than the association between the free energies of transfer of amino acid side chains from organic to aqueous phase and hydrophobicity, which is used to calculate the Bigelow average hydrophobicity of a protein, very few cases of modeling protein properties from amino acid composition have been described in the literature. Siebert previously used amino acid principal property sums to model Bigelow and exposed hydrophobicity and foam capacity. In the current study successful models of protein circular dichroism, accessible surface area, zeta potential, and hydrophobicity determined by two different assays (CPA and ANS) as functions of amino acid composition were demonstrated for the first time. Modeling of a ratio of enzyme activities in a similar manner was also unprecedented.

LITERATURE CITED

(1) Hettiarachchy, N. S.; Ziegler, G. R. *Protein Functionality in Food Systems*; Dekker: New York, 1995.

(2) Kinsella, J. E.; Rector, D. J.; Phillips, L. G. Physicochemical properties of proteins: texturization via gelation, glass and film

formation. In *Protein Structure−Function Relationships in Foods*; Yada, R. Y., Jackman, R. L., Smith, J. L., Eds.; Blackie Academic and Professional: London, U.K., 1994; pp 1−21.

(3) Pomeranz, Y. *Functional Properties of Food Components*, 2nd ed.; Academic Press: New York, 1985.

(4) Damodaran, S. Structure−function relationship of food proteins. In *Protein Functionality in Food Systems*; Hettiarachchy, N. S., Ziegler, G. R., Eds.; Dekker: New York, 1995; pp 1−37.

(5) Fligner, K. L.; Mangino, M. E. Relationship of composition to protein functionality. In *Interactions of Food Proteins*; Parris, N., Barford, R., Eds.; American Chemical Society: Washington, DC, 1991; pp 1−12.

(6) Nakai, S.; Li-Chan, E.; Hayakawa, S. Contribution of protein hydrophobicity to its functionality. *Nahrung* **1986**, *30* (3/4), 327–336.

(7) Phillips, L. G.; Whitehead, D. M.; Kinsella, J. E. *Structure−Function Properties of Food Proteins*; Academic Press: New York, 1994.

(8) Siebert, K. J. Modeling protein functional properties from amino acid composition. *J. Agric. Food Chem.* **2003**, *51*, 7792–7797.

(9) Sneath, P. H. A. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **1966**, *12*, 157–195.

(10) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure−activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126–1135.

(11) Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. Multivariate parametrization of 55 coded and non-coded amino acids. *Quant. Struct.−Act. Relat.* **1989**, *8*, 204–209.

(12) Collantes, E. R.; Dunn, W. J., III. Amino acid side chain descriptors for quantitative structure−activity relationship studies of peptide analogues. *J. Med. Chem.* **1995**, *38*, 2705–2713.

(13) 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.

(14) Siebert, K. J. 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.

(15) Yada, R. Y.; Nakai, S. Use of principal component analysis to study the relationship between physical/chemical properties and the milk clotting to proteolytic activity ratio of some aspartyl proteinases. *J. Agric. Food Chem.* **1986**, *34*, 675–679.

(16) Bigelow, C. C. On the average hydrophobicity of proteins and the relation between it and protein structure. *J. Theor. Biol.* **1967**, *16* (2), 187–211.

(17) Janin, J. Surface area of globular proteins. *J. Mol. Biol.* **1976**, *105* (1), 13–14.

(18) Buydens, L. M. C.; Reijmers, T. H.; Beckers, M. L. M.; Wehrens, R. Molecular data-mining: a challenge for chemometrics. *Chemom. Intell. Lab. Syst.* **1999**, *49* (2), 121–133.

(19) Nakai, S. Structure−function relationships of food proteins with an emphasis on the importance of protein hydrophobicity. *J. Agric. Food Chem.* **1983**, *31*, 676–683.

(20) Sklar, L. A.; Hudson, B. S.; Simoni, R. D. Conjugated polyene fatty acids as fluorescent membrane probes: model system studies. *J. Supramol. Struct.* **1976**, *4* (4), 449–465.

(21) Bulheller, B. M.; Rodger, A.; Hirst, J. D. Circular and linear dichroism of proteins. *Phys. Chem. Chem. Phys.* **2007**, *9* (17), 2020–2035.

(22) Kelly, S. M.; Jess, T. J.; Price, N. C. How to study proteins by circular dichroism. *Biochim. Biophys. Acta, Proteins Proteomics* **2005**, *1751* (2), 119–139.

(23) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(24) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; Van de Waterbeemd, H., Ed.; VCH: New York, 1995; Vol. 2, pp 309−318.

(25) Wold, S.; Sjöström, M.; Eriksson, L. Partial least squares projections to latent structures (PLS) in chemistry. In *Encyclopedia of Computational Chemistry*; von Schleyer, R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 2006−2021.

(26) Andersson, P. M.; Sjöström, M.; Lundstedt, T. Preprocessing peptide sequences for multivariate sequence−property analysis. *Chemometr. Intell. Lab. Sys.* **1998**, *42* (1−2), 41–50.

(27) Schulz, G. E.; Barry, C. D.; Friedman, J.; Chou, P. Y.; Fasman, G. D.; Finkelstein, A. V.; Lim, V. I.; Ptitsyn, O. B.; Kabat, E. A.; Wu, T. T.; Levitt, M.; Robson, B.; Nagano, K. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature* **1974**, *250*, 140–142.

(28) Stuper, A. J.; Brügger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Functions*; Wiley: Somerset, NJ, 1979.

Received for review September 23, 2008. Revised manuscript received January 13, 2009. Accepted January 23, 2009. The support of this work by the China Scholarship Council (Grant 2006-3142) is gratefully acknowledged. This material is based upon work supported by the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Project NYG 623-496.

JF802971N