Bridging Neuropeptidomics and Genomics with Bioinformatics: Prediction of Mammalian Neuropeptide Prohormone Processing

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JUNE 2006

Impact Factor: 4.25 · DOI: 10.1021/pr0504541 · Source: PubMed

6 AUTHORS, INCLUDING:

Tyler A Zimmerman
National Institute of Standards and Technolo…
16 PUBLICATIONS **322** CITATIONS
SEE PROFILE

Jonathan Sweedler
University of Illinois, Urbana-Champaign
511 PUBLICATIONS **15,111** CITATIONS
SEE PROFILE

Published in final edited form as:
J Proteome Res. 2006 May ; 5(5): 1162–1167. doi:10.1021/pr0504541.

Bridging Neuropeptidomics and Genomics with Bioinformatics: Prediction of Mammalian Neuropeptide Prohormone Processing

Andinet Amare[1], **Amanda B. Hummon**[1], **Bruce Southey**[1,2], **Tyler A. Zimmerman**[1], **Sandra L. Rodriguez-Zas**[2], and **Jonathan V. Sweedler**[1,*]

1 *Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801*

2 *Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801*

Abstract

Neuropeptides are an important class of cell to cell signaling molecules that are difficult to predict from genetic information because of their large number of posttranslational modifications. The transition from prohormone genetic sequence information to the determination of the biologically active neuropeptides requires the identification of the cleaved basic sites, among the many possible cleavage sites, that exist in the prohormone. We report a binary logistic regression model trained on mammalian prohormones that is more sensitive than existing methods in predicting these processing sites, and demonstrate the application of this method to mammalian neuropeptidomic studies. By comparing the predictive abilities of a binary logistic model trained on molluscan prohormone cleavages with the reported model, we establish the need for phyla-specific models.

Keywords

neuropeptide; prohormone processing prediction; binary logistic regression; statistical methods; mammalian prohormones

Introduction

In even simple neuronal networks, cell to cell communication is based on the interplay of a diverse set of intercellular signaling molecules, with neuropeptides making up the most complex set of such molecules. Neuropeptides are important molecules involved in many high-level functions, including maintaining homeostasis, aiding learning and memory, and influencing behavior. Large-scale neuropeptidomic studies are on the rise due to the capabilities of mass spectrometry (MS) and the increasing availability of genetic information.[1–7] With an ever-increasing number of mammalian genomes being sequenced (more than 30 as of October 2005, (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi), the need for an effective and accurate method to predict neuropeptides from genomic information has become imperative.

It is difficult to accurately predict all of the biologically active neuropeptides in a neuropeptide gene because of the large number of posttranslational modifications a prohormone undergoes en route to becoming a bioactive neuropeptide. Neuropeptides are generated from longer precursor molecules, or prohormones, via proteolytic cleavages, with most cleavages occurring at basic residues.[8, 9] However, while most cleavages occur at basic sites, only a small

*To whom correspondence should be addressed: Jonathan V. Sweedler, University of Illinois, 600 S. Mathews Ave., 63-5, Urbana IL 61801, jsweedle@uiuc.edu.

percentage of basic sites are actually cleaved.[10] Several research groups have formulated guidelines to help determine the location of these basic processing sites, with these guidelines based on the frequency of amino acids appearing nearby cleaved and non-cleaved sites.[10–13] In recent studies, Hummon et al.[14] and Duckert et al.[15] advanced the prediction of processing sites from observation to include quantitative estimation of processing probabilities. Hummon et al. employed binary logistic regression analysis, using a model trained on the processing events from 23 prohormones of the mollusk *Aplysia californica* with MS evidence of processing. Duckert et al. developed a neural network algorithm (known as ProP) trained on viral and eukaryotic proteins obtained from the Swiss-Prot database (v 39.0). Recently, Southey et al.[16] compared these two models, along with a known motif model which incorporated Arg and Lys at positions near the cleavage sites, for several RFamide peptide families. They reported that the known motif and *Aplysia* binary logistic models had higher sensitivity than the ProP model across the RFamide family in both invertebrates and vertebrates.

Here, we report a prohormone processing model developed using a binary logistic regression algorithm trained on mammalian prohormone cleavages which proves more effective than the existing *Aplysia*, known motif, and ProP models in making mammalian processing predictions. The algorithm presented herein promises to play an important role in predicting the correct proteolytic processing sites of prohormones and will, combined with the understanding of other posttranslational modifications, greatly facilitate instrument-driven characterization of neuropeptides.

Experimental

Data

The data was composed of 428 basic processing sites resulting from 39 prohormones derived from cow, human, mouse, pig and rat. For a complete list of prohormones and other details, see Table S1 of the supporting information and the website, Neuropred (http://neuroproteomics.scs.uiuc.edu/neuropred.html). To ensure high quality data, only processing sites verified by complete or partial sequencing of the peptide, MS, or immunostaining combined with an elution profile during an HPLC separation, were included. To minimize redundancy in the data, a single homologue was used for 30 of the prohormones (i.e., the same prohormone was not used from multiple animal models).

In this work, a processing site refers to a basic residue of Lys[†] (K) or Arg (R), preceded by a non-basic residue, or two of these residues in series, preceded by a non-basic residue (i.e., KK, RR, RK and KR). Cleavage was assumed to take place at the C-terminal end of a processing site. A dibasic site was never counted as two monobasic sites by considering cleavage at each basic residue; instead, it was taken as a whole unit. As tri- and tetrabasic sites occur rarely (~3%), they were not included in the data.[§] If a processing site flanked an experimentally verified neuropeptide, either at the N- or C-terminal, it was considered as a cleaved site, but if it was located within the neuropeptide, it was considered as a non-cleaved site. A neuropeptide, however, may exist in a truncated and an elongated form; thus, an internal processing site may appear as cleaved in the first case, and as non-cleaved in the latter. In such instances, the internal processing site was counted as cleaved.

[†]Three-letter abbreviations are used for single amino acids but one-letter abbreviations are used when two or more amino acids are listed consecutively such that LysArg is written as KR.

[§]In general, the basic amino acids that flank the processing sites are enzymatically removed prior to the release of the mature bioactive neuropeptide; thus, these amino acids are rarely detected by mass spectrometry. Consequently, for sites that contain multiple basic amino acids in series, the actual position of cleavage is unknown unless both peptides surrounding the cleavage site are identified.

Non-mammalian data

Insect prohormones and processing sites were obtained from *Apis*,[17] *Drosophila*,[18, 19] *Periplaneta*,[20, 21] and *Menduca*,[22] and the molluskan data was obtained from *Aplysia*.[14] The definitions of processing sites and cleaved sites for insects and *Aplysia* are the same as those for mammalian data.

The model

In order to determine the amino acids and how their positions influenced cleavage, 18 positions that surrounded the cleavage sites were examined. In accordance with prior nomenclature,[10, 14] the nine positions N-terminal to the cleavage site were designated as M1–M9, and those at the C-terminal as P1–P9, where the numbers 1 to 9 indicated increasing distance from the cleavage site (Figure 1). When there were less than 18 amino acids surrounding the cleavage site (because the site was located close to the C- or N-terminal of the prohormone), a dummy amino acid (z) was assigned to each unoccupied position. By definition, the M1 position is always occupied by a basic residue, which is the C-terminal residue of a processing site. For example, if the processing site is KR, Arg occupies the M1 position and Lys the M2 position. The KR site is not visited again using the Lys in the M1 position.

The data was randomly divided into five groups, each containing 85 or 86 processing sites. Four of these groups were combined to create a training dataset and the remaining group was used as a test set; this was repeated five times, each time using one of the five groups as a test set and combining the other four into the training set. Thus, a total of five training sets and corresponding test sets were created.

Binary logistic regression

We employed binary logistic regression analysis using the Minitab statistical software (Release 13, Minitab Inc., State College, PA) to determine the important amino acids and positions relative to the cleavage sites that influence the probability of cleavage. For each training set, only those combinations of amino acid and position that were significantly associated with cleavage ($P<0.1$) were selected. In addition, additional explanatory variables were manually selected using two criteria. First, for a given position, a residue must occur more (or less) than the average frequency of that residue (computed by dividing the total occurrence of the residue by 18 positions); and second, the ratio of cleaved to non-cleaved (or non-cleaved to cleaved) must be greater or equal to 1.75. Various combinations of the initial explanatory variables were then regressed iteratively and a cutoff p-value of 0.1 was set to identify 15 or less significant explanatory variables. To construct the final model, explanatory variables identified in at least two training sets were regressed against a full dataset containing all 428 processing sites, and the most significant explanatory variables were identified by setting the threshold p-value to 0.05.

Comparisons with other models

The final model was compared to the three other models: *Aplysia* binary logistic model (trained with prohormones from *Aplysia*),[14] ProP (with the general PC option selected),[15] and a known motif model (defined by the presence or absence of either RR, KK, KR or RxxR where x is any amino acid except Lys or Arg).[16] Using the mammalian dataset, performance of the models was compared on the basis of the number of true positive, false negative, false positive and true negative results, their sensitivity and specificity, rate of correct classification, positive and negative predictive power, and correlation. These terms are defined as follows:

True positive (TP): the number of cleaved sites that the model classified correctly.

False negative (FN): the number of cleaved sites that the model misclassified as non-cleaved.

False positive (FP): the number of non-cleaved sites that the model misclassified.

True negative (TN): the number of non-cleaved sites that the model classified correctly.

Sensitivity: the proportion of cleaved sites correctly classified, computed as TP/(TP+FN).

Specificity: the proportion of non-cleaved sites correctly classified, computed as TN/(TN+FP).

Correct classification rate: the proportion of the total number of predictions that were correct, computed as (TP+TN)/(TP+FN+FP+FN).

Positive predictive power (PPP): the proportion of the predicted cleaved sites that were correct, computed as TP/(TP+FP).

Negative predictive power (NPP): the proportion of the predicted non-cleaved sites that were correct, computed as TN/(TN+FN).

Correlation: Matthew's Correlation coefficient computed as

$$((TP * TN) - (FP * FN))/SQRT((TN+FN) * (TN+FP) * (TP+FN) * (TP+FP)).$$

Results and Discussion

The composition of the 428 basic sites for the prohormones used while generating the model are summarized in Table 1; there are 293 monobasic and 135 dibasic sites. Dibasic sites are cleaved more frequently (77%) than monobasic sites (13%); KR sites are cleaved virtually all of the time, while RR sites are cleaved 7 times out of 10.

Interpretation of the model

The final model is constructed using 5 training sets, as described in the Experimental section. The performance of each training dataset on the corresponding test dataset is presented in Table S2 of the supporting information. The final model consists of 11 explanatory variables (Table 2) whose negative (positive) coefficients indicate a decrease (an increase) in the probability of cleavage. The model constant encompasses the effect of Lys in the M1 location, and all other position and amino acid combinations that are not provided in the final model. Under the final model, the probability of cleavage with just the model constant is less than 0.2%. There is no single amino acid and position combination that is sufficient to provide a probability of cleavage greater than 50%. The R term at M1 means that Lys in the M1 position is replaced by Arg and increases the probability of cleavage from 0.2% to 5.0%. Consequently, multiple terms are required to provide a high probability of cleavage. As an example, for the probability of cleavage to exceed 50% with Lys at M1, the sum of the coefficients needs to be greater than 6.3 such that more than one explanatory variable is required. In contrast, for the probability of cleavage to exceed 50% with Arg at M1 requires that the sum of the coefficients be greater than 2.9. This is readily accomplished by including only 1 out of 4 terms whose coefficients are greater than 2.9 (e.g., Arg at M2).

Is there any biological relevance to the explanatory variable?

A group of enzymes, generally known as protein convertases (PCs), are responsible for prohormone cleavages at basic sites.[23–25] The substrate specificity of these enzymes has been studied extensively in regions surrounding the cleavage sites, mainly at the M6 to P4 positions.[13, 26–29] The explanatory variables of our mammalian model that reside within this well-studied range agree with those studies. For example, the coefficients for Arg and Lys at the M2 position have positive values suggesting that dibasic sites are favorable for cleavage.[13,

[26] The negative coefficients of Pro at the M3 position corroborate previous studies noting the general absence of Pro at this position,[10, 12, 13] although Pro is found at M2 and P1 positions of cleaved sites.[30] Also, the presence of Arg at the M4 position is required for substrate recognition by furin, one of the protein convertases.[26] One widely accepted empirical rule for prohormone processing states that protein convertases recognize sites that contain two basic residues separated by 0, 2, 4, 6 or 8 residues ([R/K]-X_n-[R/K]↓, where n = 0, 2, 4, 6, or 8, X denotes any residue except Cys, and ↓ denotes the cleavage site).[13, 25] While Arg at the M4 position agrees with the consensus rule, we find an even distribution of basic residues at positions M5 to M9 of both the cleaved and non-cleaved sites, suggesting that there is no marked preference for basic residues at even intervals beyond M4.

The wide range of positions (M8–P8) occupied by the explanatory variables of our model is surprising, suggesting that residues located far from the various cleavage sites are also important in determining cleavage, and should be included in future studies on substrate structure for these cleavage enzymes. All explanatory residues, except Pro and Met, are polar residues and many are also charged. As our model contains multiple explanatory variables in positions outside of the range of M6–P4, it is difficult to comment on their binding properties and structural significance.

Evaluation of the model

The performance of the model was assessed by setting the cleavage threshold at 50%, selected because it optimizes the combination of sensitivity and specificity; however, a different threshold can be chosen to maximize one at the expense of another. The sensitivity, specificity, and correct classification rate of the model with 11 explanatory variables, computed at 50% cleavage threshold, are 88.0%, 92.7% and 91.1%, respectively (Table 3).

This model predicts processing at dibasic sites well, especially the KK and KR sites (Table 4). It is notable that for KK sites, where there is a relatively even distribution of cleaved and non-cleaved sites, the model predicts processing sites with great accuracy. Among dibasic sites, the highest misclassification rate occurs for RR sites. As discussed in the previous section, the probability of cleavage for an RR is greater than 50%, suggesting that RR sites will always predict cleavage unless another explanatory variable with a negative coefficient is present. Indeed, the only two RR sites that are correctly predicted as non-cleaved sites contain Pro at the M3 position. Nonetheless, as one goal is to maximize sensitivity by allowing over-prediction of processed sites, the misclassification of RR sites is acceptable.

Only a small fraction of monobasic sites are actually cleaved. In fact, only 3 out of 113 monobasic Lys sites are cleaved, and only 20% of Arg sites are cleaved (Table 1). A small subset of monobasic Arg sites (24) contains additional Arg at the M4 position, and thereby satisfies the minimal sequence requirement of the proteolytic enzyme furin.[26, 31] However, PC1 and PC2, the main processing enzymes of prohormones that follow the secretory pathway, may also require a basic residue at the M4 position.[13] Although the RxxR pattern by itself signifies only a 35% cleavage probability, 10 out of the 14 cleaved sites with this motif are correctly predicted because additional explanatory variables, such as Pro at P4 and Glu at M8, were present in the sequence. Of the remaining 153 monobasic Arg sites lacking Arg at M4, 21 are cleaved and 62% are predicted correctly. If Arg is located at the M4 position, the overall increase in sensitivity for monobasic Arg sites is ~5%. Given the difficulty of distinguishing cleaved monobasic processing sites, achieving a sensitivity of 66% is indeed a significant step towards accurate prediction of prohormone processing sites, and, by extension, their bioactive neuropeptides.

Comparison with other models

The individual performance of the known motif, *Aplysia*, and ProP models is evaluated using the mammalian data (Table 3). With its overall performance edging close to our mammalian model, the known motif model is the most sensitive of the three; it is remarkable that by simply assuming that RR, KK, KR and RxxR sites are cleaved, a sensitivity of 85%, and correct classification rate of 87%, can be achieved. The known motif is good at predicting processing sites that display a marked preference for cleavage (e.g., KR) or non-cleavage (e.g., Lys) but tends to miss ambiguous cases such as KK and monobasic Arg sites that lack the additional Arg at the M4 position (xxxR). For example, the known motif incorrectly predicts all of the cleaved xxxR sites and all of the non-cleaved KK sites. In contrast, the mammalian model predicts 65% of the cleaved xxxR sites correctly and misses only one of the eight non-cleaved KK sites. Therefore, the mammalian model is better than the known motif at identifying difficult-to-predict monobasic and dibasic sites.

The *Aplysia* logistic model, although trained on *Aplysia* prohormones, correctly predicts 75% of processing events in mammalian prohormones. A comparison between the distribution of processing sites used to train the *Aplysia* and mammalian models (Table 5) provides one reason why the two models perform differently. The mammalian prohormones contain more RR, Lys and KK sites, but the latter two are cleaved less frequently in our mammalian dataset. Moreover, nearly all the dibasic sites are cleaved in *Aplysia*, but more discrimination is required in mammals. There is greater similarity in processing-site distribution between insects and *Aplysia* than there is between mammals and *Aplysia*. In addition, the organization of prohormones in mammals and *Aplysia* is often distinct; many of the *Aplysia* prohormones contain multiple short, repeating motifs flanked by dibasic sites within the prohormone, whereas fewer repeats and cleavage sites exist in mammals. These variations in processing-site distribution and prohormone organization suggest the importance for phyla-specific models in achieving more accurate prediction of processing sites.

The sensitivity, specificity and correct classification rates of ProP with the general PC option, as tested on mammalian prohormones, are 28.2%, 97.2% and 74.3%, respectively. While ProP filters out the non-cleaved sites correctly, it predicts cleaved sites rather poorly. Several proteins in the ProP training data are derived from eukaryotes and viruses, which may be one reason for the reduced performance of the model. As noted earlier, the differences that exist in *Aplysia* and mammalian prohormone organization and frequency of processing sites also may occur in other phyla. Thus, the homogenized use of sequences may cause blurring of mammalian-specific processing events, and result in reduced predictive ability. As with all models, the quality of data used for training is critical. Duckert et al.[15] excluded experimentally unsubstantiated processing sites from their data by removing proteins annotated with terms such as ‘probable’, ‘by similarity’, or ‘potential’. During our search for prohormones, we encountered entries in the Swiss-Prot database where the processing information was not experimentally confirmed, but derived by comparison with other homologues—yet this fact was not explicitly stated. Hence, the quality of data in ProP may be somewhat compromised, despite the careful measures the authors took to avoid non-empirical evidence.

Application

How can this model be used for neuropeptide prediction? Let’s take the example of the 26-RFa. This peptide, originally discovered in the frog brain,[32] belongs to the FMRFamide-related peptide family. Mammalian homologues have been identified in human, rat, and mouse by *in situ* hybridization and cloning.[32, 33] In these organisms, the peptide 26-RFa is found at the C-terminal of the 26-RFa prohormone, is preceded by a Lys and ends with one or two Arg, depending on the species. The human 26-RFa is shown to have an orexigenic effect in mice.[32] Jiang et al.[33] also found that the 26-RFa is an agonist to the orphan receptor, GCRP SP9155.

However, these peptides have not been experimentally confirmed; therefore, it is not known if 26-RFa is a naturally occurring peptide in mammals, and if there are other products of the prohormone. Our model helps us answer these questions by identifying the most likely prohormone cleavage sites in the prohormone, leading to predictions of the final peptides. The mouse prohormone contains 6 Arg, 1 Lys, 1 RR and 1 RRK sites. Our model predicts cleavage at the tribasic site that is part of the 26-RFa sequence, suggesting a truncated RFamide peptide. It also predicts cleavage at the RR and the first N-terminal Arg site. Another Arg following the sequence FRLG is predicted with a probability of 48%. Based on multiple sequence alignments, FRLGR in mouse corresponds to the internal FRFGR sequence in human, which has a cleavage probability of only 6%, indicating that an additional RFamide may not be generated from the human sequence. Lastly, two additional processing sites are predicted in the rat prohormone. Based on the predicted cleavage sites, our model not only indicates the peptides that are most likely to be observed experimentally, it suggests that processing of the 26-RFa prohormone may be different in these three organisms. Further studies on the processing of this prohormone will provide experimental data that will help to validate the predictions of the proposed model.

Conclusions

The immediate application of this model is to facilitate MS characterization of neuropeptides from genetic information. There are many more putative cleavage sites than there are actually cleaved sites. On average, the prohormones used in our study contain 10 basic sites, out of which only a third are cleaved. If one applies no knowledge of processing, then one has to search for greater than 120 peptide masses that could be generated using any three cleavage sites of the typical mammalian prohormone. This task becomes both labor and time intensive when numerous posttranslational modifications occur on the peptides after the basic site cleavages, and when many prohormones are under investigation, as is the case in many neuropeptidomic studies. With its power to accurately predict cleaved sites, the mammalian model limits the number of putative peptides expected from a novel prohormone, and thus minimizes the time and effort required to analyze MS data. Advancing these predictive approaches will reduce the time it takes to determine novel bioactive peptides from the genetic sequence information. In fact, a significant mismatch of predictions and experimental results may even be used to look for unusual processing patterns and pinpointing regions that appear to have novel processing enzymes, thereby directing additional studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This material is based upon work supported by the National Institute on Drug Abuse (NIDA), National Institutes of Health, under Award No. P30 DA018310, to the UIUC Neuroproteomics Center for Cell to Cell Signaling.

References

1. Che FY, Fricker LD. J Mass Spectrom 2005;40:238–249. [PubMed: 15706629]
2. Haskins WE, Watson CJ, Cellar NA, Powell DH, Kennedy RT. Anal Chem 2004;76:5523–5533. [PubMed: 15362916]
3. Hummon AB, Amare A, Sweedler JV. Mass Spectrom Rev 2006;25:77–98. [PubMed: 15937922]
4. Ivanov VT, Yatskin ON. Expert Rev Proteomics 2005;2:463–473. [PubMed: 16097881]
5. Paulson L, Persson R, Karlsson G, Silberring J, Bierczynska-Krzysik A, Ekman R, Westman-Brinkmalm A. J Mass Spectrom 2005;40:202–213. [PubMed: 15706622]

6. Ramstrom M, Bergquist J. FEBS Lett 2004;567:92–95. [PubMed: 15165899]
7. Svensson M, Skold K, Svenningsson P, Andren PE. J Proteome Res 2003;2:213–219. [PubMed: 12716136]
8. Seidah, NG. The Enzymes. In: Dalbey, RE.; Sigman, DS., editors. Co- and Posttranslational Proteolysis of Proteins. 22. Academic Press; San Diego, CA: 2001. p. 237-258.
9. Steiner, D. The Enzymes. In: Dalby, RE.; Sigman, DS., editors. Co- and Posttranslational Proteolysis of Proteins. 22. Academic Press; San Diego, CA: 2001. p. 163-198.
10. Devi L. FEBS Lett 1991;280:189–194. [PubMed: 2013311]
11. Lindberg, I.; Hutton, JC. Peptide Biosynthesis and Processing. Fricker, LD., editor. CRC; Boca Raton, FL: 1991. p. 141-174.
12. Rholam M, Brakch N, Germain D, Thomas DY, Fahy C, Boussetta H, Boileau G, Cohen P. Eur J Biochem 1995;227:707–714. [PubMed: 7867629]
13. Cameron, A.; Apletalina, EV.; Lindberg, I. The Enzymes. In: Dalbey, RE.; Sigman, DS., editors. Co- and Posttranslational Proteolysis of Proteins. 22. Academic Press; San Diego, CA: 2001. p. 291-332.
14. Hummon AB, Hummon NP, Corbin RW, Li L, Vilim FS, Weiss KR, Sweedler JV. J Proteome Res 2003;2:650–656. [PubMed: 14692459]
15. Duckert P, Brunak S, Blom N. Protein Eng Des Sel 2004;17:107–112. [PubMed: 14985543]
16. Southey BR, Rodriguez-Zas SL, Sweedler JV. Peptides. 2005in press
17. Hummon, AB.; Richmond, TA.; Verleyen, P.; Baggerman, G.; Huybrechts, J.; Ewing, MA.; Vierstraete, E.; Rodriguez-Zas, SL.; Schoofs, L.; Robinson, GE.; Sweedler, JV. 2006. submitted
18. Baggerman G, Boonen K, Verleyen P, De Loof A, Schoofs L. J Mass Spectrom 2005;40:250–260. [PubMed: 15706625]
19. Baggerman G, Cerstiaens A, De Loof A, Schoofs L. J Biol Chem 2002;277:40368–40374. [PubMed: 12171930]
20. Predel R. J Comp Neurol 2001;436:363–375. [PubMed: 11438936]
21. Predel R, Neupert S, Wicher D, Gundel M, Roth S, Derst C. Eur J Neurosci 2004;20:1499–1513. [PubMed: 15355317]
22. Audsley N, Weaver RJ. Peptides 2003;24:1465–1474. [PubMed: 14706525]
23. Seidah NG, Day R, Chretien M. Biochem Soc Trans 1993;21:685–691. [PubMed: 8224490]
24. Seidah NG, Chretien M. Curr Opin Biotechnol 1997;8:602–607. [PubMed: 9353231]
25. Seidah NG, Chretien M. Brain Res 1999;848:45–62. [PubMed: 10701998]
26. Molloy, SS.; Thomas, G. The Enzymes. In: Dalbey, RE.; Sigman, DS., editors. Co- and Posttranslational Proteolysis of Proteins. 22. Academic Press; San Diego, CA: 2001. p. 199-235.
27. Takahashi S, Hatsuzawa K, Watanabe T, Murakami K, Nakayama K. J Biochem (Tokyo) 1994;116:47–52. [PubMed: 7798185]
28. Watanabe T, Murakami K, Nakayama K. FEBS Lett 1993;320:215–218. [PubMed: 8462689]
29. Watanabe T, Nakagawa T, Ikemizu J, Nagahama M, Murakami K, Nakayama K. J Biol Chem 1992;267:8270–8274. [PubMed: 1569080]
30. Schwartz TW. FEBS Lett 1986;200:1–10. [PubMed: 3516723]
31. Molloy SS, Bresnahan PA, Leppla SH, Klimpel KR, Thomas G. J Biol Chem 1992;267:16396–16402. [PubMed: 1644824]
32. Chartrel N, Dujardin C, Anouar Y, Leprince J, Decker A, Clerens S, Do-Rego JC, Vandesande F, Llorens-Cortes C, Costentin J, Beauvillain JC, Vaudry H. Proc Natl Acad Sci USA 2003;100:15247–15252. [PubMed: 14657341]
33. Jiang Y, Luo L, Gustafson EL, Yadav D, Laverty M, Murgolo N, Vassileva G, Zeng M, Laz TM, Behan J, Qiu P, Wang L, Wang S, Bayne M, Greene J, Monsma FJ, Zhang FL. J Biol Chem 2003;278:27652–27657. [PubMed: 12714592]

Figure 1.
The processing site, with the positions surrounding the cleavage site (arrow) shown. Residues to the left of the cleavage site are indicated by negative numbers and those to the right by positive numbers signifying their distance relative to the cleavage site. In the text, the negative numbers are also indicated by "M" followed by a number such as M1. All cleavage is assumed to take place C-terminal to a processing site (i.e., after KR, but not between K and R). In constructing the binary logistic regression model, these eighteen positions surrounding the cleavage site are considered.

Table 1

Basic site composition of mammalian prohormones

Four hundred and twenty-eight basic sites from 39 mammalian prohormones are present in the data set.

Site[a]	Frequency[b]	Cleaved[c]	% Cleaved by site[d]	Composition of cleaved sites[e] (%)
KK	13 (3.0%)	5	38.5	3.5
KR	75 (17.5%)	73	97.3	51.4
RR	34 (7.9%)	24	70.6	16.9
RK	13 (3.0%)	2	15.4	1.4
R	183 (42.8%)	35	19.1	24.6
K	110 (25.7%)	3	2.7	2.1
Total	**428**	**142**	**33.2**	**100**

[a] Tri- and tetrabasic sites are not included in the data. For sites listed as monobasic (Lys and Arg), or dibasic (KK, RR, KR, RK) the next amino acid must be a non-basic residue.

[b] The frequency of occurrences of each site is given as both a count and percentage. For each of the processing sites the number

[c] and percentage

[d] of cleaved sites is given.

[e] The percent composition of the 142 cleaved sites is given for each site.

Table 2

Significant explanatory variables of the mammalian model

An explanatory variable is an amino acid and the position it occupies relative to the cleavage site. The letters in column 1 indicate whether the residue is located at the N-terminal (M) or at C-terminal (P) of the cleavage site, whereas the numbers indicate increased distance from the cleavage site.

Position	Amino Acid	Coeff.	P value
Constant		−6.3291	0.000
M1	R	3.3757	0.000
M2	K	5.2732	0.000
M2	R	3.0561	0.000
M3	P	−2.411	0.036
M4	M	2.9775	0.000
M4	R	2.3414	0.000
M8	E	1.7691	0.010
P1	S	1.9357	0.001
P4	D	1.5376	0.026
P7	T	3.1275	0.000
P8	E	1.8426	0.009

Table 3

Performance of different models is compared to the mammalian model

In the known motif model, the KR, RR sites and the RxxR sites are considered cleaved 100% of the time. The *Aplysia* model is a binary logistic regression model trained using data from prohormone processing from the mollusk, *Aplysia californica*[14] and tested on the mammalian dataset. ProP[15] is a neural network based processing site predictor available online (http://www.cbs.dtu.dk/services/ProP/) with the general PC model used here. Lastly, the known motif defines cleavages as occurring after RR, KK, KR or RxxR, and assumes all other combinations remain uncleaved.[16]

Model	Mammalian	*Aplysia*	Known Motif	ProP
True positive	125	99	121	40
False negative	17	43	21	102
False positive	21	65	38	8
True negative	265	221	248	278
Sensitivity	88.0%	69.7%	85.2%	28.2%
Specificity	92.7%	77.3%	86.7%	97.2%
Correct classification rate	91.1%	74.8%	86.2%	74.3%
Positive predictive power	85.6%	60.4%	76.1%	83.3%
Negative predictive power	94.0%	83.7%	92.4%	73.2%
Correlation	0.80	0.46	0.72	0.38

Table 4

The performance of the mammalian model by putative cleavage site

The mammalian model's predictive power on the 428 processing sites is listed:

	KK[a]	KR	RR	RK	K	R
True positive	5	73	24	0	0	23
False negative	0	0	0	2	3	12
False positive	1	2	8	0	0	10
True negative	7	0	2	12	107	138

[a]The processing sites are categorized as monobasic (a single Arg or Lys site preceded by a non-basic amino acid) or dibasic (KK, KR, RR and RK) sites.

Table 5

Distribution of processing sites in mammals, *Aplysia*, and insects

The numbers for *Aplysia*[14] and insects were obtained from the literature. The insect data is derived from *Apis*,[17] *Drosophila*,[18, 19] *Periplaneta*,[20, 21] and *Menduca*[22] prohormones.

	Mammals		*Aplysia*		Insects	
Site	**Total**	**% Cleaved**	**Total**	**% Cleaved**	**Total**	**% Cleaved**
KK	13	38.5	15	100	15	93.3
KR	75	97.3	365	100	100	98.0
RR	34	70.6	15	60.0	23	87.0
RK	13	15.4	2	50.0	3	33.3
R	183	19.1	266	17.3	176	27.8
K	110	2.7	60	45.0	38	20.1