# Quantitative Structure−Activity Relationship Modeling of Peptide and Protein Behavior as a Function of Amino Acid Composition

Karl J. Siebert[†]

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

A quantitative structure−activity relationship (QSAR) modeling approach based on the location of each amino acid along three axes obtained by principal component analysis (called *z* scores) was extended to physical and functional properties of proteins, where the proportion of particular amino acids rather than a precise sequence is the determining factor. Coomassie Brilliant Blue spectral responses to amino acid homopolymers ($R = 0.926$) and proteins, either as a function of their contents of six basic and aromatic amino acids ($R = 0.976$) or as a function of the contributions of these amino acids to the three *z* scores ($R = 0.935$), were modeled. The ultraviolet absorbance of proteins was modeled in terms of the *z* score contributions of tyrosine, tryptophan, and cysteine ($R = 0.995$). Modeling many protein functional properties in this manner appears to be possible. An approach to modeling peptide behaviors that depend on short sequences of amino acids was also considered.

**Keywords:** *QSAR; functional properties; principal component analysis; partial least-squares regression*

## INTRODUCTION

Peptides and proteins participate in a vast array of biological functions. They provide physical structure, enzyme activity, and transport across membranes and within organisms. They also act as hormones, enzyme inhibitors, antibodies, olfaction and taste receptors, and antimicrobial compounds.

In foods, peptides (including proteins) provide nutrition and influence organoleptic and "functional" properties. The functional properties have been described as solubility, viscosity, gelation, emulsification, and foam formation (*1*). Peptides can also form films and glasses and contribute to color and flavor. Other authors add to this list binding of flavor compounds and water (which affects viscosity and gelation), modification of surface tension and interfacial activity (which influence emulsification and foaming ability), and cohesive/adhesive properties (which affect texturization) (*2*).

For some peptide properties a precise amino acid sequence is required for a particular function. In other cases, and particularly with the functional properties described above, behavior is more dependent on the relative proportion of a particular amino acid or class of amino acids (e.g., acidic, basic, hydrophilic, hydrophobic, and aromatic). For example, hydrophobicity, either in a domain or of an entire protein, is associated with foaming, gel formation, and binding of nonpolar flavor compounds (*3*, *4*).

It has for some time been of interest to try to relate peptide structure to biological or functional properties, and this has been accomplished with some success. A fairly recent development of particular interest in the quantitative structure−activity relationship (QSAR) field is the use of amino acid "*z* scores" obtained by principal component analysis (PCA) of property data (*5*, *6*). The three *z* scores for each amino acid express its relationship to the other amino acids in terms that have chemical meaning. These primarily represent hydrophilicity (or polarity), side-chain bulk (molecular size), and electronic properties. The *z* scores have proven to be useful for modeling a number of biological effects of small peptides as a function of the *z* score values of the amino acids in each position in a peptide (*6*). For modeling the bitterness of dipeptides, for example, this resulted in a 6-term model (three *z* scores for each of the two amino acid positions), and for bradykinin potentiating activity of pentapeptides, a 15-term model resulted. More recently, the same approach was expanded to a larger set of amino acids (20 coded + 67 noncoded) and more parameters. Application of PCA resulted in a set of five orthogonal variables termed *zz* scores, of which the first three corresponded to the original *z* scores. The *zz* scores were applied to two peptide data sets, elastase substrates and neurotensin analogues, and performed well (*7*). Classification of *Escherichia coli* proteins according to cellular localization was accomplished by using auto-cross-covariances of amino acid sequence *z* scores (*8*); this approach enabled comparison of proteins of differing lengths. Obviously, as the length of a peptide increases, the number of terms and complexity of a property model rapidly increase. Due to practical limitations of modeling this means that the number of peptides for which data would be needed to build a model could soon exceed reasonable possibility.

It is, however, of interest to consider if there might not be special cases in which merely the proportion of a particular amino acid or class of amino acids in a peptide determines a property. In those instances the modeling of peptide behavior could presumably be simplified to the point of manageability.

## MATERIALS AND METHODS

The amino acid *z* scores used were those reported by Jonsson et al. (*6*).

The Coomassie Brilliant Blue (CBB) dye binding data for amino acid homopolymers are from Compton and Jones (*9*). Those for proteins are from Sedmak and Grossberg (*10*).

[†] Telephone (315) 787-2299; fax (315) 787-2284; e-mail kjs3@cornell.edu.

10.1021/jf000718y CCC: $20.00 © 2001 American Chemical Society
Published on Web 12/13/2000

The 280 nm absorbance data and formula for estimating protein UV molar absorptivity are from Gill and von Hippel (*11*).

Modeling of the property data as a function of amino acid composition (expressed as moles per mole of protein, as mole percent, or as *z* scores) was carried out by partial least-squares regression (PLS) using the SIMCA-S for Windows computer program (Umetrics Inc., Kinnelon, NJ), which also provided estimates of model fit (the multiple correlation coefficient, *R*) and predictive ability (the cross-validated correlation coefficient, *Q*). When it was necessary to linearize a response, the natural logarithm transformation was applied.

## RESULTS AND DISCUSSION

Some protein functions are thought to result from very specific localized configurations (e.g., enzyme active sites, olfaction and taste receptors, and antigen recognition sites on antibodies). In such cases the precise amino acid sequence, at least of a region of the peptide, and usually the three-dimensional configuration are thought to be important to the activity. Often, however, at least a portion of the protein is not essential to the function. This has been shown by demonstrations that synthetic constructs resembling enzyme active sites have catalytic effects similar to those of native enzymes (*12*).

Many of the functional effects of proteins, which typically are bulk physical properties, are thought not to be highly dependent on amino acid sequence but more generally a characteristic of the proportions of certain amino acids or amino acid classes (e.g., basic, acidic, hydrophobic, hydrophilic, or aromatic amino acids) in a peptide (*3*, *13*, *14*).

A number of approaches have been taken to relate protein composition to functional properties. Protein characterizations have included amino acid composition, protein geometry (globular versus fibrillar, proportions of α-helix versus β-pleated sheet conformation, etc.) or responses to specific dyes (such as those considered to indicate hydrophobicity). Functional properties are often assessed by direct physical measurements such as viscosity, light scattering, or foam performance in either model systems or foods (*15*).

Nearly all of the interactions of peptides in biological systems are noncovalent. These include enzyme functions (catalysis), binding to specific sites on cells (hormone receptors), binding to antigens (antibody function), binding to small molecules for transport or for active permeation, and binding to taste or olfaction receptors. These interactions occur with from modest to great specificity. Because there are only a limited number of mechanisms for noncovalent interactions, it seems likely that various combinations of these interactions are involved in different proportions in particular protein properties, and because this range of interactions should stem from a peptide's amino acid composition, it should influence chemical, biological, and physical properties.

In the case of small peptides an approach to modeling relationships between structure and function has been taken that has proven useful. Large amounts of structural and property data for each of the coded amino acids (*5*) as well as a number of the noncoded ones (*6*) were obtained. Principal component analysis (PCA) was applied and indicated that three principal components accounted for 84% of the variance in the property data. The three PCs represent independent (due to the nature of PCA, the PCs are orthogonal and thus uncorrelated) properties that appear to be closely aligned with the chemical concepts of hydrophilicity and molecular size and with electronic properties (this component was heavily influenced by NMR data). The three PC score values for each amino acid (which have been designated *z* scores, see Table 1) define its position along each of the three axes. The *z* scores were used to model the bitterness of 48 dipeptides (*6*). This led to a six-term model where the terms were the $z_1$, $z_2$, and $z_3$ scores for the amino acid in the N-terminal position (position 1) and the $z_1$, $z_2$, and $z_3$ scores for the amino acid in the C-terminal position (in this case position 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 sensory bitterness and the *b* values are the fitted coefficients. This model of dipeptide bitterness was quite successful, with a correlation coefficient of 0.88 (see Figure 1).

The same PC scores were used to model the bradykinin potentiating activity of pentapeptides (*6*). In this case the model had 15 terms, corresponding to the three

**Table 1. Amino Acid *z* Scores[a] (from Reference *6*)**

| amino acid | code[b] | $z_1$ score | $z_2$ score | $z_3$ score |
|---|---|---|---|---|
| Ala | A | 1.13 | −2.36 | 1.26 |
| Arg | R | 3.21 | 2.31 | −3.32 |
| Asn | N | 2.88 | 1.47 | 1.55 |
| Asp | D | 2.33 | 0.70 | 1.46 |
| Cys | C | −0.86 | −1.36 | 2.44 |
| Glu | E | 1.07 | −0.07 | −1.10 |
| Gln | Q | 1.14 | 0.27 | −1.15 |
| Gly | G | 2.39 | −4.12 | −0.42 |
| His | H | 3.49 | 2.44 | 0.34 |
| Hos |  | 0.99 | −0.85 | 0.44 |
| Hpr |  | 1.13 | 2.45 | 4.36 |
| Ile | I | −3.48 | −1.69 | −1.10 |
| Leu | L | −3.66 | −1.22 | −0.54 |
| Lys | K | 3.76 | 1.16 | −2.32 |
| Met | M | −2.88 | 0.13 | 0.31 |
| Nle |  | −3.71 | −1.31 | −0.55 |
| Nvl |  | −2.51 | −1.85 | −0.05 |
| Phe | F | −3.62 | 1.80 | 0.97 |
| Pro | P | 0.02 | 0.47 | 3.06 |
| Ser | S | 2.48 | −1.16 | 1.24 |
| Thr | T | 0.55 | −2.23 | −1.49 |
| Trp | W | −4.02 | 4.09 | 0.18 |
| Tyr | Y | −3.58 | 2.07 | −0.04 |
| Val | V | −2.27 | −2.67 | −1.32 |

[a] $z_1$ = hydrophilicity; $z_2$ = side chain bulk; $z_3$ = electronic properties. [b] Single-letter codes for amino acids.

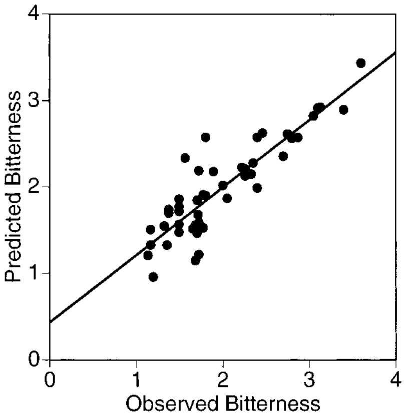

**Figure 1.** Six-term model of bitterness of dipeptides as a function of amino acid *z* scores (*R* = 0.88). Reproduced with permission from *Quant. Struct.-Act. Relat.* **1989**, *8*, 204−209. Copyright 1989 Wiley.

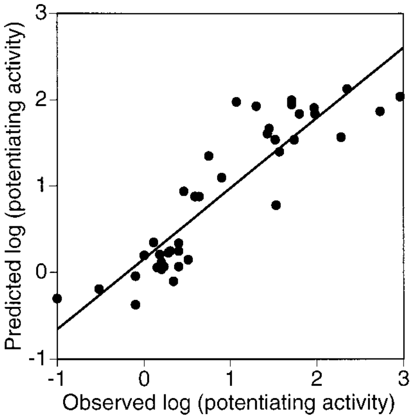

**Figure 2.** Fifteen-term model of bradykinin potentiating activity of pentapeptides as a function of amino acid *z* scores ($R = 0.90$). Reproduced with permission from *Quant. Struct.-Act. Relat.* **1989**, *8*, 204−209. Copyright 1989 Wiley.

*z* scores for the amino acids in each of the five positions in the sequence:

$$y = b_{11}z_{11} + b_{21}z_{21} + b_{31}z_{31} + ... + b_{15}z_{15} + b_{25}z_{25} + b_{35}z_{35} \quad (2)$$

This, too, was quite successful, with a correlation coefficient of 0.90 (see Figure 2).

The general form of the two models can be described by

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

where *y* is the modeled property and *n* is the length of the peptide chain.

The same approach was also used to successfully model both the oxytocic activity (88% of the variance) and the pressor activity (64% of the variance) of oxytocins, the inhibition of pepsin by pepstatins (80% of the variance) (*5*), and the inhibition of angiotensin converting activity (*16*). An attempt to model the oncostatic activity of a set of pseudopeptides was not successful.

It is conceivable to extend eq 3 to larger peptides, but the number of terms (3*n*) quickly becomes overwhelming. In the case of a peptide containing 100 amino acids (on the order of 15 kDa), there would be 300 terms to fit. With multiple linear regression fitting, good modeling practice would require data from at least 600, and preferably 900−1200, different peptide sequences to fit all of the coefficients (*17*). Clearly, this is not feasible. Stepwise multiple regression or, especially, PLS would greatly reduce the data requirement, but this would still be unwieldy for proteins. It is, however, useful to consider how modeling based on the *z* scores might be applied to larger molecules in some special cases.

A model of the type shown in eq 3 describes only the primary structure (amino acid sequence) of a peptide. Clearly, this is sufficient for the small peptides that have been modeled, as the models have quite good fits to the biological properties. This could occur either because the more complicated aspects of structure in these cases mainly are determined by the primary sequence or because the important aspect of structure for the biological activity depends on only a rather small region of the peptide (such as a few contiguous or neighboring amino acids in a sequence), so that three-dimensional (3D) structure is not so important. It is increasingly possible to successfully predict polypeptide 3D structures (regions of α-helix, β-pleated sheet, etc.) from primary amino acid sequences (*18*), indicating that the 3D structure is, to at least some extent, determined by the primary sequence.

Functional properties of peptides and proteins have often been determined empirically, by direct measurement of a physical property in a model system or an actual food as a function of protein type and concentration, matrix pH, composition, etc. In some cases it has been possible to find relationships between a functional property and some convenient measurement. For example, emulsifying and fat binding capacities have been related to hydrophobicity determined by response to some fluorescent dyes (*3*).

It has been shown in some cases that the physical and sensory properties of proteins and larger peptides are related to their content of particular amino acids (expressed either as number of moles or mole percent) rather than their precise sequence. Two examples of this are the foaming activity of beer proteins, which has been reported to be related to the content of basic amino acids (*19*), and the polyphenol-binding activity of proline-rich proteins (*20*, *21*). In both of these cases there is an interaction between particular amino acids in the peptide and a small molecule that results in a physical phenomenon.

**Property Dependent on Content of a Single Amino Acid—Polyphenol Binding.** Binding of polyphenols by proteins is responsible for the formation of haze in beverages including beer, wine, and fruit juices (*21*), perception of astringency in the mouth (*22*, *23*), and defense against the antinutritional effect of polyphenols (*23*, *24*).

The most frequent cause of haze (turbidity) formation in beer, wine, and fruit juices arises from the combination of proteins with polyphenols to form colloidal size particles that scatter light (*25*). Only peptides that contain proline demonstrate haze-forming activity, whereas peptides that lack it fail to form haze (*20*, *26*). As in the case of foam, the physical property is closely related to the amino acid composition of the peptide. Results from a model system, in which peptides were combined with catechin in buffer and heated, then followed by turbidity assessment, showed an essentially linear relationship between mole percent of proline and haze formed (see Figure 3).

In this case the proline content of a peptide was the main property that mattered and the activity was essentially independent of molecular size and the content of other amino acids. This would be equivalent to eq 3 with *b* terms for all amino acids other than proline equal to 0 or

$$y = \frac{m}{n}\sum_{i=1}^{3} b_i z_{i\mathrm{P}} \quad (4)$$

where *m* is the number of proline residues in a peptide with a total length of *n* amino acids and $z_{i\mathrm{P}}$ are the three *z* scores for proline. This could result from substituting *b* values of 0 for all amino acids other than proline in eq 3, if it is assumed that the position in the amino acid sequence which proline occupies is irrelevant, even if it falls in the C-terminal or N-terminal location (which might well behave differently).

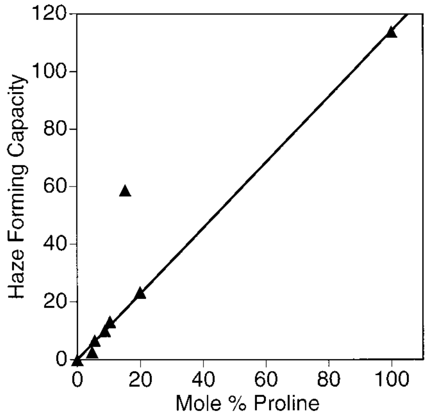

**Figure 3.** Relationship between the mol % of proline in synthetic polypeptides and natural proteins and haze formed in a model system with catechin at 100 °C. Reproduced with permission from *J. Agric. Food Chem.* **1999**, *47*, 353−362. Copyright 1999 American Chemical Society.

Free proline does not produce haze with polyphenols or compete against proteins for polyphenol binding (*25*). This could be because ionization of the carboxy and/or amino group interferes with the interaction with polyphenols (i.e., only peptidically linked proline is active). If prolines occurred in the terminal positions and were excluded, the relationship would be

$$y = \frac{m-2}{n}\sum_{i=1}^{3} b_i z_{i\mathrm{P}} \qquad (5)$$

There is evidence that peptides with as few as 15 amino acids combine with polyphenols (*27*), and the critical size could well be smaller yet. In any case, the relationship would still be generally similar to eq 4, particularly for larger peptides:

$$y \approx \frac{m}{n}\sum_{i=1}^{3} b_i z_{i\mathrm{P}} \qquad (6)$$

The situation can be conceptualized geometrically as a three-dimensional space defined by the *z* scores, in which each amino acid has its own vector (from the 0,0,0 point). Peptide properties can be projected into the same 3D space, also as vectors. The similarity of alignment of an amino acid vector with a property vector reflects the contribution of that amino acid to the property. The magnitude of the contribution is influenced by the proportion of the amino acid of an effective type in the peptide; this can be conceptualized as influencing the length of the amino acid vector.

**Property Dependent on Content of Several Amino Acids.** *CBB Dye Binding of Amino Acid Homopolymers.* Another special case is a homopolymer, where the amino acid in each position is the same. This is equivalent to eq 4, where $m = n$ and *k* represents the particular amino acid of which the homopolymer is built:

$$y = \sum_{i=1}^{3} b_i z_{ik} \qquad (7)$$

If this concept is correct, then some of the activities of polypeptides that depend on the amino acid composition should be modelable. CBB dye binding has been widely used as a method of measuring proteins larger than ~5000 Da (*28*), but it is well-known to be heavily biased in favor of basic and aromatic amino acids (*9*). This was demonstrated by determining the relative response of the dye to a variety of amino acid homopolymers (Table 2). When the amino acid *z* scores of Jonsson et al. (*6*; Table 1) were used to model the data in Table 2, the results in Figure 4 and eq 8 were obtained.

**Table 2. CBB Binding Responses of L-Amino Acid Homopolymers (from Reference *9*)**

| polymeric amino acid | mol wt (kDa) | relative response[a] |
|---|---|---|
| poly(Arg) | 40 | 36.0 |
| poly(Tyr) | 100 | 4.7 |
| poly(Try) | 5 | 4.4 |
| poly(His) | 11 | 4.2 |
| poly(Phe) | 15 | 1.9 |
| poly(Lys) | 35 | 1.0 |
| poly(Ala) | 25 | 0.0 |
| poly(Gly) | 6 | 0.0 |
| poly(Asn) | 9 | 0.0 |
| poly(Asp) | 20 | 0.0 |

[a] Absorbance difference at 595 nm relative to poly(Lys) = 1.0.

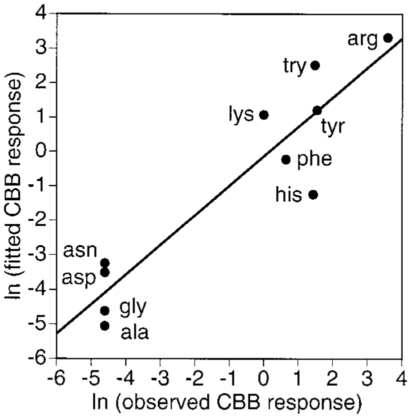

**Figure 4.** Model of CBB dye binding to amino acid homopolymers (response data from ref *9*) as a function of amino acid *z* scores from Table 1 ($R = 0.926$, $Q = 0.799$).

$$\ln(\text{rel CBB response}) = -1.446 - 0.3174z_1 + 0.7129z_2 - 1.246z_3 \qquad (8)$$

This model has quite a respectable correlation ($R = 0.926$) and cross-validated correlation ($Q = 0.799$), which indicate that the amino acid *z* scores are effective in modeling at least this property of medium to large peptide homopolymers. The *R* value indicates how well the model fits the data, whereas the *Q* value is an indicator of its predictive value for other samples. The model indicates that the greatest contribution to CBB response comes from a low $z_3$ value (electronic properties), as exhibited by arginine and lysine, and next by high $z_2$ (side chain bulk), for which the greatest contributor is tryptophan.

*CBB Dye Binding of Proteins.* If, as indicated in Table 2, only six amino acids influence CBB response, then it should also be possible to predict the CBB response of a protein from consideration of the proportion of each critical amino acid. CBB responses for a number of proteins were previously reported (see Table 3) (*10*). The contents of the basic and aromatic amino acids of these proteins were obtained from the literature (also see Table 3). Two approaches to modeling CBB binding were attempted. In the first, the mole percent of each of the six amino acids in each of the proteins was used. This resulted in a model of ln(CBB response) with $R = 0.976$ and $Q = 0.572$ (see Figure 5):

**Table 3. CBB Binding Responses of Proteins in Aqueous Perchloric Acid (PA) (from Reference *10*)[a]**

| protein | $n$[b] | no. of Arg (R) | no. of Lys (K) | no. of His (H) | no. of Phe (F) | no. of Tyr (Y) | no. of Trp (W) | CBB (Abs in PA) |
|---|---|---|---|---|---|---|---|---|
| plasma albumin (bovine) | 607 | 26 | 60 | 17 | 30 | 21 | 3 | 0.429 |
| fetuin | 359 | 13 | 15 | 11 | 12 | 8 | 2 | 0.349 |
| ovalbumin (egg white) | 386 | 15 | 20 | 7 | 20 | 10 | 3 | 0.357 |
| pepsin (porcine stomach) | 326 | 2 | 1 | 1 | 14 | 16 | 5 | 0.163 |
| trypsin inhibitor (soybean) | 181 | 9 | 10 | 2 | 9 | 4 | 2 | 0.350 |
| lysozyme (egg white) | 129 | 11 | 6 | 1 | 3 | 3 | 6 | 0.348 |
| cytochrome *c* (horse heart) | 104 | 2 | 19 | 3 | 4 | 4 | 1 | 0.465 |
| trypsin inhibitor (bovine pancreas) | 58 | 6 | 4 | 0 | 4 | 4 | 0 | 0.478 |
| insulin (porcine) | 51 | 1 | 1 | 2 | 3 | 4 | 0 | 0.437 |
| glucagon (porcine) | 29 | 2 | 1 | 1 | 2 | 2 | 1 | 0.469 |

[a] The amino acid contents are from sequence data reported on the protein database (swissprot) of the National Center for Biotechnology Information web site (http://www.ncbi.nlm.nih.gov/). [b] Total number of amino acids in the protein.

**Table 4. Data for 280 nm Molar Absorptivity of Proteins (from Reference *11*)**

| protein | mol of Trp ($n_W$) | mol of Tyr ($n_Y$) | mol of Cys ($n_C$) | molar absorptivity | |
|---|---|---|---|---|---|
| | | | | measured[a] | predicted[b] |
| aldolase (rabbit muscle) | 3 | 12 | 8 | 35074 | 33390 |
| alcohol dehydrogenase (yeast) | 5 | 14 | 8 | 48093 | 47330 |
| carboxypeptidase A (bovine) | 7 | 19 | 2 | 64698 | 64390 |
| carboxypeptidase B (bovine) | 8 | 22 | 7 | 72696 | 74520 |
| chymotrypsinogen A (beef pancreas) | 8 | 4 | 10 | 51725 | 51840 |
| glyceraldehyde-3-phosphate dehydrogenase (yeast) | 3 | 11 | 2 | 31775 | 31390 |
| glutamate dehydrogenase (bovine) | 4 | 18 | 6 | 51480 | 46520 |
| insulin (bovine) | 0 | 4 | 6 | 5677 | 5840 |
| lac repressor (*E. coli*) | 2 | 8 | 3 | 23190 | 21980 |
| α-lactalbumin (bovine) | 4 | 4 | 8 | 28796 | 28840 |
| *β*-lactoglobulin (bovine) | 2 | 4 | 5 | 17581 | 17100 |
| lysozyme (hen egg white) | 6 | 3 | 8 | 37825 | 38940 |
| lysozyme (T4) | 3 | 6 | 2 | 23900 | 24990 |
| ovalbumin (chicken) | 3 | 10 | 6 | 29972 | 30590 |
| papain | 5 | 19 | 7 | 58570 | 53610 |
| ribonuclease A (beef pancreas) | 0 | 6 | 8 | 9824 | 8640 |
| serum albumin (bovine) | 2 | 20 | 35 | 43962 | 41180 |
| serum albumin (human) | 1 | 18 | 35 | 35761 | 32930 |
| trypsinogen (bovine) | 4 | 10 | 12 | 33357 | 37000 |

[a] Where more than one result for a protein was reported by Gill and von Hippel (*11*), these are averages. [b] Calculated as protein molar absorptivity$_{280}$ = $5690n_W + 1289n_Y + 120n_C$.

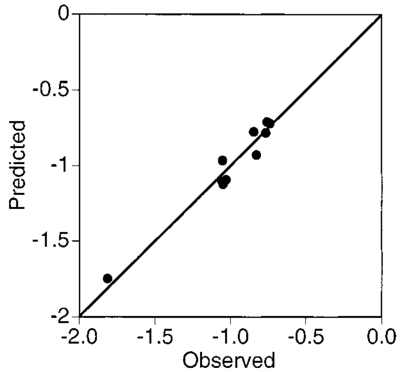

**Figure 5.** Model of CBB dye binding responses of proteins (data from ref *10*) modeled as a function of their contents of six basic and aromatic amino acids ($R$ = 0.976, $Q$ = 0.572).

$$\ln(\mathrm{CBB}) = -1.8813 + 0.081314\%_R + 0.025067\%_K + 0.15425\%_H + 0.016488\%_F + 0.0053052\%_Y - 0.043868\%_W \quad (9)$$

where the $\%_i$ values represent the mole percents of the amino acids indicated by the single-letter codes (see Table 1).

In the other approach, the contribution of each amino acid to each of the three $z$ scores was estimated. In this case the number of moles, $n$, of each of the six relevant amino acids in the peptide was multiplied by each of the $z$ scores for that amino acid and these were summed with the products for the same $z$ score from the other amino acids. The $z_1$ sum in this case would be

$$\sum z_1 = z_{1R}n_R + z_{1K}n_K + z_{1H}n_H + z_{1F}n_F + z_{1Y}n_Y + z_{1W}n_W \quad (10)$$

and the general case could be expressed as

$$\sum z_i = \sum_{j=1}^{6} n_j z_{ij} \quad (11)$$

This was repeated for each of the $z$ scores, resulting in three $z$ score sums. These too were quite successful in modeling CBB response (see Figure 6):

$$\ln(\mathrm{CBB}) = -0.7675 + 0.004741\sum z_1 - 0.005024\sum z_2 - 0.00248\sum z_3 \quad (12)$$

The correlation ($R$ = 0.935) is slightly weaker in this case, but this is offset by the simpler model (fewer terms always resulting in less error in fitting coefficients) and often, as in this case, better predictive ability ($Q$ = 0.890). According to this equation, low $z_2$ (as found in lysine, phenylalanine, and tyrosine) and high $z_1$ (exhibited by lysine, histidine, and arginine) make the greatest contributions to CBB response.

**Ultraviolet Absorbance of Proteins.** Another widely used property of proteins is light absorbance near 280

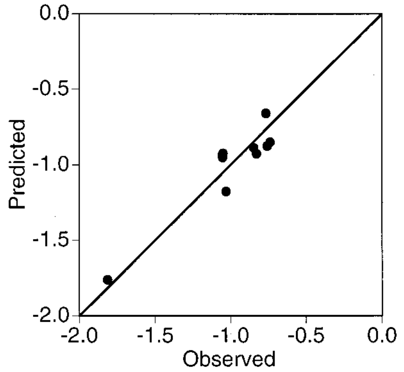

**Figure 6.** Model of CBB dye binding responses of proteins (data from ref *10*) modeled as a function of the *z* score contributions of six basic and aromatic amino acids ($R = 0.935$, $Q = 0.890$).

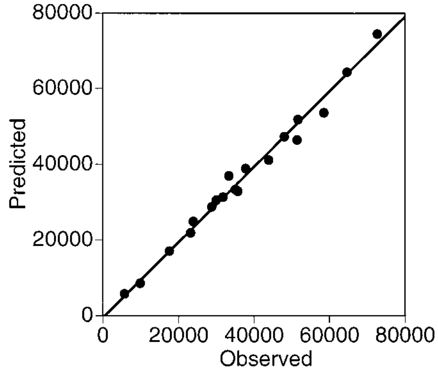

**Figure 7.** Comparison of protein 280 nm molar absorptivity measured and predicted (from contents of Tyr, Trp, and Cys). Data and model from ref *11* ($R = 0.995$, $Q = 0.988$).

nm. This has previously been shown to depend on only three amino acids (cysteine, tyrosine, and tryptophan), and it was seen that from knowledge of the number of each of these in a protein and their respective molar absorptivities, it was possible to very successfully model protein UV molar absorptivity, $\epsilon$ (*11*):

$$\epsilon_{280} = 5690n_{\mathrm{W}} + 1289n_{\mathrm{Y}} + 120n_{\mathrm{C}} \qquad (13)$$

where $n_i$ represents the number of moles of amino acid $i$ in the protein (see Figure 7). This resulted in a very good fit ($R = 0.995$) with high predictive ability ($Q = 0.988$).

The data fit equally well to the form of eq 3 using the three *z* scores for each of the three relevant amino acids ($R = 0.995$). Although that approach employed nine terms and had the disadvantage of increased error in fitting more terms, in this case the predictive ability did not decline ($Q = 0.988$).

This situation was also approached by calculating the *z* score sums for each of the three critical amino acids. The model produced (see eq 14 and Figure 8)

$$\epsilon_{280} = -59.569 + 693.18\sum z_1 + 1930.58\sum z_2 + 1367.97\sum z_3 \qquad (14)$$

was as good as that originally described ($R = 0.995$) and had a slightly higher predictive ability ($Q = 0.992$). The greatest contributions to absorbance come from high $z_2$ (tryptophan and tyrosine) and high $z_3$ (cysteine) values.

In the geometric conceptualization of this modeling approach, the similarity of alignment of an amino acid vector with a property vector projected into the *z* score space indicates the extent to which an amino acid contributes to the property. Presumably in a case like that of proline in polyphenol interaction, the alignment between the proline vector and the polyphenol interaction vector is very close. In a multiple interaction situation like that of 280 nm absorption, the vector is presumably intermediate between the vectors of cysteine, tyrosine, and tryptophan, but closer to the latter two (because of their much stronger influence).

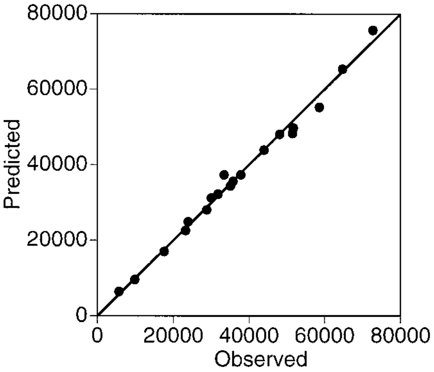

**Figure 8.** Comparison of protein 280 nm molar absorptivity measured and predicted (from *z* score sums of Tyr, Trp, and Cys). Data from ref *11* ($R = 0.995$, $Q = 0.992$).

**Property Dependent on a Short Amino Acid Sequence.** Suppose that a peptide "recognition site" or "specificity site" depends on two (or several) amino acids in sequence in a larger polymer and that any of these amino acids in isolation is ineffective. Furthermore, suppose that there could be multiple occurrences of the "significant sequence" in a peptide, especially in tandem repeats such as are well-known in the salivary proline-rich proteins (*23*) and in the "antifreeze" proteins of cold water fish (*29*). Then the activity should depend (at least once the critical size is passed) on the proportion (or frequency of occurrence) of the significant sequence.

For a two amino acid significant sequence and eq 3, all of the *b* values would be 0 except when the two critical amino acids were adjacent in the correct sequence.

$$y = b_{1i}z_{11} + b_{2i}z_{21} + b_{3i}z_{31} + b_{1i}z_{12} + b_{2i+1}z_{22} + b_{3i}z_{32} \qquad (15)$$

where $i$ represents any position in the amino acid sequence and $i$+1 is the adjacent position, and there could be multiple occurrences of this sequence

$$y = \frac{m}{n}\sum_{i=1}^{3}(b_{ji}z_{j1} + b_{ji+1}z_{j2}) \qquad (16)$$

where $m$ indicates the number of occurrences of the significant sequence in a peptide of length $n$ amino acids, $z_{j1}$ for $j = 1-3$ represents the three *z* scores for the first amino acid in the critical sequence, and $z_{j2}$ for $j = 1-3$ represents the three *z* scores for the second amino acid in the critical sequence.

Then for a peptide in which the significant sequence is longer, the situation would be similar

$$y = \frac{m}{n}\sum_{j=1}^{3}(b_{ji}z_{j1} + ... + b_{ji+k}z_{jk}) \qquad (17)$$

where there are $k + 1$ amino acids in the critical sequence, or

$$y = \frac{m}{n}\sum_{j=1}^{3}\sum_{l=1}^{k} b_{ji}z_{ji+l} \qquad (18)$$

**Limitations of the Approach.** There is one situation in which the approach described here would not lead to good models (low *R* and/or Q) for either peptides or proteins. That is the case where the nature of the protein interaction in question is not represented, or not represented well, by the properties described by the *z* scores. Recent efforts to broaden the *z* scales by extending the approach to more amino acids (67 noncoded plus 20 coded) and up to five scales appear to have resulted in additional useful information (*7*).

For proteins or longer peptides another problem could occur. That is the case when a critical amino acid for a property is not more or less uniformly distributed throughout the peptide length (as it certainly is in a homopolymer) but rather is concentrated in a region that is folded into the interior. Although the critical amino acid would be assessed by composition or sequence analysis, it might be inaccessible for participation in a chemical interaction. This should not be a problem for spectral characteristics (such as 280 nm absorbance) or in analyses when the protein is largely denatured (as in the catechin binding experiments, which involved heating to 100 °C). Small molecules, such as dyes, may also be able to penetrate, at least somewhat, into a protein structure, but tight, globular protein domains might be inaccessible.

Still another situation in which the approach would fail would be when an interaction site is formed by amino acids that are not close together in the linear sequence of the peptide but are physically close in the folded structure.

**Dye Binding Response as a Property Indicator.** In a number of cases, protein dye binding response, where a dye binds to a protein and undergoes a chromic shift, has been found to be a convenient marker of some functional or compositional property. Examples include dyes that bind preferentially to hydrophobic proteins (*3*) or those that bind to lysine-rich proteins (*15*). In these cases both the dye binding response and the functional or other property are presumably dependent variables that are functions of the same aspect of amino acid composition of the proteins and, thus, correlated with one another. This can be conceptualized as two vectors with similar alignments in the same 3D *z* space. For example, in beer the predominant protein present originates from barley hordein, is very poor in basic and aromatic amino acids, and gives very little response to CBB, whereas the much less prominent foam active protein, known to be rich in basic amino acids, responds strongly (*30*). In this case both the CBB dye response and the foaming activity are strongly influenced by the basic amino acid content of the protein.

**Future Possibilities.** It may be possible to select a set of dyes that could be used to characterize different aspects of a test protein. Results from its application could reveal information both about the general nature of the protein (e.g., indicating the presence of regions of basic, aromatic, acidic, nonpolar, or aromatic amino acids) and about the functional properties that are known to be related to these characteristics. This could be useful for screening proteins for possession of a desirable functional property or combination of functional properties. It could also have application in formulating a food product with desirable characteristics. A more remote possibility would be its use as a guide in the modification of a protein to enhance certain characteristics or even in the design of a peptide sequence with desirable functionality.

**Conclusions.** The amino acid *z* scores, which were previously used to develop sequence-dependent models of small peptide biological properties, were successfully applied to model some polypeptide properties. A model of CBB dye binding response of amino acid homopolymers was developed. The *z* scores were also employed to model behavior that is mainly a function of the proportion of one or a few amino acids in a peptide. This was demonstrated with polyphenol binding activity as a function of proline content, and ultraviolet light absorption as a function of a protein's contents of tyrosine, tryptophan, and cysteine. The CBB dye binding response of proteins as a function of the three basic and three aromatic amino acids was also demonstrated. In all of these cases it was possible to develop models using as many terms as the relevant amino acids, but it was also possible to use sums of the three *z* score contributions of each of the relevant amino acids to produce a simpler model (fewer terms). It appears that a number of other functional properties of proteins are likely to be modelable in this fashion.

It may be possible, although it is not yet proven, to extend the modeling approach described here to encompass situations where the property of interest is a short amino acid sequence (from two to a modest number of amino acids) that may or may not recur in the peptide.

## LITERATURE CITED

(1) *Protein Functionality in Food Systems*; Hettiarachchy, N. S., Ziegler, G. R., Eds.; Dekker: New York, 1994; p 519.

(2) 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.

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

(4) Phillips, L. G.; Whitehead, D. M.; Kinsella, J. *Structure−Function Properties of Food Proteins*; Academic Press: San Diego, CA, 1994.

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

(6) 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.

(7) 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.

(8) Sjöström, M.; Rännar, S.; Wieslander, Å. Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. *Chemom. Intell. Lab. Syst.* **1995**, *29*, 295−305.

(9) Compton, S. J.; Jones, C. G. Mechanism of dye response and interference in the Bradford protein assay. *Anal. Biochem.* **1985**, *151*, 369−374.

(10) Sedmak, J. J.; Grossberg, S. E. A rapid, sensitive, and versatile assay for protein using Coomassie brilliant blue G250. *Anal. Biochem*. **1977**, *1*, 544−552.

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

(12) Atassi, M. Z.; Manshouri, T. Design of peptide enzymes (pepzymes): surface-simulation synthetic peptides that mimic the chymotrypsin and trypsin active sites exhibit the activity and specificity of the respective enzyme. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 8282−8286.

(13) Morr, C. V. Emulsifiers: milk proteins. In *Protein Functionality in Foods*; Cherry, J. P., Ed.; American Chemical Society: Washington, DC, 1981; pp 201−215.

(14) St. John Coghlan, D.; Woodrow, J.; Bamforth, C. W.; Hinchliffe, E. Polypeptides with enhanced foam potential. *J. Inst. Brew*. **1992**, *98*, 207−213.

(15) Lin, S.; Lakin, A. L. Thermal denaturation of soy proteins as related to their dye-binding characteristics and functionality. *J. Am. Oil Chem. Soc.* **1990**, *67*, 872−878.

(16) Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström, M.; Skagerberg, B.; Wold, S.; Andrews, P. Minimum analogue peptide sets (MAPS) for quantitative structure−activity relationships. *Int. J. Pept. Protein Res.* **1991**, *37*, 414−424.

(17) Beebe, K. R.; Pell, R. J.; Seascholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998.

(18) Crawford, O. H. A fast, stochastic threading algorithm for proteins. *Bioinformatics* **1999**, *15*, 66−71.

(19) Asano, K.; Hashimoto, N. Isolation and characterization of foaming proteins in beer. *J. Am. Soc. Brew. Chem*. **1980**, *38*, 129−137.

(20) Asano, K.; Shinagawa, K.; Hashimoto, N. Characterization of haze-forming proteins of beer and their roles in chill haze formation. *J. Am. Soc. Brew. Chem*. **1982**, *40*, 147−154.

(21) Siebert, K. J. Effects of protein−polyphenol interactions on beverage haze, stabilization, and analysis. *J. Agric. Food Chem*. **1999**, *47*, 353−362.

(22) Luck, G.; Hua, L.; Murray, N. J.; Grimmer, H. R.; Warminski, E. E.; Williamson, M. P.; Lilley, T. H.; Haslam, E. Polyphenols, astringency and proline-rich proteins. *Phytochemistry* **1994**, *37*, 357−371.

(23) Baxter, N. J.; Lilley, T. H.; Haslam, E.; Williamson, M. P. Multiple interactions between polyphenols and a salivary proline-rich protein repeat result in complexation and precipitation. *Biochemistry* **1997**, *36*, 5566−5577.

(24) Mehansho, H.; Butler, L. G.; Carlson, D. M. Dietary tannins and salivary proline-rich proteins: interactions, induction, and defense mechanisms. *Annu. Rev. Nutr.* **1987**, *7*, 423−440.

(25) Siebert, K. J.; Carrasco, A.; Lynn, P. Y. Formation of protein−polyphenol haze in beverages. *J. Agric. Food Chem*. **1996**, *44*, 1997−2005.

(26) Siebert, K. J.; Troukhanova, N. V.; Lynn, P. Y. Nature of polyphenol−protein interactions. *J. Agric. Food Chem.* **1996**, *44*, 80−85.

(27) Outtrup, H.; Fogh, R.; Schaumburg, K. The interaction between proanthocyanidins and peptides. *Proceedings, European Brewery Convention 21st Congress*, Madrid; IRL Press: Oxford, U.K., 1987; pp 583−590.

(28) Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem*. **1976**, *72*, 248−254.

(29) Lillford, P. J.; Holt, C. B. Antifreeze proteins. *J. Food Eng.* **1994**, *22*, 475−482.

(30) Siebert, K. J.; Knudson, E. J. The relationship of beer high molecular weight protein and foam. *Tech. Q. Master Brew. Assoc. Am*. **1989**, *26*, 139−146.

Received for review June 9, 2000. Revised manuscript received October 16, 2000. Accepted October 29, 2000. 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.

JF000718Y