JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Modeling Protein Functional Properties from Amino Acid Composition

KARL J. SIEBERT*

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

Physicochemical and functional properties of proteins were modeled as a function of the contributions of each of the 20 coded amino acids to three (*z*-scores) or five (extended *z*-scores) amino acid principal properties using partial least squares regression. The five term models were in all cases stronger in both fit and prediction than the three term models, indicating that useful information is contained in the fourth and fifth property scores. Models predicting protein hydrophobicity ($R = 0.932$), viscosity ($R = 0.737$), and foam capacity ($R = 0.880$) from amino acid composition rather than sequence were obtained. It is likely that additional functional and physicochemical properties of proteins can be modeled in this way.

KEYWORDS: QSAR; functional properties; principal components analysis; partial least squares regression

INTRODUCTION

It was previously shown that it is possible to model biological properties of small peptides as a function of amino acid principal properties (*1*, *2*). These were derived by carrying out principal components analysis (PCA) of numerous amino acid properties for all of the common coded amino acids as well as a number of noncoded ones. The authors called the first three principal component scores of each amino acid its z_1, z_2, and z_3 scores or principal properties. These were interpreted to represent largely hydrophilicity, side chain bulk/molecular size, and electronic properties, respectively. The three principal properties for the amino acid in each position in a peptide were then used to construct models. For example, a model of dipeptide bitterness of the following form was developed (*2*)

$$y = b_{11}z_{11} + b_{21}z_{21} + b_{31}z_{31} + b_{12}z_{12} + b_{22}z_{22} + b_{32}z_{32} \quad (1)$$

where y is the bitterness, the b values are regression coefficients, and the z values are the three z-scores for each of the two amino acids in the dipeptide sequence (e.g., z_{21} is the z_2 score for the first, or N-terminal, amino acid in the dipeptide). The multiple correlation coefficient, R, was 0.88 for the dipeptide bitterness model. Bradykinin potentiating activity of pentapeptides was fit to a 15-term model (three z-scores $\times$ five amino acids) with $R = 0.90$. The general formula for this approach can be written as the summation

$$y = \sum_{j=1}^{n}\sum_{i=1}^{3} b_{ij}z_{ij} \quad (2)$$

for a peptide of length n amino acids.

This approach produced good models for small peptides but has the disadvantage for those larger than a few amino acids than the number of terms to fit, and as a result, the number of peptides needed to construct a model is large.

There are, however, some special cases in which large peptides can be modeled with fewer terms (*3*). In a homopolymer, all of the amino acids in the peptide are the same; it was possible to model the Coomassie blue dye binding response to homopolymers as a function of the three z-scores for the amino acid comprising the homopolymer ($R = 0.926$).

Some peptide properties result from the proportion of only one or a few amino acids. It was possible to model two of these situations from the number of moles of the relevant amino acids in the protein (*3*). In the following fits Q, the cross-validated multiple correlation coefficient, is given as well as R. The Q is considered to represent the predictive ability of a model, while R is an estimate of model fit (*4*). R increases with increasing model complexity (terms or components) and can be overoptimistic. As model complexity increases, Q reaches a maximum at the point where complexity and fit are balanced. Coomassie blue dye binding of proteins was modeled in terms of their contents of three basic and three aromatic amino acids ($R = 0.976$; $Q = 0.572$). UV absorbance of proteins depends only on tyrosine, tryptophan, and cysteine (*5*); this had $R = 0.995$ and $Q = 0.988$ (*3*). It was also possible to construct models based on the algebraic sums of the contributions of the relevant amino acids to each of the three principal properties (*3*). In the case of UV absorbance at 280 nm, the contributions to z_1 of tyrosine (Y), tryptophan (W), and cysteine (C) were estimated by multiplying the number of moles of each of these amino acids in the protein by its z_1-score and summing algebraically:

$$\sum z_1 = z_{1Y}n_Y + z_{1W}n_W + z_{1C}n_C \quad (3)$$

* To whom correspondence should be addressed. Tel: 315 787-2299. Fax: 315 787-2284. E-mail: kjs3@cornell.edu.

10.1021/jf0342775 CCC: $25.00
Published on Web 11/14/2003

The contributions to z_2 and z_3 ($\sum z_2$ and $\sum z_3$) were computed in an analogous manner. A model expressing the molar absorptivity of a protein at 280 nm, ϵ, as a function of these z-score sums was then developed (*3*).

$$\epsilon = b_0 + b_1\sum z_1 + b_2\sum z_2 + b_3\sum z_3 \qquad (4)$$

This had an identical fit to the model based simply on the moles of each amino acid ($R = 0.995$) and a marginally better prediction ability ($Q = 0.992$). When the same approach was applied to Coomassie blue response to proteins, the fit ($R = 0.935$) was not as good as with the six term model but the predictive ability ($Q = 0.890$) was much better (*3*).

Recently, the amino acid principal property approach was expanded to a larger set of amino acids (20 coded + 67 noncoded) and more parameters (*6*). Application of PCA resulted in a set of five orthogonal axes termed extended z-scales (ext-z), of which the first three largely corresponded to the original z-scales. The ext-z_1-scale was interpreted to largely represent hydrophilicity (similar to the original z_1-scale). The ext-z_2-scale corresponded to molecular size (similar to z_2). The ext-z_3-scale was similar to z_3 and was described as representing electronic properties. The ext-z_4- and ext-z_5-scales were more complicated. The ext-z_4-scale was associated positively with heat of formation and negatively with electronegativity, both estimated from molecular orbital calculations. The ext-z_5-scale was associated with "hardness" and the energy levels of the highest occupied and the lowest unoccupied molecular orbitals, again from molecular orbital calculations, and NMR α-proton shift observations. The extended z-scores were applied in modeling two peptide data sets, elastase substrates and neurotensin analogues, and performed well. This suggested that more than three principal properties may be useful to represent behavior in some cases.

One of the situations where the proportion of amino acids of different types rather than a precise sequence is thought to impact protein properties is in what are called by food chemists functional properties. Various functional properties have been listed by different authors and include solubility, wettability, gelation, fat binding, water binding, emulsifying capacity, and foam, film, and glass formation (*7*−*11*). A number of physicochemical properties (hydrophobicity, melting point, etc.) have been related to the proportions of individual amino acids or particular classes of amino acids (e.g., acidic, basic, hydrophilic, hydrophobic, aromatic, etc.) in a protein (*8*). A number of the functional properties have in turn been related to protein physicochemical properties (*8*, *12*). For example, hydrophobicity, either in a domain or of an entire protein, is associated with foaming, gel formation, and binding of nonpolar flavor compounds (*8*, *13*).

It was of interest to see if it was possible to model physicochemical and functional properties of proteins in terms of their amino acid composition and principal property scores.

MATERIALS AND METHODS

Data reported by Townsend and Nakai (*14*) for Bigelow hydrophobicity, exposed hydrophobicity, viscosity, and foam capacity under several conditions of pH and ionic strength for a number of well-characterized proteins were used. The amino acid sequences for the proteins were obtained from the Swiss-Prot Protein Knowledgebase (Swiss Institute for Bioinformatics and European Bioinformatics Institute, http://www.wbi.ac.uk/swissprot). The sequences were converted into composition (i.e., the number of moles of each amino acid in the protein) using the Southampton Bioinformatics Data Server web site (http://molbiol.soton.ac.uk) of the University of Southampton, U.K.

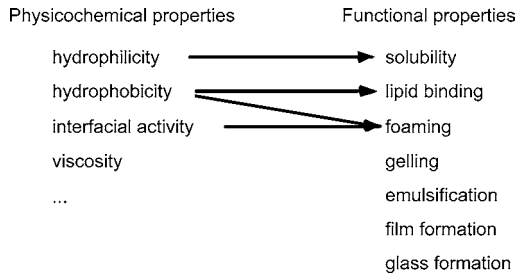

Figure 1. Concept of relationship between peptide functional and physicochemical properties. Various physicochemical properties either singly or in combination are responsible for functional properties.

The amino acid z-scores used were those reported by Jonsson et al. (*2*). The extended z-scores were from Sandberg et al. (*6*). Sums of each of the three z-scores ($\sum z_i$) and each of the five extended z-scores ($\sum$ext-z_i) for a protein were computed by multiplying the appropriate score (e.g., z_1) for each amino acid by the number of moles (n) of that amino acid in the protein and then summing algebraically. So, for $\sum z_i$:

$$\sum z_i = \sum_{X=1}^{20} z_{iX} n_X \qquad (5)$$

where X represents each of the 20 coded amino acids in a protein.

Modeling of the property data as a function of amino acid composition was carried out by partial least squares regression (PLSR) using the SIMCA-S for Windows computer program v 6.01 (Umetrics Inc., Kinnelon, NJ); this provided estimates of model fit (the multiple correlation coefficient, R) and predictive ability (the cross-validated correlation coefficient, Q). The SIMCA-S program uses cross-validation of models calculated with increasing numbers of components to determine the best prediction model (the minimum number of components needed to achieve the best Q). Comparison of the relative influence of the terms was made using the variable importance in the projection (VIP) calculation provided by the SIMCA-S program (*15*).

RESULTS AND DISCUSSION

In functional properties of proteins as in nearly all of the interactions of peptides in biological systems, the interactions are noncovalent. We know of only a limited number of mechanisms for noncovalent interactions: mainly ionic bonding, hydrophobic bonding, hydrogen bonding, and van der Waals interactions. It seems likely that various combinations of these phenomena are involved in different proportions in particular protein properties. If only a limited number of mechanisms are involved, the number of quantities needed to represent the behavior may be modest.

There are many reports relating the functional properties of proteins to their physicochemical nature (*7*, *8*, *16*−*18*); see the concept in **Figure 1**. Functional properties are thought to arise from various combinations of physicochemical properties (*11*, *17*, *19*). Physicochemical properties include viscosity, surface activity, hydrophobicity, hydrophilicity, adhesion, and cohesion, among others. The physicochemical properties themselves result from the structure of a protein, and some have been related to the proportions of amino acids of different classes (*8*). For example, the hydrophobic nature of proteins, which has been associated with foaming and emulsification, results from the proportion of amino acids with nonpolar side chains. The ionic nature of proteins, which influences solubility and water binding, results from the proportion of amino acids with basic and acidic side chains. The content of some amino acids determines aspects of protein folding, which impacts protein shape and rigidity; these in turn influence viscosity. So, ultimately, the amino acid content of a protein largely determines

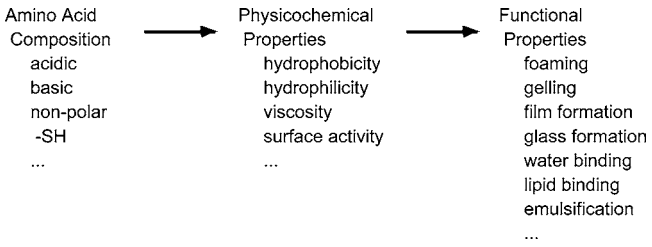

Figure 2. Concept of relationship between amino acid composition and peptide functional and physicochemical properties. Contents of amino acids of different classes determine physicochemical properties, which in turn determine functional properties.

the physicochemical properties, which determine the functional properties; see **Figure 2**.

Conceptually then, relationships between amino acid composition and physicochemical and functional properties can be thought of as follows. Amino acid composition (i.e., the contents of amino acids of various classes), as opposed to sequence, influences physicochemical properties (e.g., hydrophobicity, hydrophilicity, surface activity, viscosity, etc.), which in turn, in various combinations, determines functional properties (foaming, gelling, formation of films or glasses, binding of water or lipids, emulsification, etc.). So, if physicochemical properties can be predicted from amino acid composition, it should also be possible to directly predict protein functional properties from amino acid composition; see **Figure 3**.

As described earlier, using *z*-score sums to model a peptide property that is known to be a function of a limited number of amino acids worked well (*3*). The properties modeled (UV absorbance and dye binding), however, were not functional properties. In many cases with functional properties, there may not be prior knowledge of particular amino acids that influence the property. A broad brush approach is possible, however, by computing *z*-score contributions (three *z*-sums) or ext-*z*-score contributions (five ext-*z*-sums) from all of the amino acids that a peptide contains (in most cases, some or all of the 20 coded amino acids). This approach was attempted using data from a paper that reported both physicochemical data (viscosity and hydrophobicity) and functional property data (foam capacity) for a number of well-characterized proteins (*14*). The principal property sums were computed for the proteins used in this study; see **Tables 1** and **2**.

Models of the physicochemical properties as a function of the *z*-sums and ext-*z*-sums were calculated using PLSR. The models were of the form:

$$\text{property} = b_0 + b_1\sum z_1 + b_2\sum z_2 + b_3\sum z_3 \quad (6)$$

or

$$\text{property} = b_0 + b_1\sum z_1 + b_2\sum z_2 + b_3\sum z_3 + b_4\sum z_4 + b_5\sum z_5 \quad (7)$$

PLSR can employ different numbers of components, also called latent variables (*4*). As the number of latent variables increases, the fit (R) increases, but after some point, the model validity, usually expressed as Q, declines or fails to increase further. The number of latent variables corresponding to this point is considered to represent a reasonable balance between fit and prediction ability known as the best prediction model. Results for the best prediction models, chosen by cross-validation, that relate amino acid principal property sums of proteins to physicochemical properties are shown in **Table 3**. It is quite apparent that the fits produced with the five ext-*z*-sums were in each case much stronger than those obtained with the three *z*-sums and that this approach to modeling was quite successful. Clearly, information that is valuable for modeling these protein properties resides in the fourth and fifth (as well as the first three) amino acid PCs. The relationships between each observed property and the equivalent values predicted from the ext-*z*-sum models are shown in **Figures 4**−**6**. The coefficients for

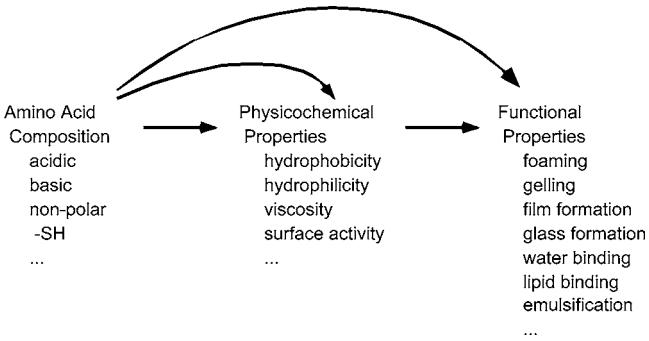

Figure 3. Concept of modeling peptide physicochemical and functional properties from amino acid composition.

Table 1. *z*-Score Sums of the Proteins Used in the Townsend and Nakai Study

protein	no.[a]	$\sum z_1$	$\sum z_2$	$\sum z_3$
pepsin	371	18.71	−300.99	34.71
conalbumin	686	305.19	−363.08	−75.37
RNAase I	119	81.46	−14.5	−28.01
lysozyme	129	71.79	−54.23	−0.58
ovomucoid	186	115.06	−121.11	24
ovalbumin A	385	23.51	−221.54	−43.41
serum albumin	583	181.25	−199.95	−75
κ-casein	169	29.94	−65.28	0.26
β-casein	209	−25.22	−71.28	12.97
β-lacto-globulin	162	−5.29	−79.91	−31.04
trypsin	224	57.29	−166.69	8.32

[a] Number of amino acids in the protein.

Table 2. Ext-*z*-Score Sums of the Proteins Used in the Townsend and Nakai Study

protein	$\sum z_1$	$\sum z_2$	$\sum z_3$	$\sum z_4$	$\sum z_5$
pepsin	9.66	−280.86	12.98	−274.37	96.59
conalbumin	290.1	−344.85	−84.45	−294.96	118.57
RNAase I	99.25	−11.25	−24.2	−20.77	3.97
lysozyme	72.66	−50.31	−2.69	−32.44	16.62
ovomucoid	145.76	−117.33	44.22	−108.6	8.5
ovalbumin A	0.58	−207.54	−81.63	−189.61	98.85
serum albumin	203.48	−188.79	−85.94	−257.84	91.65
κ-casein	8.56	−58.47	−34.26	−85.12	70.2
β-casein	−76.65	−69.11	−45.01	−92.27	106.17
β-lacto-globulin	−10.5	−76.81	−51.61	−90.86	48.15
trypsin	49.48	−157.16	−3.76	−128.19	51.71

Table 3. Summaries of PLS Best Prediction Fits Relating Amino Acid Principal Property Sums of Proteins to Physicochemical Properties (Data from Ref *14*)

	z-sum models			ext-*z*-sum models		
property	comps[a]	*z*-sum *R*	*z*-sum *Q*	comps[a]	ext-*z*-sum *R*	ext-*z*-sum *Q*
Bigelow hydrophobicity	1	0.512	0	2	0.856	0.763
exposed hydrophobicity	2	0.701	0.339	2	0.932	0.862
viscosity	1	0.483	0.385	1	0.737	0.682

[a] Number of PLS components.

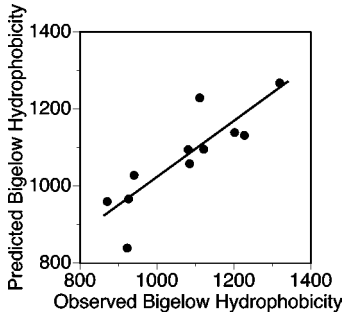

Figure 4. Model of Bigelow hydrophobicity data from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.856$; $Q = 0.763$.

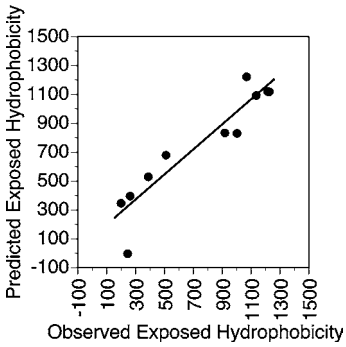

Figure 5. Model of exposed hydrophobicity data from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.932$; $Q = 0.862$.

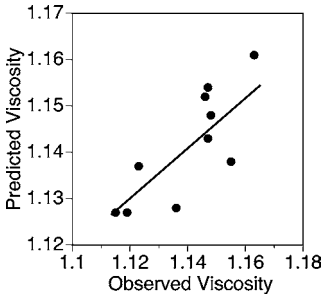

Figure 6. Model of viscosity data from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.737$; $Q = 0.682$.

Table 4. Regression Coefficients for the *z*-Score Sum Fits to Physicochemical Properties in **Table 3**

property	b_0	b_1	b_2	b_3
Bigelow hydrophobicity	7.452	–0.713	–0.167	–0.274
exposed hydrophobicity	1.768	–0.544	–0.327	–0.826
viscosity	72.107	0.084	–0.275	–0.240

Table 5. Regression Coefficients for the Ext-*z*-Score Sum Fits to Physicochemical Properties in **Table 3**

property	b_0	b_1	b_2	b_3	b_4	b_5
Bigelow hydrophobicity	7.452	–0.481	0.124	–0.361	–0.023	0.494
exposed hydrophobicity	1.768	–0.176	0.065	–0.566	–0.056	0.465
viscosity	72.107	0.002	–0.165	–0.233	–0.190	0.281

the fitted equations are shown in **Tables 4** and **5**. All three *z*-sum coefficients (**Table 4**) for the two protein hydrophobicity measures have negative signs, indicating lower hydrophobicity with increasing $\sum z_1$, $\sum z_2$, or $\sum z_3$ (corresponding to amino acid hydrophilicity, molecular size, and electronic properties). The viscosity model, on the other hand, had a positive sign for the $\sum z_1$ coefficient but a very small magnitude. The arithmetic signs

Table 6. VIP Values for the *z*-Score Sum Fits to Physicochemical Properties in **Table 3**

property	$\sum z_1$	$\sum z_2$	$\sum z_3$
Bigelow hydrophobicity			
exposed hydrophobicity	0.373	0.903	1.430
viscosity	0.389	1.273	1.109

Table 7. VIP Values for the Ext-*z*-Score Sum Fits to Physicochemical Properties in **Table 3**

property	$\sum z_1$	$\sum z_2$	$\sum z_3$	$\sum z_4$	$\sum z_5$
Bigelow hydrophobicity	1.041	0.662	1.056	0.664	1.386
exposed hydrophobicity	0.468	0.821	1.323	0.864	1.269
viscosity	0.009	0.833	1.174	0.960	1.417

of the ext-*z*-sum coefficients (**Table 5**) also had identical patterns for the two hydrophobicity measures (positive for $\sum z_2$ and $\sum z_5$ and negative for the other three). In the viscosity model, the signs were reversed for the $\sum z_1$ and $\sum z_2$ coefficients, although the $\sum z_1$ coefficient was again quite small.

The particular terms that are the most influential in the models can be represented by a calculation made by the SIMCA-S program that is called the VIP. The VIP values provide a better indication of the influence of terms on the response than comparing the magnitudes of the coefficients (*15*). The VIP values for the *z*-sum and ext-*z*-sum models are shown in **Tables 6** and **7**. The *z*-sum model for Bigelow hydrophobicity was too weak to consider ($Q = 0$). The exposed hydrophobicity was a better fit, and VIP values showed that $\sum z_3$ was most influential followed by $\sum z_2$. These parameters are thought to represent electronic properties and molecular size. The ext-*z*-sum VIPs (**Table 7**) show in part why the five term models were stronger than the three term models; the $\sum z_5$ term, considered among other things to represent electrophilicity, was very important in all three modeled quantities. The $\sum z_3$ term, representing electronic properties, was also very influential.

Of the three physicochemical properties modeled, viscosity had the poorest fit. Viscosity must surely depend on protein molecular size, one major aspect of which is the number of amino acids in the molecule, as well as flexibility. Neither is represented in this method of calculation, which may in part explain the relatively poorer modeling performance for this property. When an additional term indicating the number of amino acids in a protein was added, it resulted in a slightly stronger model for viscosity with the *z*-sum fit ($Q = 0.450$ vs 0.385) but a slightly weaker model for the ext-*z*-sum fit ($Q = 0.658$ vs 0.682). The *z*-sum term with the greatest influence on viscosity was $\sum z_2$ (**Table 6**), which represents amino acid molecular size, but the relationship was inverse (negative signed coefficient in **Table 4**). The ext-*z*-sum term with greatest influence on viscosity (**Table 7**) was $\sum z_5$ followed by $\sum z_3$ and $\sum z_4$.

Townsend and Nakai (*14*) also reported foam capacity measurements at different combinations of ionic strength and pH and showed logarithmic relations to hydrophobicity. Separate models of log(foam capacity) were constructed for each set of conditions. The results for the foam capacity at different ionic strengths, all at pH 7, are shown in **Table 8**. In each case, the ext-*z*-sum model was stronger (in both *R* and *Q*) than the equivalent *z*-sum model. With both *z*-score and ext-*z*-score models, the model strength (*R* and *Q*) improved with increasing ionic strength. The strongest model obtained was the ext-*z*-sum model at ionic strength 0.20 (see **Figure 7**). The VIPs for the

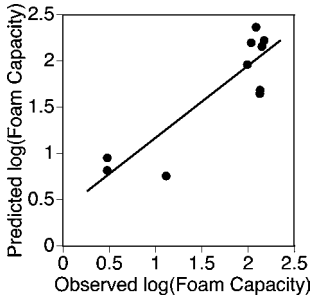

Figure 7. Model of foam capacity data for pH 7.0 and ionic strength 0.20 from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.880$; $Q = 0.831$.

Table 8. Summaries of PLS Best Prediction Fits Relating Amino Acid Principal Property Sums of Proteins to log(Foam Capacity) at Different Ionic Strengths at pH 7 (Data from Ref *14*)

ionic	*z*-sum models			ext-*z*-sum models		
strength	comps[a]	*z*-sum *R*	*z*-sum *Q*	comps[a]	ext-*z*-sum *R*	ext-*z*-sum *Q*
0.01	1	0.517	0.362	2	0.828	0.760
0.05	2	0.717	0.533	2	0.867	0.812
0.20	2	0.738	0.581	2	0.880	0.831

[a] Number of PLS components.

Table 9. VIP Values for the *z*-Score Sum Fits to log(Foam Capacity) in **Table 8**

ionic strength	$\sum z_1$	$\sum z_2$	$\sum z_3$
0.01	0.012	1.662	0.488
0.05	1.041	1.270	0.549
0.20	1.056	1.257	0.552

Table 10. Regression Coefficients for the Ext-*z*-Score Sum Fits to log(Foam Capacity) in **Table 8**

ionic strength	b_0	b_1	b_2	b_3	b_4	b_5
0.01	2.106	–0.317	–0.124	–0.149	–0.234	0.446
0.05	2.117	–0.343	–0.135	–0.186	–0.234	0.451
0.20	2.379	–0.358	–0.138	–0.202	–0.229	0.454

Table 11. VIP Values for the Ext-*z*-Score Sum Fits to log(Foam Capacity) in **Table 8**

ionic strength	$\sum z_1$	$\sum z_2$	$\sum z_3$	$\sum z_4$	$\sum z_5$
0.01	0.789	0.930	0.738	1.056	1.361
0.05	0.813	0.925	0.773	1.041	1.343
0.20	0.832	0.918	0.788	1.027	1.337

z-sum models are shown in **Table 9** and indicate that $\sum z_2$ (molecular size) had the greatest influence followed by $\sum z_1$ (hydrophilicity) except for the 0.01 ionic strength. The regression coefficients and VIP values for the ext-*z*-sum models are shown in **Tables 10** and **11**. Considering the VIP magnitudes and the arithmetic signs of the coefficients, it can be seen that the terms most influential in modeling foam capacity are in each case $\sum z_5$ (where the positive signed coefficient indicates increasing foam capacity with increasing electrophilicity), followed by $\sum z_4$ (here, the coefficient has a negative sign, indicating increasing foam capacity with decreasing electronegativity) and $\sum z_2$ (increasing foam capacity with decreasing molecular size). Once again, the utility of the fourth and fifth terms was evident.

Townsend and Nakai (*14*) also determined foam capacity at one ionic strength (0.05) at three different pH values. These

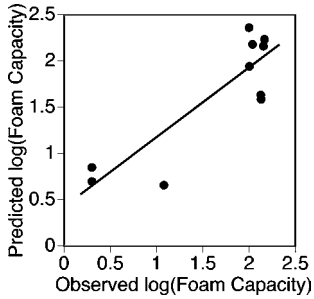

Figure 8. Model of foam capacity data for pH 7.0 and ionic strength 0.05 from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.866$; $Q = 0.812$.

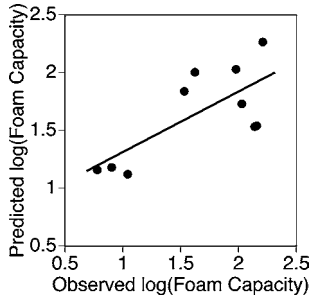

Figure 9. Model of foam capacity data for pH 9.0 and ionic strength 0.05 from ref *14* as a function of ext-*z*-score sums of all 20 coded amino acids. $R = 0.725$; $Q = 0.682$.

Table 12. Summaries of PLS Best Prediction Fits Relating Amino Acid Principal Property Sums of Proteins to log(Foam Capacity) at Different pH Values at Ionic Strength 0.05 (Data from Ref *14*)

	z-sum models			ext-*z*-sum models		
pH	comps[a]	*z*-sum *R*	*z*-sum *Q*	comps[a]	ext-*z*-sum *R*	ext-*z*-sum *Q*
5	1	0.650	0.556	1	0.672	0.547
7	2	0.716	0.535	2	0.866	0.812
9	1	0.420	0.237	1	0.725	0.682

[a] Number of PLS components.

Table 13. Regression Coefficients for the *z*-Score Sum Fits to log(Foam Capacity) at 0.05 Ionic Strength and Various pH Values in **Table 12**

pH	b_0	b_1	b_2	b_3
5	1.173	0.210	–0.393	–0.187
7	2.116	–0.524	–0.854	–0.113
9	2.958	0.056	–0.232	–0.231

data were also modeled (see **Table 12**). At the pH 5 condition, the best *z*-sum and ext-*z*-sum models had similar predictive abilities (Q). At the other two pH values, the ext-*z*-sum models were much stronger. The model with the highest Q was the ext-*z*-sum model at pH 7 (see **Figure 8**), followed by the ext-*z*-sum model at pH 9 (see **Figure 9**). The coefficients and VIPs for the *z*-sum models are shown in **Tables 13** and **14**. The relative VIP rankings showed the same patterns at the different pH values. $\sum z_2$ was most important (and inversely related with foam capacity) followed by $\sum z_1$ (although the signs were different at the different pH values).

The ext-*z*-sum coefficients and VIPs are shown in **Tables 15** and **16**. Here, too, the rankings of the VIPs were different at different pH values. At pH 5, the $\sum z_2$ and $\sum z_4$ VIPs were similar and much higher than the other terms (both were inversely related with foam capacity). At pH 7, $\sum z_5$ was most important,

Table 14. VIP Values for the *z*-Score Sum Fits to log(Foam Capacity) at 0.05 Ionic Strength and Various VIP Values in **Table 12**

pH	$\sum z_1$	$\sum z_2$	$\sum z_3$
5	0.753	1.408	0.671
7	1.040	1.272	0.548
9	1.205	1.210	0.290

Table 15. Regression Coefficients for the Ext-*z*-Score Sum Fits to log(Foam Capacity) at 0.05 Ionic Strength and Various pH Values in **Table 12**

pH	b_0	b_1	b_2	b_3	b_4	b_5
5	4.1731	0.154	–0.261	–0.048	–0.253	0.082
7	2.116	–0.343	–0.136	–0.185	–0.233	0.451
9	2.958	–0.016	–0.141	–0.245	–0.180	0.292

Table 16. VIP Values for the Ext-*z*-Score Sum Fits to log(Foam Capacity) at 0.05 Ionic Strength and Various VIP Values in **Table 12**

pH	$\sum z_1$	$\sum z_2$	$\sum z_3$	$\sum z_4$	$\sum z_5$
5	0.845	1.439	0.265	1.393	0.452
7	0.815	0.925	0.772	1.040	1.343
9	0.0798	0.709	1.231	0.906	1.469

with $\sum z_2$ and $\sum z_4$ similar in magnitude and quite a bit less influential. The $\sum z_1$ coefficient reversed sign as compared with pH 5. At pH 9, the $\sum z_5$ and $\sum z_3$ term VIPs were the largest. Obviously, the pH affects the charge on the amino acid side chains and that impacts their influence on foam capacity and modelability. The strength of the foam capacity models is definitely influenced by the conditions under which the measurements were made.

It was possible to model both physicochemical and functional properties of proteins from their contents of all 20 coded amino acids and amino acid principal properties. Clearly, it is possible to directly model protein functional properties from amino acid composition without the need to determine physicochemical properties. It appears that a number of other functional properties of proteins are likely to be modelable in this fashion.

LITERATURE CITED

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

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

(3) 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.

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

(5) Gill, S. C.; von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **1989**, *182*, 319−326.

(6) 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.

(7) Pomeranz, Y. *Functional Properties of Food Components*, 2nd ed.; Academic Press: New York, 1991.

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

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

(10) *Protein Functionality in Food Systems*; Hettiarachchy, N. S., Ziegler, G. R., Eds.; Marcel Dekker: New York, 1994.

(11) 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.

(12) MacRitchie, F. Physicochemical properties of wheat proteins in relation to functionality. *Adv. Food Nutr. Res.* **1992**, *36*, 1−87.

(13) Nakai, S.; Li Chan, E.; Hayakawa, S. Contribution of protein hydrophobicity to its functionality. *Nahrung* **1986**, *30*, 327−336.

(14) Townsend, A. A.; Nakai, S. Relationship between hydrophobicity and foaming characteristics of food proteins. *J. Food Sci.* **1983**, *48*, 588−594.

(15) Umetrics. *Users Guide to SIMCA-S Version 6.0*; Umetri AB: Umea, Sweden, 1996.

(16) Kinsella, J. E.; Srinivasan, D. Nutritional, chemical and physical criteria affecting the use and acceptability of proteins in foods. *Criteria of Food Acceptance*; Foster Verlag: Switzerland, 1981.

(17) Nakai, S.; Li-Chan, E.; Hirotsuka, M.; Vazquez, M. C.; Arteaga, G. Quantitation of hydrophobicity for elucidating the structure−activity relationships of food proteins. In *Interactions of Food Proteins*; Parris, N., Barford, R., Eds.; American Chemical Society: Washington, DC, 1991; pp 42−58.

(18) Nakai, S.; Powrie, W. D. Modification of proteins for functional and nutritional improvements. In *Cereals: A Renewable Resource: theory and practice*; Munck, L., Ed.; American Association of Cereal Chemists: St. Paul, MN, 1981; pp 217−242.

(19) 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 & Professional: London, 1994; pp 1−21.

Received for review March 20, 2003. Revised manuscript received September 23, 2003. Accepted October 9, 2003. 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.

JF0342775