research articles Journal of proteome research

From Precursor to Final Peptides: A Statistical Sequence-Based Approach to Predicting Prohormone Processing

Amanda B. Hummon,[†] Norman P. Hummon,[‡] Rebecca W. Corbin,[†,||] Lingjun Li,[†,⊥]
Ferdinand S. Vilim,[§] Klaudiusz R. Weiss,[§] and Jonathan V. Sweedler*,[†]

Department of Chemistry and the Beckman Institute, University of Illinois, Urbana, Illinois 61801, Department of Sociology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029

Received July 1, 2003

Predicting the final neuropeptide products from neuropeptides genes has been problematic because of the large number of enzymes responsible for their processing. The basic processing of 22 *Aplysia californica* prohormones representing 750 cleavage sites have been analyzed and statistically modeled using binary logistic regression analyses. Two models are presented that predict cleavage probabilities at basic residues based on prohormone sequence. The complex model has a correct classification rate of 97%, a sensitivity of 97%, and a specificity of 96% when tested on the *Aplysia* dataset.

Keywords: proteolytic processing • neuropeptides • MALDI MS • prediction algorithms

Introduction

Understanding brain function requires knowledge of the intercellular signaling molecules used by neurons. Neuropeptides are the most diverse and largest category of signaling molecules. The library of neuropeptide genetic information has exploded with current molecular biology techniques. However, the translation of genetic information into knowledge of biologically active neuropeptides is considerably more difficult. As of yet, expert systems cannot transform gene sequences into accurate predictions of final peptide products. Large prohormones are processed into smaller bioactive components under the action of large numbers of enzymatic processing steps. Many of the final products are incompletely characterized, and it is often difficult to determine the processing steps taken to produce these neuropeptides. Biochemical characterization of neuropeptides in the brain is a difficult task made easier by the advances in mass spectrometry (MS) techniques. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS) is a simple and elegant method to profile the peptide contents of a single cell[1−3] and can provide a snapshot of the peptide content of a single cell, from the intermediate prohormone products to the final bioactive peptides. As prohormones can be produced in a cell specific manner, the ability to examine cells individually is crucial for an understanding of the tissue-specific processing of neuropeptides.

The cleavage of bioactive peptides from a prohormone is controlled by a series of proteolytic enzymes. The most common cleavages occur at basic processing sites by endoproteases, typically at mono- and dibasic amino acid residues or at furin sites (R−X−K/R−R).[4] The proprotein convertases responsible for this activity are part of the subtilisin family of enzymes and are highly conserved across species.[5] Additional nonbasic processing sites have been observed.[6] In our laboratory, the complete processing of multiple prohormones of the marine mollusc, *Aplysia californica*, has been determined using MALDI MS.[7−9] The MS profile contains partially processed intermediates and the final posttranslationally modified peptide products. In this study, the processing of multiple prohormones was examined using previously published MS data, and the cleavage percentages of the common processing sites compiled. In addition, the amino acids surrounding and including the basic site were analyzed for trends. These residues were statistically examined using regression analyses and two models were generated. These models were tested first on the *Aplysia* dataset and then on data from other species to test their predictive power. This information aids in the prediction of peptides from gene sequences.

Materials and Methods

Data Tabulations. All *Aplysia* prohormones used have been published previously: proELH,[9,10] proR3−14 and proR15,[11] proL5−67,[12] proNPY,[13] proCerebrin,[14] proAMRP,[7] proInsulin,[8] proPRQ,[15] proAPGWs,[16] proPEP,[17] proCP2,[18] proCCK,[19] proFMRFa,[20] proMyomodulin,[21] proFRFa,[22] proFCMP,[23] proATRP,[24] proALK,[25] proEnterins,[26] and proSCP.[27] To ensure equivalent high quality processing data, only prohormone processing confirmed by mass spectrometry has been used. The data were compiled into a database with Microsoft Access.

Statistical Analyses. The binary logistic regression models were constructed using Minitab Statistical Software (Release 12, Minitab Inc.). To determine the significant amino acids, binary logistic regression analyses were performed for each position in the sequence being evaluated.

[†] Department of Chemistry and the Beckman Institute, University of Illinois.

[‡] Department of Sociology, University of Pittsburgh.

[§] Department of Physiology and Biophysics, Mount Sinai School of Medicine.

[||] Present Address: Department of Chemistry/Geology/Physics, Ashland University, Ashland, OH 44805.

[⊥] Present Address: Departments of Pharmaceutical Sciences and Chemistry, University of Wisconsin, Madison, WI 53705-2222.

10.1021/pr034046d CCC: $25.00

Published on Web 10/04/2003

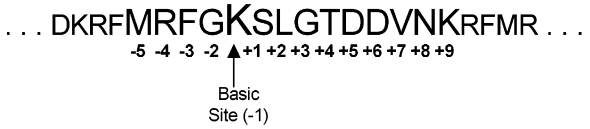

Figure 1. Segment from the *Aplysia* FMRFa prohormone sequence labeled to demonstrate the nomenclature for sequences.

The sequence nomenclature was adapted from previous studies.[28,29] The scissile bond was defined as the bond C-terminal to the furthest C-terminal basic amino acid. The sequence examined included the 9 amino acids N-terminal to the scissile bond (including the basic cleavage site), and the 9 amino acids C-terminal to the basic site. The basic site is defined as the − 1 position (dibasic sites occupy the − 2 and − 1 positions, tribasic, −3, −2, −1, etc.). The amino acid position immediately adjacent to a monobasic site on the N-terminal side was designated −2 (−3 for a dibasic site). The further N-terminal amino acids were labeled sequentially. A symmetrical nomenclature was used with the C-terminal amino acid positions: + 1 for the position adjacent to the basic site, + 9 for the furthest position. This nomenclature system is illustrated in Figure 1.

Results

Many elegant studies have described neuropeptide processing to develop rules of processing prohormones in vivo.[28,29] However, since the advent of methods that allow the measurement of the full processing of a prohormone into final peptide products, there have been few studies describing the processing of multiple prohormones to verify these processing observations. In addition, the current study involves a considerably larger data set than use in previous studies[28,29] to summarize neuronal prohormone processing.

In the cases of prohormones containing repeating sequences, each peptide produced from the prohormone was counted individually. For a site to be counted as processed, at least one of the two fragments with the exact mass (within one part in one thousand) that corresponds to that cleavage must be detected with mass spectrometry. An unprocessed site was determined by detection of a defined mass corresponding to a peptide fragment with the site uncleaved. Some sites are partially processed, that is, both processed and nonprocessed forms are detected. Although such incomplete processing may be real (a site is only partially processed), a likely explanation is the nature of MS data. The MS approach obtains a snapshot of the peptides at a particular time in a cell at the instant the cell is acidified with the MALDI matrix. Because of this, partially processed sites are counted as processed.

The basic residues considered are Lys and Arg. For these tables, the monobasic sites are single Lys or Arg residues, with two or three adjacent basic residues comprising the dibasic and tribasic sites. Not surprisingly, the most frequently used cleavage sites are dibasic residues, with 100% of the KR residues cleaved and only 1 of the 2 RK sites cleaved. Over 375 dibasic sites were examined from 21 prohormones and the results are tabulated in Table 1. Although dibasic residues are the most common cleavage locations, monobasic sites are used as well. The usage percentages for mono- and tribasic residue cleavage as well as furin cleavages in *Aplysia* prohormones are also presented (Table 2). Only sites that conform to the R−X−K/R−R cleavage consensus site[30] were counted as furin sites.

To determine the amino acids that significantly affect the probability of cleavage, a binary logistic regression analysis was performed for each sequence position in the −9 to +9 positions (Analysis 1). This analysis contains too many variables to be useful. Thus, the complexity of the model was reduced using multiple iterations and selecting the most significant amino acids in a stepwise fashion at each position for predicting cleavage. Specifically, the amino acids with *p* values of 0.000 in Analysis 1 from the positions −9 through +9 were analyzed together in a regression analysis (Analysis 2). The majority of significant amino acids were located in the −5 through +5 positions, so additional analyses focused primarily on that

Table 1. Processing Percentages at Dibasic Residues in *Aplysia* Prohormones

	RR		RK		KR		KK	
prohormone	used	not used	used	not used	used	not used	used	not used
proELH	1	1			2			
proR3−14	1				1			
proR15		1		1	2		1	
proL5−67					1			
proNPY					1			
proCerebrin					2			
proAMRP	4				11		4	
proInsulin		1			2		1	
proPRQ					66			
proAPGWs		1	1		12			
proPEP					23			
proCP2					1			
proCCK		2			23		2	
proFMRFa	1				31			
proMyomodulin					17		2	
proATRP					2			
proALK					81		2	
proEnterins	1				39		1	
proFCAP					37		2	
proOMMs	1				3			
proFRFa					8			
totals	9	6	1	1	365	0	15	0
	60%		50%		100%		100%	

Table 2. Processing Percentages at Monobasic and Tribasic Residues and Furin-Like Sites in *Aplysia* Prohormones

	monobasic sites				tribasic sites		furin sites	
	R		K		KKR or RKR		R−X−K/R−R	
prohormone	used	not used	used	not used	l	not used	used	not used
proELH	3	11	2	2	1		1	
proR3−14	1	7		1				
proR15		5		1				
proL5−67		3						
proNPY		6				1		
proCerebrin		2		2				1
proAMRP	16	21		1	11			
proInsulin	2	6	1	1				
proPRQ		37						
proAPGWs				2			1	
proCP2	2	5	1	2				
proSCP	2	3						
proCCK	2	38		7				
proFMRFa	5	31	21	3			2	
proMyomodulin	1	17					1	
proATRP		2						
proALK		5		5	4			
proEnterins	3		1	6				
proFCAP	1	1			3		1	
proOMMs	8	14	1					1
proFRFa		6						
totals	46	220	27	33	19	1	6	2
	17%		45%		95%		75%	

Table 3. Comparison of Dibasic Cleavage Percentages for *Aplysia*, Mammal, and Insect Sequences

	Aplysia		mammals[a]		insects[b]	
	pairs analyzed	percent of pairs cleaved	pairs analyzed	percent of pairs cleaved	pairs analyzed	percent of pairs cleaved
KR	365	**100**	27	**100**	39	**95**
RR	15	**60**	13	**62**	18	**33**
RK	2	**50**	8	**25**	9	**0**
KK	15	**100**	8	**38**	5	**20**

refs: [a]Schwartz, 1983; [b]Veenstra, 2000.

portion of the sequences. Considering this portion, all of the amino acids with p values < 0.05 from Analysis 1 were analyzed in a regression and the amino acids with p values < 0.300 were deemed significant (Analysis 3). From these analyses, a subset of the most significant amino acids at predicting cleavage was constructed.

This subset is composed of the significant amino acids from Analysis 3 (−5 through +5). In addition, Lys in the +9 position, the only amino acid from a position further removed from the cleavage site remaining in the simplified model (−9 to −6 and +6 to +9) to have a p value < 0.100 in Analysis 2, remained in this model. A regression analysis was performed with this subset (Analysis 4). This analysis, Analysis 4, was used as the basis in constructing the final models. Analysis 4 provided p values for every amino acid in the subset relative to one another. Using varying cutoff p values, portions of the subset could be selected and analyzed together to construct the final models. Model 1 was constructed with a cutoff p value < 0.500 from Analysis 4 and Model 2 used a cutoff p value < 0.050.

A summary of Model 1 is presented in Table 4 while the results of Model 2 are detailed in Table 5. Additional measures of accuracy[31] for the two models are provided in Table 6. The models were also tested on published sequences (Supporting Table 1) and the measures of accuracy are also provided in Table 6. Finally, the processing of mosquito prohormones is predicted with Models 1 and 2 and is compared with recently published predictions (Supporting Tables 2 and 3).

Discussion

One of the most important family of enzymes responsible for prohormone processing are the proprotein convertases (PC), a group of subtilisin-like endoproteases, that cleave at the carboxyl terminus of basic residues.[4] Following cleavage by the protease, carboxypeptidases remove the basic residues from the C-terminus of the nascent neuropeptide.[32] The proteases are highly conserved across species and primarily are present in the secretory vesicles.[33] Some of the PCs have complementary activity, for example, PC1 and PC2 both act in the processing of proinsulin, though cleaving the molecule at different sites.[4] One of the PCs, furin, cleaves predominantly at the consensus sequence R−X−K/R−R,[30] although it has been demonstrated through site-directed mutagenesis that the endoprotease will cleave at other basic combinations as well, though less efficiently.[34] In *Aplysia*, four PCs have been cloned, aPC1A,[33] aPC1B,[33,35] aPC2,[33,35,36] and aFurin.[33,37] Also, two *Aplysia* carboxypeptidases, carboxypeptidase D[38] and E,[39] have been characterized, indicating that the expected suite of processing enzymes are found in *Aplysia*.

A protein molecule can be considered from multiple perspectives; for example, each protein has unique primary, secondary, and tertiary structures. When examining complex molecular interactions such as enzyme recognition resulting in protein cleavage, each of these different types of structural information may be evaluated to assist in predicting this processing event. Certainly, explanations that encompass multiple structural elements are valuable; however, in many cases, valid information can be gleaned from primary structure alone.

Table 4. Quantitative Values for Basic Cleavage from Model 1[a]

factors that assist cleavage

position	amino acid	coef.	*Z* value	*P* value	position	amino acid	coef.	*Z* value	*P* value
−5	H	1.178	0.76	0.446	+1	Q	1.341	1.57	0.117
−5	P	1.037	1.01	0.313	+1	S	2.454	3.08	0.002
−4	F	2.089	1.86	0.062	+3	P	2.508	3.68	0.000
−4	L	1.932	1.83	0.067	+5	V	3.176	2.12	0.034
−4	V	4.207	3.12	0.002	+9	K	0.897	1.08	0.280
−3	L	1.435	1.68	0.092					
−2	G	3.163	4.20	0.000					
−2	K	8.643	6.44	0.000					
−1	R	0.884	1.39	0.166					

factors that hinder cleavage

position	amino acid	coef.	*Z* value	*P* value	position	amino acid	coef.	*Z* value	*P* value
−3	R	−2.888	−1.23	0.217	+1	F	−2.787	−3.15	0.002
−2	M	−1.455	−0.96	0.335	+1	L	−3.370	−2.64	0.008
					+1	N	−4.770	−2.83	0.005
					+1	Y	−1.617	−1.09	0.276
					+2	F	−2.372	−1.73	0.084
					+4	F	−0.807	−0.82	0.411
					+4	M	−1.670	−1.01	0.312
					+5	K	−2.279	−1.45	0.146

[a] When tested on the *Aplysia* dataset, 472 basic sites were predicted to be cleaved with Model 1 that were observed to be cleaved with mass spectrometry, 10 sites were predicted to be cleaved that were not observed with MS, 16 sites were not predicted to be cleaved that were observed to be cleaved by MS, and 252 sites were not predicted to be cleaved that were not observed to be cleaved by MS. The model constant value is −2.8334, the *Z* value is −4.18, the p value is 0.000, the percentage of concordant pairs is 99.3%, the percentage of discordant pairs is 0.5%, and the percentage of ties is 0.1%.

Table 5. Quantitative Values for Basic Cleavage from Model 2[a]

factors that assist cleavage									
position	amino acid	coef.	*Z* value	*P* value	position	amino acid	coef.	*Z* value	*P* value
−4	V	3.179	3.16	0.002	+1	S	2.362	3.46	0.001
−2	G	2.986	4.77	0.000	+3	P	3.493	6.35	0.000
−2	K	8.551	7.33	0.000	+5	V	2.528	2.35	0.019
factors that hinder cleavage									
					+1	F	−2.869	−3.72	0.000
					+1	L	−2.340	−2.16	0.031
					+1	N	−3.161	−1.80	0.073

[a] When tested on the *Aplysia* dataset, 472 basic sites were predicted to be cleaved with Model 1 that were observed to be cleaved with mass spectrometry, 13 sites were predicted to be cleaved that were not observed with MS, 16 sites were not predicted to be cleaved that were observed to be cleaved by MS, and 249 sites were not predicted to be cleaved that were not observed to be cleaved by MS. The model constant values is −2.1901, the Z value is −7.71, the p value is 0.000, the percentage of concordant pairs is 98.0%, the percentage of discordant pairs is 0.6%, and the percentage of ties is 1.4%.

Table 6. Assessment of Prediction Errors for Models 1 and 2 with *Aplysia* Prohormones and Sequences from Other Species[a]

measure	calculation	simple model *Aplysia* dataset	complex model *Aplysia* dataset	complex model other species[b]
prevalence	$(a + c)/N$	65.1%	65.1%	56.6%
overall diagnostic power	$(b + d)/N$	35.9%	34.9%	43.4%
correct classification rate	$(a + d)/N$	96.1%	96.5%	89.8%
sensitivity	$a/(a + c)$	96.7%	96.7%	87.2%
specificity	$d/(b + d)$	95.0%	96.2%	93.1%
false positive rate	$b/(b + d)$	4.9%	3.8%	6.9%
false negative rage	$c/(a + c)$	3.3%	3.3%	12.8%
positive predictive power	$a/(a + b)$	97.3%	97.9%	94.3%
negative predictive power	$d/(c + d)$	95.0%	94.0%	84.8%
misclassification rate	$(b + c)/N$	3.9%	3.5%	10.2%
odds-ratio	$(a*d)/(c*b)$	565.0	743.4	91.6

[a] The confusion matrix terms are defined as: **a**: basic sites that were predicted to be cleaved with Model 1 that were also observed to be cleaved with mass spectrometry; **b**: sites that were predicted to be cleaved by Model 1 that were not observed to be cleaved with MS; **c**: sites that were not predicted to be cleaved but were observed to be cleaved by MS; **d**: sites that were not predicted to be cleaved that were also not observed to be cleaved by MS. [b] Other Species: Includes Cockroach allatostatin (*Periplaneta americanus*), Loligo FMRFa (*Loligo opalescens*), Snail CDCH (*Lymnaea stagnalis*), and several Fruitfly prohormones (*Drosophila melanogaster*) (166 total sites).

As shown in Table 1, the most striking result is the utilization percentage of Lys-Arg (KR) and Lys-Lys (KK) sites; all of the 365 KR and 15 KK sites surveyed were cleaved, making these residues reliable predictors of cleavages from new prohormones. Tribasic sites are also frequently cleaved (95%, Table 2), although they occur less commonly in prohormone sequences. Thus, KR and KK sites are an excellent example of the predictive power of primary structure as these specific combinations of amino acids are sufficient for endoprotease action, regardless of the local secondary and tertiary structure.

Another interesting property of these cleavage patterns is their conservation in diverse metazoa.[28,40] The utilization percentages for common cleavage sites have also been compiled for insects and mammals (Table 3)[28,41] and show noticeable consistency between different animals. In a study by Paganetti and Scheller, the *Aplysia* atrial gland ELH-related precursor (A-ELH) was transfected in anterior pituitary tumor (AtT-20) cells. In the mammalian cells, processing of the precursor occurred with cleavages produced at the same monobasic and dibasic amino acids sites that are natively utilized in *Aplysia*.[5] The largest difference between phyla are the less commonly employed basic residues such as KK, which may be because of the poor statistics from uncommon processing sites when using data sets of limited size.

As shown in Table 2, although common in prohormones, the monobasic residues are less frequently used cleavage sites. In this case, the primary structure does not appear sufficient to accurately predict cleavage, perhaps due to more elaborate secondary structure near such sites or the properties of the enzymes. As a consequence, previous studies have examined the sequences surrounding monobasic sites.[29,42,43] These studies produced systems of rules that are fairly accurate in the prediction of neuropeptides. These rules have been compiled by observing the frequencies of individual amino acids occurring in specific positions, tabulating their frequencies, and using those frequencies to infer the significance of the amino acid for recognition and cleavage by an endopeptidase. To our knowledge, the current study represents the first quantitative statistical analysis on this type of dataset. Our statistically based results agree with previous studies, but provide additional quantitative information, especially useful when confounding amino acids are present near a basic site (those that predict cleavage and those that hinder cleavage).

One of the primary observations in these prior studies is a higher cleavage percentage at single arginine residues than at single lysine residues. We observe the opposite trend; in the *Aplysia* prohormones, single lysine residues are cleaved 45% of the time while single arginine residues are used only 17% of the time. However, the higher percentage for lysine is partially due to the large number of single lysine sites cleaved in the proFMRFa precursor (21 of the 27 cleaved lysine sites are from the FMRFa prohormone). Single arginines are more common, but lysines are a more likely site for cleavage.

To determine more complex sequence trends concerning basic cleavage, statistical analyses were performed. Regression analyses investigate and model the relationship between a response variable and its predictor variables. In binary logistic regression analyses, the data is dichotomized into a binary

$$g(\pi_j)=\beta_o+x_j'\beta_j+x_k'\beta_k+\ldots=\mathrm{Log}_e(\pi_j/(1-\pi_j))$$

$g(\pi_j)$ = the Estimated Linear Function of the Model
Π_j = Probability of a response
B_o = the Intercept (the Model Constant Value)
x_j' = A vector of predictor variables associated with the jth factor/covariate pattern
(The number of times a predictor appears in a Consensus Sequence, 0 or 1)
B_j = A vector of unknown coefficients associated with the predictors
(Coefficient for a predictor)

Figure 2. Logit functions that are used to construct binary logistic regression models. The quantitative values provided in Tables 4 and 5 can be used to compute the probability values for any combination of predictors described by the models.

function. In this case, the surrounding amino acids, the predictor variables, are used to evaluate the outcome of a binary response variable, cleaved/not cleaved.

For each of the predictor variables, a coefficient (coef.), a *z* value, and a *p* value is provided. The coefficient represents the marginal contribution to the logit function for a specific variable. The distribution of the data is “linked” to the model by means a logit link function. The logit function expresses the probability of cleavage (see Figure 2). One unit of change in a coefficient results in one unit of change in the value of the logit function. The *z* value is the test statistic for the coefficient of each predictor variable. A larger *z* value indicates a more significant variable. Both the coefficients and the *z* values can be expressed as positive and negative values. In these models, positive values reflect increased likelihood of cleavage; negative values indicate less likelihood. The *p* value is associated with the test statistic and is the probability that the coefficient is statistically different from zero.

The accuracy of the models is also provided. Once the data is linked to the logit function, the model is used to predict the response variable for each data point in the given data set. A probability value for each set of predictor variables is calculated and compared against a threshold probability value. The threshold probability value is the second derivative of the logit function. For a binary logistic regression, the threshold probability value is 0.50 or 50%. The percentage of concordant pairs indicates the percentage of trials when the model accurately predicted the binary response variable, given the predictor variables. The discordant pairs indicate the percentage of inaccurate predictions. When the value of the binary response variable matches the threshold probability value, it is recorded as a tie. For these models, calculated values for the logit function above 0.50 indicate a cleaved prediction and values below 0.50 predict that a specific sequence will not be cleaved.

Two models are presented in this report. Model 1 (Table 4) is more complex, containing more predictor variables, and has a higher degree of accuracy. Model 2 (Table 5) is considerably simpler, less accurate, but contains only statistically relevant predictor variables in a few key positions. Accuracy measurements for the two models[31] are provided in Table 6.

Model 1 has a 96.5% correct classification rate with the *Aplysia* dataset. The model indicates the significance of 24 different predictor variables occupying 11 of the different sites in the sequence. Some trends that are apparent in the model: cleavage is assisted by the presence of bulky, hydrophobic residues in the −4 position but hindered by their presence C-terminus to the basic site. Lys in the −2 position is a powerful predictor of cleavage, not surprising considering the 100% cleavage percentages observed for both KR and KK sites. Small amino acids adjacent to the basic site promote cleavage, whereas large residues are detrimental to cleavage. Many bioactive neuropeptides have amidated C-termini, the result of a glycine residue that has been converted by peptidylglycine alpha-hydroxylating monooxygenase to form the amide.[44,45] Because of this important biological function, glycines commonly precede cleavage sites. A large percentage of *Aplysia* prohormone cleavage sites are preceded by glycine residues; structurally, this may be because Gly allows a wider range of unusual main chain conformations.[46] When glycine precedes a basic site, the likelihood of cleavage is significantly increased. Proline aids cleavage when it is present in positions a few amino acids from the basic site: −5 and +3 positions. Perhaps its unique structure contributes a bend to the peptide configuration, permitting easier access to the scissile bond. Other trends can also be inferred, but are not as striking.

Similar trends can be inferred from Model 2. This model is considerably simpler and indicates that with only 9 amino acid predictors in 5 positions, a 96.1% correct classification rate can be obtained. In this model, all but one of the predictor variables are statistically significant and all but three aid cleavage. Although Model 1 contains numerous amino acids factors and high accuracy, this model was designed to be a blend of simplicity and accuracy. Model 2 contains only a few powerful predictors for quick prediction.

The predictions of Model 2 are similar to those from Model 1 and make sense from a chemical standpoint. Small amino acids, glycine and serine, adjacent to the basic site allow the endoprotease access and promote cleavage. Bulky hydrophobic residues, like phenylalanine, leucine, and valine, aid cleavage when positioned far from the basic site (−4 and +5 positions) but are detrimental when found closer (+1 position). In fact, the +1 position contains four of the nine strong amino acid predictors, indicating the importance of this site for enzyme recognition. As with Model 1, Lys is an extremely strong predictor of cleavage in the −2 position. The variables in these models are linear, that is, they predict cleavage independent of the presence of other amino acids. Using multiple variables simultaneously was not necessary to achieve accurate predictions and would complicate the analysis.

To test the statistical models on a different dataset, the processing of published prohormones from other species were examined. The processing of cockroach allatostatin (*Periplaneta americanus*[47,48]), squid FMRFa (*Loligo opalescens*[49]), snail CDCH (*Lymnaea stagnalis*[50]), and multiple fruit fly prohormones (*Drosophila melanogaster*[51]) were previously determined using MALDI and ESI−MS and so the data quality is similar to those used in constructing the model. The sequences surrounding and including each of the basic sites in the prohormones were examined and Model 1 was used to make cleavage predictions. These predictions were then compared against the published results (Supporting Table 1); out of 166 basic sites, 149 were predicted correctly with Model 1, giving a correct classification rate of 89.8%. Additional measures of accuracy are described in Table 6. These results imply that the models described here have relevance to organisms besides *Aplysia* in a wide variety of metazoa with only a slightly poorer predictive ability.

For the allatostatin prohormone, the model and the MS data agreed in all cases. As a general rule, arginine was not found to be a strong predictor of cleavage when found in the −2 position, perhaps because RR and RK sites had 60% and 50% cleavage percentages in the *Aplysia* dataset, respectively, as compared to the 100% cleavage percentages when a lysine is found in the −2 position (Table 1). There were six other false

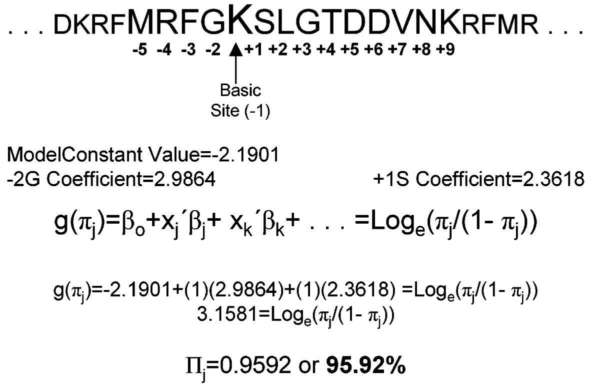
. . . DKRFMRFGKSLGTDDVNKRFMR . . .
-5 -4 -3 -2 +1 +2 +3 +4 +5 +6 +7 +8 +9
Basic Site (-1)
ModelConstant Value=-2.1901
-2G Coefficient=2.9864
+1S Coefficient=2.3618
g(πj)=βo+xj´βj+ xk´βk+ . . . =Loge(πj/(1- πj))
g(πj)=-2.1901+(1)(2.9864)+(1)(2.3618) =Loge(πj/(1- πj))
3.1581=Loge(πj/(1- πj))
Πj=0.9592 or 95.92%

Figure 3. Example calculation to predict the probability of cleavage for a segment of the *Aplysia* FMRFa prohormone using the logit function for Model 2. The segment is predicted to cleave based on the calculation, which predicts a probability of 95.92%.

negatives corresponding to sequences with arginine occupying the −2 position; three tribasic KRR sites, one from proCDCH and two from the *Drosophila* prohormones and 3 RR sites, all from the *Drosophila* prohormones. Another source of multiple false negative results was a repeating sequence from the FMRFa prohormone. In the prohormone, a sequence with glycine in the −2 position, a single basic residue, and either asparagine or aspartate in the +1 position (G−K/R−N/D) repeated five times. These sites were cleaved according to the MALDI data; however, Model 1 did not predict these events, due in large measure to the presence of Asn/Asp in the +1 position, which is a strong predictive factor for hindering cleavage in the *Aplysia* prohormones.

There were only five cases of false positives, where Model 1 predicted cleavage and it was not observed in the MALDI or ESI data. One case involved a sequence where a dibasic KK was not observed to be cleaved when examined by ESI−MS. As lysine is a strong predictor of cleavage in the Aplysia dataset, this site was predicted to be cleaved. In general, Model 1 tends to predict on the conservative side, rarely suggesting cleavage where it does not occur.

The rules compiled by previous studies have been successful at predicting prohormone processing into peptides.[28,52] In general, our rules agree with previous prediction rules. Devi noted that aliphatic amino acids in the −2 position and bulky aromatic residues in the +1 position are detrimental to cleavage. In addition, small amino acids were found to assist cleavage when present in the +1 position. Our results agree with these rules, although our models include methods to take into account the presence of both positive and negative factors.

One of the main advantages of these statistical analyses is that the results are quantitative. As described earlier, the predictions of binary logistic regression models are based on comparisons of probability values for a specific set of predictors against a threshold value. Using the logit function (Figure 2), the Model's Constant Value, and the Coefficients (coef.) for each of the individual predictors, a probability can be calculated for every combination of predictors. This value can be compared against the threshold probability value, 0.50 or 50%, to determine the likelihood of cleavage for any combination of predictors. This enables new prohormones to be tested using either model to determine the most likely basic site cleavages. A sample calculation is shown in Figure 3.

With the completion of the mosquito genome in October of 2002, predictions for putative neuropeptides were made,[52] although no experimental data exists at this time to determine the accuracy of their predictions. To compare our statistical models against these predictions, some of the same sequences were analyzed, representing 221 basic sites (Supporting Tables 2 and 3). Riehle's predictions were more similar to those predictions made with the simpler model, Model 2. Using Model 2, 186 predictions matched those of Riehle and coworkers while with the more complex model, Model 1, 183 predictions were identical. In both cases, the predictions made with the statistical models presented here appear to be more conservative than those made by Riehle: there were more cases of basic sites predicted to be cleavage sites than with our models. Some of the differences observed with the predictions may be due to our ability to evaluate between conflicting factors in a sequence. At this time, actual information on the peptides in mosquitoes is lacking, but as this information becomes available, it will be interesting to determine which of the disparate predictions are correct.

Multiple trends that affect basic site cleavage patterns are presented in this report. Differential cleavage sites provide an elegant mechanism by which multiple bioactive products can be derived from a single source.[28] While this pluralism allows for rich diversity, it complicates the process of making predictions from novel sequences. Only through detailed observations and techniques such as statistical modeling can we accurately describe prohormone processing. Perhaps most surprising is that such cleavage trends can be modeled with high predictive success without examination of additional factors such as secondary structure and the large number of enzymes that are responsible for the prohormone processing.

As the endoproteases responsible for the cleavage patterns are highly conserved across phyla, predictions of this nature are applicable to a wide range of organisms. This information will also aid in understanding enzyme function and, by determining the organisms and tissues for which these models do not work well, may point to cells that contain unknown processing enzymes. With the rapidly expanding array of completed genomes, accurate algorithms capable of predicting final bioactive products are a key step in the advancement of proteomics.

Acknowledgment. This study was supported by the generous funding of the National Institutes of Health (NIH) Grants NS31609 and DA14879 to J.V.S., MH50235, MH01427, and DA133330 to F.S.V. and K.R.W.

Supporting Information Available: Models tested on published sequences and processing of mosquito prohormones predicted with Models 1 and 2 compared with recently published predictions (three tables). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Li, L.; Garden, R. W.; Sweedler, J. V. *Trends Biotechnol.* **2000**, *18*, 151−160.
(2) van Veelan, P. A.; Jimenez, C. R.; Li, K. W.; Wildering, W. C.; Gerearts, W. P.; Tjaden, U. R.; van der Greef, J. *Organic Mass Spectrom.* **1993**, *28*, 1542−1546.
(3) Jimenez, C. R.; Li, K. W.; Dreisewerd, K.; Spijker, S.; Kingston, R.; Bateman, R. H.; Burlingame, A. L.; Smit, A. B.; van Minnen, J.; Gerearts, W. P. *Biochemistry* **1998**, *37*, 2070−2076.
(4) Canaff, L.; Bennett, H. P. J.; Hendy, G. N. *Mol. Cell. Biol.* **1999**, *156*, 1−6.
(5) Paganetti, P.; Scheller, R. H. *Brain Res.* **1994**, *633*, 53−62.
(6) Hummon, A. B.; Huang, H.; Kelley, W. P.; Sweedler, J. V. *J. Neurochem.* **2002**, *82*, 1398−1405.

(7) Fujisawa, Y.; Furukawa, Y.; Ohta, S.; Ellis, T. A.; Dembrow, N. C.; Li, L.; Floyd, P. D.; Sweedler, J. V.; Minakata, H.; Nakamaru, K.; Morishita, F.; Matsushima, O.; Weiss, K. R.; Vilim, F. S. *J. Neurosci.* **1999**, *19*, 9618−9634.

(8) Floyd, P. D.; Li, L.; Rubakhin, S. S.; Sweedler, J. V.; Horn, C. C.; Kupfermann, I.; Alexeeva, V.; Ellis, T. A.; Dembrow, N. C.; Weiss, K. R.; et al. *J. Neurosci.* **1999**, *19*, 7732−7741.

(9) Garden, R. W.; Shippy, S. A.; Li, L.; Moroz, T. P.; Sweedler, J. V. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 3972−3977.

(10) Nagle, G. T.; Painter, S. D.; Blankenship, J. E.; Kurosky, A. *J. Biol. Chem.* **1988**, *263*, 9223−9237.

(11) Garden, R. W.; Moroz, T. P.; Gleeson, J. M.; Floyd, P. D.; Li, L.; Rubakhin, S. S.; Sweedler, J. V. *J. Neurochem.* **1999**, *72*, 676−681.

(12) Li, L.; Moroz, T. P.; Garden, R. W.; Floyd, P. D.; Weiss, K. R.; Sweedler, J. V. *Peptides* **1998**, *19*, 1425−1433.

(13) Rajpara, S. M.; Garcia, P. D.; Roberts, R.; Eliassen, J. C.; Owens, D. F.; Maltby, D.; Myers, R. M.; Mayeri, E. *Neuron* **1992**, *9*, 505−513.

(14) Li, L.; Romanova, E. V.; Floyd, P. D.; Rubakhin, S. S.; Sweedler, J. V.; Alexeeva, V.; Ellis, T. A.; Dembrow, N. C.; Weiss, K. R.; Vilim, F. S. *J. Neurochem.* **2001**, *77*, 1569−1580.

(15) Furukawa, Y.; Fujisawa, Y.; Minakata, J.; Nakamaru, K.; Morishita, F.; Matsushima, O.; Li, L.; Sweedler, J. V.; Ellis, T. A.; Dembrow, N. C.; Weiss, K. R.; Vilim, F. S. *Society for Neuroscience, 30th Annual Meeting*, 2000.

(16) Fan, X.; Croll, R. P.; Wu, B.; Fang, L.; Shen, Q.; Painter, S. D.; Nagle, G. T. *J. Comput. Neurol.* **1997**, *387*, 53−62.

(17) Alexeeva, V.; Jing, J.; Morris, L. G.; Hurwitz, I.; Hummon, A. B.; Sweedler, J. V.; Weiss, K. R.; Vilim, F. S. *Society for Neuroscience, 31st Annual Meeting*, 2001.

(18) Vilim, F. S.; Alexeeva, V.; Li, L.; Moroz, T. P.; Sweedler, J. V.; Weiss, K. R. *Peptides* **2001**, *22*, 2027−2038.

(19) Vilim, F. S.; Jing, J.; Alexeeva, V.; Church, P. J.; Hummon, A. B.; Sweedler, J. V.; Weiss, K. R. *Society for Neuroscience, 31st Annual Meeting*, 2001.

(20) Schaefer, M.; Picciotto, M. R.; Kreiner, T.; Kaldany, R. R.; Taussig, R.; Scheller, R. H. *Cell* **1985**, *41*, 457−467.

(21) Lopez, V.; Wickham, L.; DesGroseillers, L. *DNA Cell Biol.* **1993**, *12*, 51−59.

(22) Rybak, J.; Alexeeva, V.; Brezina, V.; Cropper, E. C.; Kupfermann, I.; Orekhova, I.; Price, D. A.; Vilim, F. S.; Weiss, K. R. *Society for Neuroscience, 26st Annual Meeting*, 1996.

(23) Sweedler, J. V.; Li, L.; Rubakhin, S. S.; Alexeeva, V.; Dembrow, N. C.; Dowling, O.; Jing, J.; Weiss, K. R.; Vilim, F. S. *J. Neurosci.* **2002**, *22*, 7797−7808.

(24) Vilim, F. S.; Park, J. H.; Dembrow, N. C.; Alexeeva, V.; Jing, J.; Weiss, K. R. *Society for Neuroscience, 30th Annual Meeting*, 2000.

(25) Weiss, K. R.; Saunders, S.; Cropper, E. C.; Alexeeva, V.; Jing, J.; Church, P. J.; Vilim, F. S. *Society for Neuroscience; 30th Annual Meeting*, 2000.

(26) Furukawa, Y.; Nakamaru, K.; Wakayama, H.; Fujisawa, Y.; Minakata, J.; Ohta, S.; Morishita, F.; Matsushima, O.; Li, L.; Romanova, E. V.; Sweedler, J. V.; Park, J. H.; Romero, A.; Cropper, E. C.; Dembrow, N. C.; Jing, J.; Weiss, K. R.; Vilium, F. S. *J. Neurosci.* **2001**, *21*, 8247−8361.

(27) Mahon, A. C.; Lloyd, P. E.; Weiss, K. R.; Kupfermann, I.; Scheller, R. H. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 3925−3929.

(28) Veenstra, J. *Arch, Insect Biochem. Physiol.* **2000**, *43*, 49−63.

(29) Devi, L. *FEBS Lett.* **1991**, *280*, 189−194.

(30) Roebroek, A. J. M.; Creemers, J. W. M.; Ayoubi, T. A. Y.; Van de Ven, W. J. M. *Biochemie* **1994**, *76*, 210−216.

(31) Fielding, A. H.; Bell, J. F. *Environ. Conserv.* **1997**, *24*, 38−49.

(32) Seidah, N. G.; Chretien, M. *Trends Endocrinol. Metab.* **1992**, *3*, 133−140.

(33) Chun, J. Y.; Korner, J.; Kreiner, T.; Scheller, R. H.; Axel, R. *Neuron* **1994**, *12*, 831−844.

(34) Rehemtulla, A.; Kaufman, R. J. *Blood* **1992**, *79*, 2349−2355.

(35) Ouimet, T.; Castellucci, V. F. *J. Neurochem.* **1997**, *68*, 1031−1040.

(36) Nagle, G. T.; Garcia, A. T.; Knock, S. L.; Gorham, E. L.; Van Heumen, W. R.; Kurosky, A. *DNA Cell Biol.* **1995**, *14*, 145−154.

(37) Nagle, G. T.; Garcia, A. T.; Gorham, E. L.; Knock, S. L.; Van Heumen, W. R.; Spijker, S.; Smit, A. B.; Geraerts, W. P.; Kurosky, A. *DNA Cell Biol.* **1995**, *14*, 431−443.

(38) Fan, X.; Qian, Y.; Fricker, L. D.; Akalal, D. B.; Nagle, G. T. *DNA Cell Biol.* **1999**, *18*, 121−132.

(39) Juvvadi, S.; Fan, X.; Nagle, G. T.; Fricker, L. D. *FEBS Lett.* **1997**, *408*, 195−200.

(40) Rholam, M.; Brakch, N.; Germain, D.; Thomas, D. Y.; Fahy, C.; Boussetta, H.; Boileau, G.; Cohen, P. *Eur. J. Biochem.* **1995**, *227*, 707−714.

(41) Schwartz, T. W.; Wittels, B.; Tager, H. S. In *Proceedings of the Eighth American Peptide Symposium*; Darby, V. J., Ed.; Pierce Chemical: Rockford, IL, Univ of Arizona, 1983; pp 229−238.

(42) Schwartz, T. W. *FEBS Lett.* **1986**, *200*, 1−10.

(43) Gomez, S.; Boileau, G.; Zollinger, L.; Nault, C.; Rholam, M.; Cohen, P. *EMBO* **1989**, *8*, 2911−2916.

(44) Merkler, D. J. *Enzyme Microb. Technol.* **1994**, *16*, 450−456.

(45) Bradbury, A. F.; Smyth, D. G. *Physiol. Bohemoslov* **1988**, *37*, 267−274.

(46) Brandon, C.; Tooze, J. *Introduction to Protein Structure*, 2nd ed.; Garland: New York, 1999.

(47) Predel, R. *J. Comput. Neurol.* **2001**, *436*, 363−375.

(48) Predel, R.; Kellner, R.; Rapus, J.; Gade, G. *Reg. Peptides* **1999**, *82*, 81−89.

(49) Sweedler, J. V.; Li, L.; Floyd, P. D.; Gilly, W. *J. Exp. Biol.* **2000**, *203*, 3565−3573.

(50) Jimenez, C. R.; van Veelan, P. A.; Li, K. W.; Wildering, W. C.; Gerearts, W. P.; Tjaden, U. R.; van der Greef, J. *J. Neurochem.* **1994**, *62*, 404−407.

(51) Baggerman, G.; Cerstiaens, A.; De Loof, A.; Schoofs, L. *J. Biol. Chem.* **2002**, *277*, 40 368−40 374.

(52) Riehle, M. A.; Garczynski, S. F.; Crim, J. W.; Hill, C. A.; Brown, M. R. *Science* **2002**, *298*, 172−175.

PR034046D