Protein Engineering, Design & Selection vol. 17 no. 1 pp. 107–112, 2004
DOI: 10.1093/protein/gzh013

Prediction of proprotein convertase cleavage sites

Peter Duckert, Søren Brunak and Nikolaj Blom[1]

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

[1]To whom correspondence should be addressed.
E-mail: nikob@cbs.dtu.dk

Many secretory proteins and peptides are synthesized as inactive precursors that in addition to signal peptide cleavage undergo post-translational processing to become biologically active polypeptides. Precursors are usually cleaved at sites composed of single or paired basic amino acid residues by members of the subtilisin/kexin-like proprotein convertase (PC) family. In mammals, seven members have been identified, with furin being the one first discovered and best characterized. Recently, the involvement of furin in diseases ranging from Alzheimer's disease and cancer to anthrax and Ebola fever has created additional focus on proprotein processing. We have developed a method for prediction of cleavage sites for PCs based on artificial neural networks. Two different types of neural networks have been constructed: a furin-specific network based on experimental results derived from the literature, and a general PC-specific network trained on data from the Swiss-Prot protein database. The method predicts cleavage sites in independent sequences with a sensitivity of 95% for the furin neural network and 62% for the general PC network. The ProP method is made publicly available at http://www.cbs.dtu.dk/services/ProP.

Keywords: furin/neural network/propeptide/protease specificity/secretory precursor

Introduction

Post-translational processing by limited proteolysis of inactive secretory precursors to produce active proteins and peptides is an ancient mechanism that enables cells to regulate the level of specific bioactive polypeptides and to generate diverse products from precursor molecules. Many biologically active proteins and peptides are initially synthesized as larger, inactive precursors, usually in the form of pre-pro-proteins, which are post-translationally modified to generate the mature molecule. The N-terminal pre-regions are signal peptides, which direct the precursors to the appropriate cellular compartment, while the prodomains can have several functions. Some prodomains act as intramolecular chaperones that mediate correct folding of the newly synthesized proteins, while other prodomains are only indirectly involved in folding and have other functions such as transport and localization, oligomerization, regulation of activity (Shinde and Inouye, 2000) and quality control of folding (Bauskin *et al*., 2000). Precursor cleavage frequently occurs at motifs containing multiple basic residues, arginine or lysine, by limited endoproteolysis of the corresponding precursor proteins immediately C-terminally of the basic amino acid motifs (Seidah and Chretien, 1999).

Most of the enzymes responsible for this type of processing belong to a family of evolutionary conserved dibasic- and monobasic-specific Ca^{2+}-dependent serine proteases called subtilisin/kexin-like proprotein convertases (PCs) (Seidah *et al*., 1998). Since the identification of the yeast enzyme Kex2 (kexin), seven mammalian kexin-related proprotein convertases have been identified: PC1, PC2, furin, PC4, PC5, PACE4 and PC7. Furin was the first PC to be discovered; a database search identified furin as the first mammalian kexin homolog (Fuller *et al*., 1989), and furin was soon shown to correctly process the precursors of von Willebrand factor and beta-nerve growth factor (Bresnahan *et al*., 1990; van de Ven *et al*., 1990).

The PCs are the major endoproteolytic processing enzymes of the secretory pathway in mammals (Steiner, 1998). These enzymes process precursors at sites, which usually contain the consensus sequence [R/K]–X_n–[R/K]↓, where X indicates any amino acid residue, *n*, the number of spacer amino acid residues, is 0, 2, 4 or 6, [R/K] indicates either an arginine or a lysine residue, and the arrow (↓) indicates the site of cleavage (Seidah and Chretien, 1999). Furin has a more stringent substrate specificity and preferentially recognizes sites that contain the sequence motif R–X–[R/K]–R↓ (Nakayama, 1997).

Recently, the first reports on the crystal structure of members of the PC family, namely furin and Kex2, were published (Henrich *et al*., 2003; Holyoak *et al*., 2003). The detailed analysis of the proteolytic domains explains the preference for basic residues, in particular arginine, at the substrate cleavage site P1. In addition, the preference observed for other positions, in particular for P4 to P2 in furin, fits well with the observed enzyme–substrate crystal structure.

The PCs are involved in the activation of a large variety of proteins like peptide hormones, neuropeptides, growth and differentiation factors, adhesion factors, receptors, blood coagulation factors, plasma proteins, extracellular matrix proteins, proteases, exogenous proteins such as coat glycoproteins from infectious viruses (e.g. HIV-1 and influenza virus) and bacterial toxins (e.g. diphtheria and anthrax toxin). PCs not only catalyze removal of prodomains, but are also involved in processing of multifunctional precursors like POMC and proglucagon. Both undergo differential processing, dependent on their sites of production (Steiner, 2002). Therefore, PCs play an essential role in many vital biological processes like embryonic development and neural function, and in viral and bacterial pathogenesis. In addition, PCs are implicated in a number of pathologies such as cancer and neurodegenerative diseases (Thomas, 2002).

From a medical and biotechnological viewpoint it can be of interest to control the production of peptides involved in various diseases by specifically inhibiting the enzymatic

activity of the PCs (Bergeron *et al*., 2000; Thomas, 2002). Several naturally occurring sequences are known to inhibit PCs, including the prodomains of the proteases themselves and the neuroendocrine proteins 7B2 and proSAAS (Cameron *et al*., 2002; Rockwell *et al*., 2002).

We present here work on the characterization and prediction of PC cleavage sites. Since not all single or paired basic amino acid residues are potential cleavage sites of PCs, we examined the sequence patterns characteristic of experimentally verified sites and describe a neural network based method for predicting whether a given site is a potential cleavage site for the PC enzymes. Not all furin cleavage sites contain the furin consensus sequence and not all sites containing the consensus sequence are cleavage sites for furin (Nakayama, 1997). Furin also has a preference for basic amino acid residues at P3, P5 and P6 (even at P7 and P8), while other specific requirements exist at P1′ and P2′. For instance, lysine residues are not accepted at the latter positions (Henrich *et al*., 2003). Favorable residues at P2 and P6 can compensate for less favorable ones at P1 and P4 (Molloy and Thomas, 2002). Together, these facts indicate that the furin recognition motif is non-linear and more complicated than previously believed. Therefore, a method capable of classifying complex patterns, such as a neural network, may be more suitable than a simple consensus sequence, which is not able to include correlated effects often termed 'exceptions'. In the case of non-furin cleavage sites, that are primarily characterized by a dibasic motif, the problem of false predictions by using a simple dibasic consensus pattern becomes evident and necessitates the use of a more sophisticated classifier. The neural network based method described in this paper can be used to predict PC cleavage sites in single proteins, but may also be used when large databases are scanned for novel growth factors, hormones and secreted peptides containing propeptides or for automated annotation of large protein sets.

Materials and methods

Two data sets were collected: (i) furin cleavage sites and (ii) general PC cleavage sites. The latter should ideally represent cleavage sites for all PCs. The furin cleavage data were taken from the literature (Nakayama, 1997; Lehmann *et al*., 1998) and consisted of 38 proteins, which contained reported furin sites, including proteins from mammals and from pathogenic bacteria and viruses.

The general proprotein convertase data set was based on Swiss-Prot annotations (release 39.0, 05/00; Bairoch and Apweiler, 2000) and also included the furin cleavage data described above, totaling 235 sites in 227 proteins. The following search criteria were used in the extraction of data from the Swiss-Prot database: signal peptide present, experimentally determined propeptide region not having comments like 'PROBABLE', 'BY SIMILARITY', or 'POTENTIAL' (Junker *et al*., 1999), and a basic residue (K or R) at P1, i.e. the first position N-terminal to the cleavage site. Proteins were from eukaryotes and viruses only, and redundancy in the data set was reduced by removal of doublet sequences, which had identical 13-mer amino acid sequences centered around the cleavage site.

Sequence logos of precursor proteins, aligned by their cleavage sites, were used for displaying the position-specific features of multiple sequence alignments (Schneider and Stephens, 1990).

In order to avoid training and testing on homologous sequences, each data set was divided into four partitions of approximately equal size based on phylogenetic trees, which were constructed using multiple alignments and the neighbour-joining algorithm of the ClustalX program (Thompson *et al*., 1997). This further eliminates the redundancy issue, and the risk that the predictive performance is overestimated owing to training and test set similarities.

The neural network architecture was feed-forward fully connected, with zero or one layer of two, four, eight or 16 hidden units (Baldi and Brunak, 2001). The neural networks were trained by back-propagation (Rumelhart *et al*., 1986), and the sequence data were presented to the network using sparsely encoded moving windows (Qian and Sejnowski, 1988; Brunak *et al*., 1991). Only positions with lysine or arginine were used during training and testing. Symmetric input windows with size varying from five to 23 positions were tested. The learning rate was 0.005, and the decision threshold value for the output unit was 0.5 for all networks.

A correlation coefficient (Matthews, 1975) was calculated from the numbers of correctly and incorrectly predicted positive and negative sequence windows generated with the selected threshold value. The correlation coefficients of both the training and test sets were monitored during training, and the performance at the training cycle with the maximal test set correlation was recorded for each training run. The test performances were calculated by 4-fold cross-validation: every network run was carried out with one part as the test data and the other three parts as the training data. For each of the four combinations, one neural network architecture was chosen based on the test set correlation coefficients. If more than one network architecture gave rise to approximately the same correlation coefficient, we chose the smallest network with respect to input window size and the number of hidden units. The combined performance for each of the four subsets when used as test sets was then calculated. The cross-validated Matthews correlation coefficient for the combined ensemble was calculated from the total numbers of true positives, true negatives, annotated cleavage sites and annotated non-cleavage sites.

The trained networks provide scores between zero and one for each arginine (R) or lysine (L) residue in an amino acid sequence. The combined prediction scores were then calculated as an average over the scores from the four networks. By default, for scores above 0.5 we predict peptide bond cleavage at the C-terminal side of the amino acid residue in question.

To illustrate the prediction performance for varying thresholds and prediction sensitivities, we present receiver operating characteristic (ROC) curves, plotting sensitivity on the *x*-axis and false-positive rate on the *y*-axis.

Results

Sequence logos

The sequence pattern at the cleavage sites is shown as sequence information logos in Figure 1. Sequence logos were generated for each of the two data sets, with P1 as the central position and seven flanking amino acid residues on each side. The sequence logos emphasize amino acid residues that are frequently found at the propeptide cleavage sites, but cannot reveal correlated effects. Note that the data sets were redundancy reduced prior to the logo analysis.

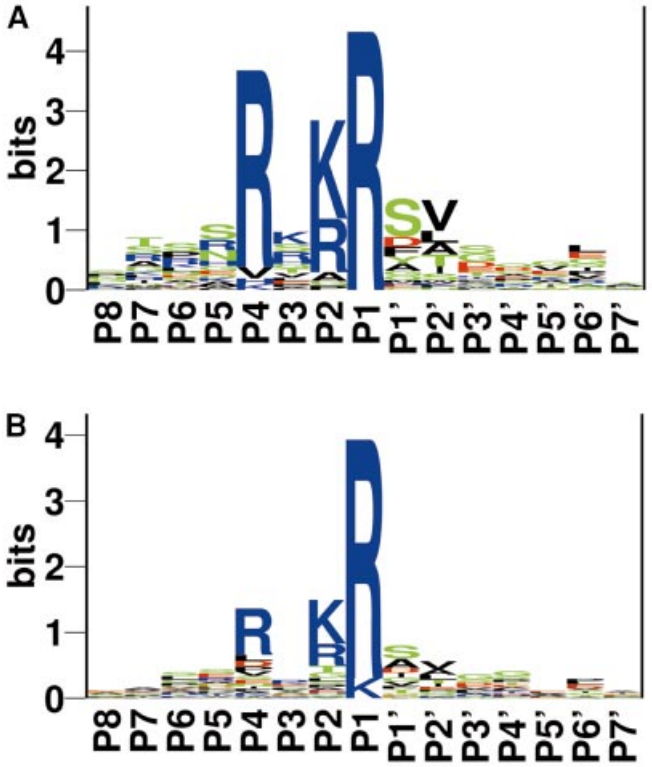

Fig. 1. Sequence logos of aligned propeptide cleavage sites centered at P1, where cleavage takes place between P1 and P1′. Shannon information is shown in units of bits: (**A**) 38 experimentally determined furin sites found in the literature; (**B**) 234 general PC sites extracted from the Swiss-Prot database.

Furin sequence logo

Furin preferentially recognizes the cleavage site sequence R–X–[R/K]–R↓ (Nakayama, 1997). This consensus cleavage sequence is clearly visible from the logo, and all furin cleavage sites in the data set had an arginine residue at position P1. The frequency of arginine at position P4 was 89%, while only one of the cleavage motifs (3%) contained a lysine residue at P4. Of the remaining amino acids only histidine (3%) and valine (5%) were found at this position. At P2, R and K were the most frequent, 32 and 58%, respectively, and the remaining amino acids found at this position were alanine (5%), glycine (3%) and proline (3%). Serine was the most frequent amino acid residue at P1′ (42%), while leucine, isoleucine and valine did not occur at this position, but these amino acids were frequent at P2′, 55% in total. Both phenylalanine and tyrosine were found at P1′, four (11%) and three (8%), respectively, but not tryptophan, arginine, lysine, asparagine, proline or threonine. The acidic amino acids glutamic acid and aspartic acid occurred at P1′ in one (3%) and four (11%) of the cleavage motifs, and histidine was found in one (3%) of the cleavage motifs. At P2′, arginine was absent, and only one lysine residue was found at this position.

Of the 38 cleavage motifs in the furin data set, 31 (81%) had the R–X–[R/K]–R consensus sequence. Eleven (29%) of these were R–X–R–R and 20 (52%) were R–X–K–R. The P2 basic amino acid residue is not essential, but can greatly enhance the processing efficiency (Thomas, 2002). Therefore, the minimal furin cleavage sequence is R–X–X–R, but a less favorable amino acid residue at P4 can be compensated for by arginine or lysine residues at P2 and P6. Three (8%) of the cleavage motifs contained only the minimal furin cleavage sequence, while the remaining four (11%) did not have an arginine residue at P4. In earlier studies, the following general rules were proposed for furin specificity: (i) an arginine residue is essential at the P1 position; (ii) in addition to the P1 arginine residue, at least two out of the three residues at P2, P4 and P6 are required to be basic for efficient cleavage; (iii) at the P1′ position, an amino acid residue with a hydrophobic aliphatic side chain (i.e. leucine, isoleucine or valine) is not suitable (Nakayama, 1997).

These rules are consistent with most of the furin cleavage sites in the data set: arginine was always present at P1 and leucine, isoleucine or valine were never observed at P1′. In contrast, two of the cleavage sites did not follow the rule of at least two basic amino acid residues at P2, P4 and P6, namely human insulin-like growth factor IA precursor (LKPAKSAR↓) and shiga toxin A-chain precursor (HHASRVAR↓). In these cases, a lysine at P7 or two histidines at P7 and P8, respectively, may have compensatory effects.

General PC sequence logo

The general PC sequence logo is based on 234 of the 235 cleavage sites, which were selected as having R or K at P1. One of the 235 cleavage sites is located only three residues from the C-terminus of the protein and therefore, for that single site, it was not possible to extract seven residues C-terminally of the cleavage site as required for generating the sequence logo in Figure 1B.

Arginine was by far the most frequent amino acid residue at P1, corresponding to 92%. At P2 the frequencies of R and K were 22 and 43%, respectively, while the frequency of R was 50% at P4. Only six lysine residues were found at this position, corresponding to 3%. At the P1′ position the frequency of serine was 24%, while the frequency of the hydrophobic, aliphatic amino acids leucine, isoleucine and valine was 17% in total. The furin consensus sequence R–X–[R/K]–R can also be recognized in this logo, but to at much lesser extent than in the furin logo (Figure 1A). The furin consensus sequence was found in 104 (44%) of the cleavage sites, and the minimal furin consensus sequence R–X–X–R appeared in 117 (50%) of the cleavage sites. These observations indicate that up to 50% of the cleavage sites in the general data set may be substrates of non-furin PCs.

Performance of the neural networks

The furin network was trained on a data set containing 38 experimentally verified cleavage sites out of a total number of 3004 sites containing R or K at the P1 position, of which 1589 had an R at P1. Using a 4-fold cross-validated training approach, the values for the optimal symmetric window size/hidden units were 13/2, 19/4, 17/8 and 11/2, respectively, for the four networks. Thus, optimal window size range from 11 to 19 residues, which correspond well to the contact area of other known protein–protein interactions, e.g. kinase–substrate contact (10–11 residues; Blom *et al.*, 1999). The fact that there were at least two hidden units in the networks indicates that the PC cleavage site prediction problem is indeed non-linearly separable only owing to the correlations and compensatory effects in the cleavage motifs (Baldi and Brunak, 2001). The number of true positives predicted reached 94.7% with a specificity of 83.7%, equivalent to a Matthews correlation coefficient of 0.89.

From a practical point of view, the most important aspect of a prediction method is its ability to make correct predictions.

As prediction methods are never perfect, one is always faced with the dilemma of choosing between making few false-positive predictions and having a high sensitivity, i.e. correctly identifying as many positive examples as possible. This trade-off can be visualized as what is known as the ROC curve in which the rate of false positives is plotted as a function of the sensitivity by varying the score threshold used for making positive predictions. Figure 2A shows the ROC curve for the furin site predictor, where the standard threshold of 0.5 corresponds to a false-positive rate of 0.2% at a sensitivity of 94.7%. When the sensitivity is increased further, the false-positive rate rises steeply and reaches ~30% at a sensitivity of 100%, i.e. when all cleavage sites are predicted correctly. For neural networks without hidden units, the highest correlation coefficient was 0.83, and the specificity of the corresponding network was 76.1% at the default threshold of 0.5.

Four of the known cleavage sites in the furin data set lacked the arginine residue at the P4 position characteristic of the furin consensus sequence. Two of these sites were correctly learned by the neural network and classified correctly, even when being part of an independent test set: human vitamin-K-dependent protein C precursor (KKRSHLKR^{199}↓DTED; score 0.797) and human parathyroid hormone precursor (DGKSVKKR^{31}↓SVSE; score 0.851). These findings indicate that the neural network has picked up sequence correlations, which indicate furin cleavage in non-consensus cleavage sites. This also may explain the fact that a hidden layer of neurons is required for optimal performance of the neural network.

Two of the known cleavage sites in the furin data set were not predicted correctly, i.e. they were false negatives when being part of an independent test set: human insulin-like growth factor IA precursor (LKPAKSAR^{119}↓SVRA) and human serum albumin precursor (YSRGVFRR^{24}↓DAHK). The cleavage scores produced by the furin-specific neural network were low, 0.100 and 0.128, respectively. In the case of human insulin-like growth factor IA precursor cleavage has been observed in furin-deficient cell types indicating that another processing enzyme may be acting *in vivo* (Duguay *et al*., 1995). For human serum albumin precursor, PCs other than furin, namely PACE4 and PC7, have been suggested to be involved in processing (Mori *et al*., 1999), and this may explain the low cleavage site score of the furin-specific network. In this sense, the network may be more correct than the statistics on the experimental data indicate.

Another caveat when using the furin-type cleavage site method is that prediction of processing by furin does not mean that the substrate is actually cleaved by furin *in vivo*. Three other proprotein convertases, PACE4, PC5 and PC7, have sequence specificities similar to that of furin (Nakayama, 1997). Since there seems to be a considerable overlap in substrate specificity between the different PC members, predictions for a given PC, e.g. furin, should be interpreted with this in mind.

After implementation of the neural network prediction method for furin-type cleavage sites, we scanned the published literature for recent reports on furin-mediated cleavage in proteins, which were not used in our training data. We found three examples reported as furin cleavage sites and correctly predicted by our furin-type network. All reported sites matched the R–X–[R/K]–R consensus: human ectodysplasin-A RVRRNKR^{159}↓SK (score 0.819) (Chen *et al*., 2001), Marburg virus glycoprotein VYFRRKR^{435}↓SI (score 0.712) (Volchkov *et al*., 2000) and mouse ZP3 glycoprotein

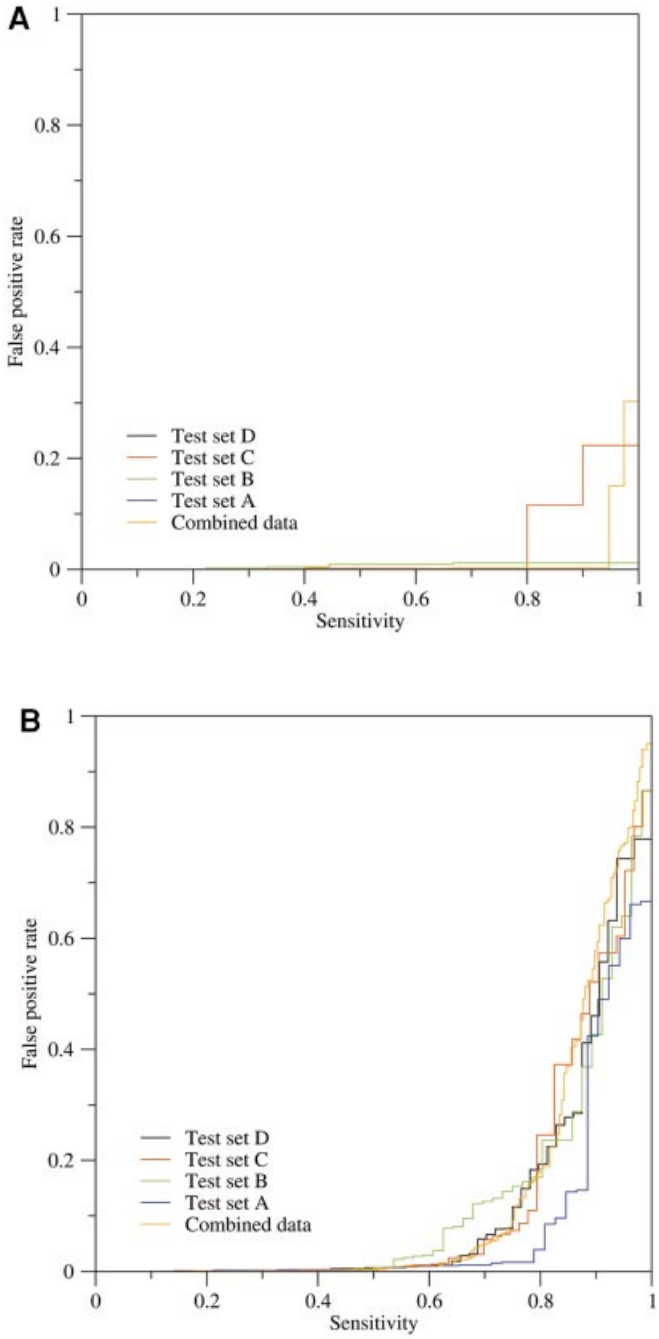

Fig. 2. The predictive performance shown as sensitivity versus false-positive rate in a ROC diagram. The plot was constructed from results obtained for the independent test sets, and corresponds to the expected performance for novel, unrelated proteins. (**A**) The furin network performance; (**B**) the general PC network. For a given category, e.g. the combined data for the general PC neural network, a sensitivity of 80% can be achieved with a false-positive rate of 20%, corresponding to 80% correct prediction on both categories, i.e. true positives and true negatives. Random performance would correspond to a line along the diagonal.

LVSRNRR^{353}↓HV (score 0.762) (Williams and Wassarman, 2001). More interestingly, our neural network method was able to correctly classify consensus-matching non-cleavage sites in two of these proteins, namely human ectodysplasin-A ESRRVRR^{156}↓NK (score 0.482) and Marburg virus glycoprotein LVCRLRR^{561}↓LA (score 0.446). Again, these findings indicate that the neural network algorithm is able to incorporate

sequence features other than consensus sequence alone and may be well suited for rejecting putative furin cleavage sites which match the consensus sequence perfectly.

It was recently claimed, based on the crystal structure of Kex2, that substrates with a glutamate residue at P3 might be selected against (Holyoak *et al*., 2003). In our data set of negative furin sites we found an example of a site that strongly matches the furin consensus site but nevertheless is predicted as being non-cleaved. In human integrin alpha-3 precursor (ITA3_HUMAN) the site RDRR^{525}↓PP (score 0.207) contains a glutamate (D) at P3, which may, in addition to a non-favored proline at P1′, result in the correct, non-cleaved classification.

The general PC network was trained on 235 cleavage sites out of a total of 11 399 R or K residues. Of these, 372 were false positives eliminated during the learning process in order to enhance performance (339 R and 33 K), leaving 10 792 negative sites with R or K at position P1. The elimination of potentially false positives, which had a detrimental effect on the predictive performance, was done by monitoring in which order the network did learn the positives and negatives (Blom *et al*., 1999). A 4-fold cross-validated training gave the following optimal network architecture (window size versus number of hidden units): 13/16, 13/4, 13/4 and 11/8, respectively, for the four networks. At the standard threshold of 0.5, the sensitivity was 61.7%, the specificity was 59.7%, and the Matthews correlation coefficient was 0.60 for the best ensemble of networks. The ROC curve (Figure 2B) showed that the 0.5 threshold was a good choice, since the false-positive rate was 0.9% at a sensitivity of 61.7%. The rise in false-positive rate is moderate until a sensitivity of ~75%, where the false-positive rate is ~7%, but this is still a high number because the number of true negative sites is much larger than the number of true positive sites. The importance of hidden units is again evidenced by the fact that the highest correlation coefficient for a network without hidden units was 0.41, and the specificity of this network was only 28.9%.

We estimate that furin-like sites are frequently represented in the general PC data set, since the furin consensus sequence, R–X–[R/K]–R, is present in 44% of the cleavage sites. This may explain the difficulties when predicting PC1 and PC2 sites, as can be seen from the following examples. Comparing our prediction performance on PC1 and PC2 proprotein processing sites, which were reported late in our study, we observed that not all reported sites could be correctly predicted (Cameron *et al*., 2002). In particular, of the 12 PC1/PC2 sites in rat proenkephalin and six sites in rat proglucagon, our method correctly predicts five sites and one site, respectively. Almost all sites contain a dibasic motif at P2 and P1, but lack a basic residue on P4. Thus, assuming that our general data set is dominated by furin-type cleavage sites, we do not expect very high prediction performance on bona fide PC1/PC2 cleavage sites.

Discussion

Proteases are highly relevant both as biotechnological tools and as pharmaceutical targets. In both aspects, a deeper understanding of protease specificity and the ability to rapidly scan thousands of protein sequences for potential candidates are desirable.

One of the obstacles when mammalian cells are designed for overproduction of secreted bioactive proteins and peptides is the relatively low level of endogenous proprotein convertases (Seidah and Chretien, 1997). Recombinant furin has been used successfully to increase the yield of over-expressed secreted bioactive proteins in engineered eukaryotic cells (Seidah and Chretien, 1997; Himmelspach *et al*., 2000). Design of recombinant proteins containing highly efficient furin cleavage sites, while still retaining native sequence after cleavage, may be facilitated by the fast methods described in this paper.

Understanding substrate specificity may in particular be useful for designing specific PC inhibitors. PC inhibitors may be designed for treatment of diseases dependent on normal or aberrant PC function, such as anthrax, Ebola virus infections or cancer (Bergeron *et al*., 2000; Rockwell *et al*., 2002; Thomas, 2002). Being able to predict and rapidly screen large sets of synthetic cleavage substrates of certain proteases such as the PCs may substantially speed up the selection of highly efficient cleavable peptides. The sequence and deduced structure of these peptides may be used as lead structures for the development of peptidomimetic protease inhibitors.

In addition, the recently published structures on related PC enzymes will allow for a more focused search for lead structures for potential PC inhibitors (Henrich *et al*., 2003; Holyoak *et al*., 2003). A deeper understanding of the structural requirements of PC substrates may also be used to improve on the predictions of potential substrates, e.g. by penalizing side chains which are incompatible with the protease structure.

Ideally, one would like to train neural networks for each of the PC members. However, the amount of experimentally verified sites, where the bona fide PC is known, is still limited. This stems from the fact that it is much easier to identify the site of cleavage in a protein, e.g. by N-terminal sequencing, than to identify the nature of the physiologically active PC. Since we were able to extract only a limited number of verified cleavage sites for PCs other than furin, we decided to include all basic proprotein cleavage sites in our general approach. This situation is in many ways similar to our earlier experiences in characterizing phosphorylation sites at a time when few kinase-specific data were available (Blom *et al*., 1999).

The performance values obtained for the general type PC network compared with the furin-specific network confirmed our expectations. Although a basic cleavage site preference is observed for all PCs, slight differences in substrate specificity imply that a single method is not able to predict all PC sites with the same performance. In particular, we expect that the performance of the general network on PC1 and PC2 sites may be sub-optimal owing to an under-representation of these sites in the data set. A natural development from this will be to design PC1 and PC2 data sets, individual or combined, and train specific neural networks on these data.

Three other PCs, PC5, PACE4 and PC7, have substrate specificities which are similar to furin (Nakayama, 1997) and may be quite difficult to identify *in vitro* owing to cross-specificity and redundancy. Sites predicted by the furin network may therefore, in some cases, be physiological substrates of a furin-like PC.

The prediction of protease cleavage sites may also be used as functional fingerprints, for example, as an additional functional feature input to systems biology approaches attempting to classify the function of orphan proteins (Jensen *et al*., 2002). Also, in systematic screenings for novel secreted factors, the presence of a strong predicted processing site may be used as a search criterion.

Acknowledgements

We thank Kristoffer Rapacki for assistance with database extraction, Lars Juhl Jensen for profitable discussions and Neil Taylor for helpful advice. The work was supported by the National Danish Research Foundation and a grant from NsGene A/S.

References

Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd edn. MIT Press, Cambridge, MA.

Bauskin,A.R., Zhang,H.P., Fairlie,W.D., He,X.Y., Russell,P.K., Moore,A.G., Brown,D.A., Stanley,K.K. and Breit,S.N. (2000) *EMBO J.*, **19**, 2212–2220.

Bergeron,F., Leduc,R. and Day,R. (2000) *J. Mol. Endocrinol.*, **24**, 1–22.

Blom,N., Gammeltoft,S. and Brunak,S. (1999) *J. Mol. Biol.*, **294**, 1351–1362.

Bresnahan,P.A., Leduc,R., Thomas,L., Thorner,J., Gibson,H.L., Brake,A.J., Barr,P.J. and Thomas,G. (1990) *J. Cell Biol.*, **111**, 2851–2859.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol.*, **220**, 49–65.

Cameron,A., Apletalina,E. and Lindberg,I. (2002) In Dalbey,R.E. and Sigman,D.S. (eds), *The Enzymes*, 3rd edn. Academic Press, San Diego, CA, Vol. 22, pp. 291–332.

Chen,Y., Molloy,S.S., Thomas,L., Gambee,J., Bachinger,H.P., Ferguson,B., Zonana,J., Thomas,G. and Morris,N.P. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 7218–7223.

Duguay,S.J., Lai-Zhang,J. and Steiner,D.F. (1995) *J. Biol. Chem.*, **270**, 17566–17574.

Fuller,R.S., Brake,A.J. and Thorner,J. (1989) *Science*, **246**, 482–486.

Henrich,S., Cameron,A., Bourenkov,G.P., Kiefersauer,R., Huber,R., Lindberg,I., Bode,W. and Than,M.E. (2003) *Nat. Struct. Biol.*, **10**, 520–526.

Himmelspach,M., Pfleiderer,M., Fischer,B.E., Plaimauer,B., Antoine,G., Falkner,F.G., Dorner,F. and Schlokat,U. (2000) *Thromb. Res.*, **97**, 51–67.

Holyoak,T., Wilson,M.A., Fenn,T.D., Kettner,C.A., Petsko,G.A., Fuller,R.S. and Ringe,D. (2003) *Biochemistry*, **42**, 6709–6718.

Jensen,L.J. *et al*. (2002) *J. Mol. Biol.*, **319**, 1257–1265.

Junker,V.L., Apweiler,R. and Bairoch,A. (1999) *Bioinformatics*, **15**, 1066–1067.

Lehmann,M., Andre,F., Bellan,C., Remacle-Bonnet,M., Garrouste,F., Parat,F., Lissitsky,J.C., Marvaldi,J. and Pommier,G. (1998) *Endocrinology*, **139**, 3763–3771.

Matthews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.

Molloy,S.S. and Thomas,G. (2002) In Dalbey,R.E. and Sigman,D.S. (eds), *The Enzymes*, 3rd edn. Academic Press, San Diego, CA, Vol. 22, pp. 199–235.

Mori,K., Imamaki,A., Nagata,K., Yonetomi,Y., Kiyokage-Yoshimoto,R., Martin,T.J., Gillespie,M.T., Nagahama,M., Tsuji,A. and Matsuda,Y. (1999) *J. Biochem. (Tokyo)*, **125**, 627–633.

Nakayama,K. (1997) *Biochem. J.*, **327**, 625–635.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Rockwell,N.C., Krysan,D.J., Komiyama,T. and Fuller,R.S. (2002) *Chem. Rev.*, **102**, 4525–4548.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) In Rumelhart,D.E., McClelland,J.L. and the PDP Research Group (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA, Vol. I, pp. 318–362.

Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.

Seidah,N.G. and Chretien,M. (1997) *Curr. Opin. Biotechnol.*, **8**, 602–607.

Seidah,N.G. and Chretien,M. (1999) *Brain Res.*, **848**, 45–62.

Seidah,N.G., Day,R., Marcinkiewicz,M. and Chretien,M. (1998) *Ann. N. Y. Acad. Sci.*, **839**, 9–24.

Shinde,U. and Inouye,M. (2000) *Semin. Cell Dev. Biol.*, **11**, 35–44.

Steiner,D.F. (1998) *Curr. Opin. Chem. Biol.*, **2**, 31–39.

Steiner,D.F. (2002) In Dalbey,R.E. and Sigman,D.S. (eds), *The Enzymes*, 3rd edn. Academic Press, San Diego, CA, Vol. 22, pp. 163–198.

Thomas,G. (2002) *Nat. Rev. Mol. Cell Biol.*, **3**, 753–766.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.

van de Ven,W.J., Voorberg,J., Fontijn,R., Pannekoek,H., van den Ouweland,A.M., van Duijnhoven,H.L., Roebroek,A.J. and Siezen,R.J. (1990) *Mol. Biol. Rep.*, **14**, 265–275.

Volchkov,V.E., Volchkova,V.A., Stroher,U., Becker,S., Dolnik,O., Cieplik,M., Garten,W., Klenk,H.D. and Feldmann,H. (2000) *Virology*, **268**, 1–6.

Williams,Z. and Wassarman,P.M. (2001) *Biochemistry*, **40**, 929–937.

Received July 31, 2003; revised October 28, 2003; accepted October 28, 2003

Edited by Valerie Daggett