Bioinformatics, 31(8), 2015, 1204–1210
doi: 10.1093/bioinformatics/btu810
Advance Access Publication Date: 9 December 2014
Original Paper

OXFORD

Sequence analysis

State of the art prediction of HIV-1 protease cleavage sites

Thorsteinn Rögnvaldsson[1,*], Liwen You[1] and Daniel Garwicz[2]

[1]CAISR, School of Information Science, Computer and Electrical Engineering, Halmstad University, Halmstad, Sweden and [2]Division of Clinical Chemistry and Pharmacology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 30, 2014; revised on November 27, 2014; accepted on December 4, 2014

Abstract

Motivation: Understanding the substrate specificity of human immunodeficiency virus (HIV)-1 protease is important when designing effective HIV-1 protease inhibitors. Furthermore, characterizing and predicting the cleavage profile of HIV-1 protease is essential to generate and test hypotheses of how HIV-1 affects proteins of the human host. Currently available tools for predicting cleavage by HIV-1 protease can be improved.

Results: The linear support vector machine with orthogonal encoding is shown to be the best predictor for HIV-1 protease cleavage. It is considerably better than current publicly available predictor services. It is also found that schemes using physicochemical properties do not improve over the standard orthogonal encoding scheme. Some issues with the currently available data are discussed.

Availability and implementation: The datasets used, which are the most important part, are available at the UCI Machine Learning Repository. The tools used are all standard and easily available.

Contact: thorsteinn.rognvaldsson@hh.se

1 Introduction

Globally, an estimated 35 million people were living with human immunodeficiency virus (HIV) infection at the end of 2012, and roughly as many have died of HIV-related illnesses since the beginning of the epidemic (World Health Organization, 2014). HIV-1 protease, also known as HIV-1 retropepsin (EC 3.4.23.16), is an aspartyl protease belonging to the retroviral protease (retropepsin) family. The protease plays a crucial role in the life cycle of HIV, the causative agent of acquired immune deficiency syndrome (AIDS); it cleaves the HIV-1 polyproteins in multiple sites to create mature protein components of the virions, the infectious HIV particles (Sundquist and Kräusslich, 2012).

Because the formal discovery of HIV in the early 1980s and the ensuing characterization of HIV-1 protease, successful attempts have been made to create drugs that inhibit the protease (Hughes *et al.*, 2011). These slow down or even stop the progression of HIV infection to AIDS (however, the effectiveness of the treatment can decrease with time due to mutations in the virus). Today, combinations of drugs with different mechanisms of action are often used to achieve high efficacy against the virus but as low toxicity as possible to the patient. Although HIV therapy is one of the most successful pharmacotherapeutical achievements in the history of medicine it is not curative, but rather transforms a deadly infection into a chronic one, often with a prolonged asymptomatic phase if properly treated.

A reliable predictor of cleavage by HIV-1 protease can be used to aid in the identification of novel HIV-1 protease substrates in human host cells (cf. Devroe *et al.*, 2005) and aid in the understanding of the specificity and the development of even more tightly fitting and more potent HIV-1 protease inhibitors in the future, but hopefully with less severe side effects, such as metabolic syndrome and gastrointestinal symptoms. The work may also be of help in prediction and understanding of other viral proteases in the future. The HIV-1 protease specificity is considered to be both broad and specific; it cleaves a variety of sequences but also processes the HIV-1 gag and gag-pol polyproteins accurately (Darke *et al.*, 1988).

There are two different approaches to predicting cleavage by HIV-1 protease: molecular modeling and sequence analysis. It has been argued that the HIV-1 protease recognizes shape rather than a specific amino acid sequence (Prabu-Jeybalan *et al.*, 2002), which supports aiming for the molecular modeling approach. However, the method is cumbersome and no large scale study has been done on the accuracy of molecular modeling approaches so it is very unclear if the approach is, or will be, competitive with the sequence-based approach. This article demonstrates the current state-of-the-art prediction, which uses the sequence-based approach.

Several bioinformatics researchers have attacked this problem during the last 20 years, using a diversity of methods (for the most recent review see Rögnvaldsson *et al.*, 2007). It was early on claimed that the problem required non-linear methods. However, 10 years ago it was demonstrated that the relatively few experimental data (362 octamers at the time) did not support a non-linear model (Rögnvaldsson and You, 2004). Five years later, when more experimental data were available, linear methods [linear support vector machines (LSVMs)] still performed better than non-linear ones (Rögnvaldsson *et al.*, 2009). This was when the methods were evaluated through out-of-sample testing on a large dataset from human proteins (Schilling and Overall, 2008). It was therefore speculated (see Rögnvaldsson *et al.*, 2007 for the discussion) that linearity could be a characteristic for the HIV-1 protease cleavage problem. (Note that linear and non-linear relate to when the standard orthogonal encoding is used.)

A number of papers have been published on the subject over the last 5 years. The most common theme in them is to introduce new features plus a feature selection scheme and show that this yields slightly better prediction accuracy when evaluated with cross-validation. Surprisingly few studies have used the experimental data from the article by Schilling and Overall (2008), which is more than double the size of any other available dataset.

Oğul (2009) used variable context Markov chains (Bejerano and Yona, 2001) to construct a generative model for the HIV-1 cleavage specificity. He reported the highest ever prediction accuracies with this method, evaluated with cross-validation.

Nanni and Lumini (2009) created a number of different features and fused classifiers, among others using genetic programming (GP). Their proposed classifiers performed better than a non-optimized LSVM with standard encoding, when evaluated using cross-validation. Their software is available.

Jaeger and Chen (2010) suggested new biophysical features and fused several classifiers [neural networks, support vector machines (SVMs) and decision trees]. They reported that they often achieved better performance than just using a single classifier from this, when evaluating with cross-validation.

Kim *et al.* (2010) suggested a feature selection method where a multilayer perceptron was trained and then used to compute the effect of the different inputs so that the best inputs were selected. They reported that this gave a much smaller feature set and better prediction accuracy. They tested this on a small dataset with cross-validation.

Li *et al.* (2010) mapped the amino acid sequences to a local kernel space and reduced the dimensionality together with a LSVM classifier. They reported that this was better than other methods when evaluated using cross-validation.

Newell (2011) studied the specificity using cascade detection on two larger datasets and concluded that favorable cooperativity between sites is weak. Newell used the larger dataset from Schilling and Overall in his study.

Gök and Özcerit (2012) studied several encoding schemes and suggested the OETMAP encoding scheme, based on amino acid features, together with a linear classifier. This encoding improved the prediction performance significantly compared with standard amino acid encodings when evaluated on two larger datasets using cross-validation. Gök and Özcerit used the larger dataset from Schilling and Overall. Their encoding schemes are available on a web server (http://yufes.yalova.edu.tr/).

Song *et al.* (2012) presented a web server for predicting cleavage by many different proteases, using support vector regression together with many different features. Features were encoded with bi-profile Bayesian feature extraction and selected using a Gini score. They used the larger dataset from Schilling and Overall plus other published data on cleavage of full proteins. Their methods are available (https://prosper.erc.monash.edu.au/).

Niu *et al.* (2013) used a correlation-based feature subset selection method combined with genetic algorithms to search for the best subset in a large set of features. This gave better performance than the standard methods when evaluated with cross-validation.

Öztürk *et al.* (2013) used a sequence representation and introduced a feature selection method (that removed features). They reported improved prediction results with this when tested with cross-validation on a small dataset with only cleaved octamers. Their software is designed to work with the Waikato Environment for Knowledge Analysis (WEKA) (Hall *et al.*, 2009) and is available by email.

In summary, many have claimed improvements over the LSVM using standard orthogonal encoding. However, few have taken the effort to check if their new features, feature selection method or model combination method really does better on out-of-sample data or if they are better than improvements suggested by others. This is a significant weakness since there is a real risk of being overly optimistic about one's own algorithm if one has access to the data that it is tested on.

There are two motivations for mining HIV-1 protease cleavage data: one is to describe the available experimental data, e.g. with sequence motifs or cleavage rules, the other is to design a method for predicting new cleavage sites. In the latter case, which is by far the most common motivation, the test data must be different from the data used to train the algorithm. A correct evaluation of methods must be done on test data that have not in any way been involved in the training of the algorithms. This is typically not done.

2 Methods

The HIV-1 cleavage problem is described in detail in (Rögnvaldsson *et al.*, 2007) together with discussions on different encoding schemes. Only a concise description is given here. The classification task is to tell whether a given octamer (sequence of eight amino acids) will be cleaved or not between the fourth and the fifth position. The octamer is represented using an orthogonal encoding where each amino acid is represented by a 20-bit vector with 19 bits set to zero and one bit set to one (other encodings have been suggested, see later). This maps each octamer to an 8 by 20 binary matrix that is transformed into a 160-dimensional vector. However, the dimensionality of the problem is 152 since there are eight linear constraints (each position must be occupied by one amino acid).

The inputs were centered so that the zero bits were set to -1 before presenting them to the classifier. The outputs were similarly coded as $\{-1,1\}$: minus one for uncleaved octamers and plus one for cleaved octamers. The OETMAP encoding (Gök and Özcerit, 2012) and the GP1 encoding (Nanni and Lumini, 2009) were also tried, in addition to the standard orthogonal encoding. The OETMAP was used by calling the web server mentioned earlier. The GP1 encoding was created by using the scripts provided by the authors, the inputs

were centered to {−1,+1} before presented to the classifier. The software and feature selection from Öztürk *et al.* (2013) was tried but crashed and is therefore not included in the comparison.

The libsvm 3.18 library (Chang and Lin, 2011) was used, with the multi-class C-SVC method, to train SVMs and called from within the MATLAB (MATLAB, 2013) environment. Both linear and radial basis kernels were tried. The hyperparameters C and γ (the latter for the non-LSVM) were optimized by search and 10-fold cross-validation. For the LSVM, the C parameter was varied over the set $\log(C) = \{-5, -4.75, -4.5, -4.25, \ldots, 5\}$. For the non-LSVM the C parameter was varied over the set $\log(C) = \{0, 0.25, 0.5, 0.75, \ldots, 7\}$ and the γ parameter was varied over the set $\log(\gamma) = \{-5, -4.75, -4.5, -4.25, \ldots, 0\}$. The final values were selected by looking at the mean and median area under the receiver operator characteristic (ROC) curve (this area is called AUC). The goal was to select them to maximize the cross-validation AUC. Sometimes the median and mean were not maximal for the same value(s) of C (and γ) and then a subjective choice was made. One final SVM model was trained with the full training dataset and the optimal hyperparameter values.

Two predictors on the web were used as comparison: HIVcleave (Shen and Chou, 2008), at http://www.csbio.sjtu.edu.cn/bioinf/HIV/#, and PROSPER (Song *et al.*, 2012), at https://prosper.erc.monash.edu.au/. The HIVcleave predictions were done by concatenating the peptides and octamers in the datasets and submitting them in suitable chunks to the web server. This does not present any problem for the HIVcleave predictor since it cuts up the sequence into octamers. PROSPER is designed to predict cleavage in full proteins. The comparison with PROSPER was therefore done on full protein sequences, which were submitted to the PROSPER server.

Algorithm performances were, when possible, compared using the full ROC curve and the test with correlated data described in DeLong *et al.* (1988).

3 Results

3.1 The HIV-1 PR datasets

Four different datasets were used in our experiments: one with 746 octamers; one with 1625 octamers; one with 3272 octamers and one with 947 octamers. The 746 dataset was presented in 2005 (You *et al.*, 2005). The 1625 dataset was presented in 2007 (Kontijevskis *et al.*, 2007). The third and largest dataset was collected from the work of Schilling and Overall (2008) on human proteins. It was presented in 2009 (Rögnvaldsson *et al.*, 2009). The details of how the first three datasets were collected are described in the references.

The three older datasets needed some corrections. The two datasets with 746 and 1625 octamers were corrected according to the comment in Rögnvaldsson *et al.* (2009), i.e. the octamer SQNYAIVQ was labeled as cleaved. Furthermore, we discovered some errors in the supplementary material of Schilling and Overall (2008), from which the largest dataset had been created. These are mostly of the form that a character has been lost compared with the sequence in the databases. The sequences were corrected so that they matched the database sequences. The changes are listed in Table 1. This produced a slightly larger dataset than the one used in Rögnvaldsson *et al.* (2009). The dataset is denoted as the Schilling dataset.

The fourth dataset was collected specifically for this study from four publications: Impens *et al.* (2012); Gerenčer and Burek (2004); Álvarez *et al.* (2006) and Nie *et al.* (2007). The latter contributed with one cleaved octamer: PKVF*FIQA (the asterisk marks the cleavage site). Furthermore, there are two uncleaved peptides reported in Nie *et al.* (2007): RGYCLIINNHNFAK and KGIIYGTDGQEAPIYELTSQFTGLK. Uncleaved octamers were created from them by running an eight position long window over, yielding {RGYCLIIN, GYCLIINN, ..., INNHNFAK} and {KGIIYGTD, GIIYGTDG, ..., TSQFTGLK}. The same process was applied to all cleavage sites and non-cleaved peptides listed in Impens *et al.* (2012). Five octamers (one cleaved and four non-cleaved) were added from Gerenčer and Burek (2004) by taking the identified cleavage site, shifting two positions toward either side of the cleavage site and labeling the resulting octamers as non-cleaved. In a similar fashion, ten octamers (two cleaved and eight non-cleaved) were added from the publication of Álvarez *et al.* (2006). This produced a total of 947 unique octamers, most of which are not in the other three datasets. We denote this the the Impens dataset.

Table 1. Corrections done to Supplementary Tables S19 and S20 (Schilling and Overall, 2008). The asterisk (*) marks the HIV-1 PR cleavage site

Table	Page	Original sequence	Corrected sequence
19	191	SGETEDTFA*	SGETEDTFIA*
19	193	FRPDNFF*	FRPDNFVF*
19	193	EENLCPEL*	EENLDCPEL*
19	193	QSPLLQYFG*	QASPLLQYFG*
19	196	LASQPVDGF*	LASQPGVDGF*
19	196	GETEDTFAD*	GETEDTFMAD*
19	197	ENDID*	YCID*
19	198	LAAQDPEVM*	LAAMQDPEVM*
19	199	NSSF*	NSSYF*
19	199	NMSDDDGWF*	NMSDDDGWQF*
19	199	LLEGNDIEL*	LILEGNDIEL*
19	200	ETEDTFADL*	ETEDTFIADL*
19	200	VETGLKPGM*	VETGVLKPGM*
20	202	ELASQPVDG*	ELASQPGVDG*
20	203	FRPDNFF*	FRPDNFVF*
20	204	VDSLEN*	VDSLLEN*
20	204	NESTPSEE*	NESTPPSEE*
20	205	LASQPVDGF*	LASQPGVDGF*
20	205	GETEDTFAD*	GETEDTFMAD*
20	205	ELFQFG*	QPSFLG*
20	207	NSSF*	NSSYF*
20	208	LSRPQDLEG*	LSRPQDALEG*

Some characteristics of the four datasets are summarized in Table 2. There are a few interesting observations. First, three of the datasets can be dichotomized by a linear classifier when the standard orthogonal encoding is used. Second, the rank of the data covariance matrix, which is the dimension of the subspace occupied by the data, is equal to the full 152 dimensions for three of the datasets, but the fourth has less variation. Third, using the OETMAP encoding (Gök and Özcerit, 2012) does not change the data dimension, which indicates that the OETMAP encoding (which is binary and 240-dimensional) does not add information compared with the standard orthogonal encoding.

Some of the datasets overlap and there are some conflicts as listed in Table 3. The 746 and the 1625 datasets overlap a lot since they were collected from almost the same sources. There are seven conflicts between the 746 and the 1625 dataset. Two of the conflicts (AEAMSQVT and FRSGVETT) are between inferred cleavage sites (Poorman *et al.*, 1991) and experimental data (Tözsér *et al.*, 1991). Three conflicts (GRINVALV, SGVFSVNG and SGVYQLSA) are different interpretations of weak cleavage (Beck *et al.*, 2001). Two conflicts (AAAMSSAI and ARVLAQAM) originate from conflicting

Table 2. Characteristics of the four HIV-1 PR datasets. The 'Octamers' column lists the total number of octamers and the 'C' and 'NC' columns list the number of cleaved and non-cleaved octamers, respectively. The 'Linear' column indicates if the dataset is separable with a linear classifier or not. The 'Rank OR' and 'Rank OET' columns show the ranks of the data covariance matrices, when using the orthogonal or the OETMAP encodings

Dataset	Octamers	C	NC	Linear	Rank OR	Rank OET
746	746	401	345	Yes	152	152
1625	1625	374	1251	Yes	152	152
Schilling	3272	434	2838	No	152	152
Impens	947	149	798	Yes	147	147

Table 3. Overlaps between the four HIV-1 PR datasets. The column labeled 'Joint' lists the number of octamers that are common between the datasets. The column 'Conflicts' lists any conflicts between the datasets

Datasets	Joint	Conflicts	Comment
746 and 1625	659	AAAMSSAI	NC in 746, C in 1625
		AEAMSQVT	C in 746, NC in 1625
		ARVLAQAM	NC in 746, C in 1625
		FRSGVETT	C in 746, NC in 1625
		GRINVALV	NC in 746, C in 1625
		SGVFSVNG	NC in 746, C in 1625
		SGVYQLSA	NC in 746, C in 1625
746 and Schilling	0		
746 and Impens	0		
1625 and Schilling	20	EENFAVEA	NC in 1625, C in Schilling
		QEEMLQRE	NC in 1625, C in Schilling
1625 and Impens	0		
Schilling and Impens	71	VEEGIVLG	C in Schilling, NC in Impens

experimental results (Boross *et al.*, 1999; Cameron *et al.*, 1992; Kádas *et al.*, 2004; Ridky *et al.*, 1998).

The 1625 and Schilling data share 20 octamers, which all come from vimentin (VIME_HUMAN). There are two conflicts: Schilling and Overall (2008) report two cleavage sites in places where Shoeman *et al.* (1990) did not find any cleavage.

The Schilling and Impens datasets overlap a bit. There is one conflict, VEEGIVLG, which is a cleavage site according to Schilling and Overall (2008) but not according to Impens *et al.* (2012). The octamer comes from a heat shock protein (CH60_HUMAN).

3.2 Orthogonal encoding and LSVM

A LSVM was trained on each one of the HIV-1 PR datasets and tested on the other HIV-1 PR datasets (the training procedure is described in the Methods section). The results are shown in Table 4.

Table 4 shows that the cross-validation result (underlined) on the training data is better than the out-of-sample result when the same data is used as test data, except if there is a large overlap between the training and test sets. Thus, the cross-validation result is a poor indicator of out-of-sample test performance. Another observation is that the two datasets derived from human proteins, the Schilling and Impens data, seem to be more similar to each other than to the 746 and 1625 datasets, which contain mostly octamers

Table 4. AUC values when using LSVM and orthogonal coding. The rows show the training data and the columns the test data. The overlapping sequences (see Table 3) were not removed. The underlined numbers on the diagonal (i.e. when training and test data are the same) are the cross-validation AUC values

Training\test	746	1625	Schilling	Impens
746	0.980	0.982	0.870	0.833
1625	0.981	0.987	0.855	0.811
Schilling	0.933	0.935	0.969	0.911
Impens	0.902	0.894	0.929	0.932

generated by varying single amino acids in cleaved sequences. The 746 and 1625 datasets have large overlaps. Table 4 also shows that it is of great importance what data one uses, both for training and test. The 1625 dataset seems to be the worst to use for training if one desires to model cleavage in human proteins.

Figure 1 shows the ROC curves for the case when the LSVM models are tested on the Schilling data and trained on any of the other three datasets. The differences between the three curves are all significant with P-values < 0.05. The smallest difference is between the LSVM trained with the 746 data and the LSVM trained with the 1625 data. The P-value for this difference is 0.02.

The test datasets were kept the same for all training sets in Table 4 to be able to compute the P-value for the differences when the test data are correlated (DeLong *et al.*, 1988). However, there are overlaps between the dataset (see Table 3), some of which are in conflict, and this can affect the results. The results when the overlapping sequences have been removed from the test data are shown in Table 5 for reference. Clearly, the overlaps have little or no effect on the results, except for the 746 and 1625 data that share many octamers.

3.3 LSVM and Radial Basis Function SVM (RBFSVM)

The Schilling dataset is the only dataset that is not linearly separable when using the orthogonal encoding. A non-LSVM with radial basis kernel was therefore trained on this data and the test results are listed in Table 6, compared with the results with LSVM. There are no significant differences between the prediction results achieved with non-LSVM and LSVM (except for the cross-validation result on the Schilling data itself). This finding is also confirmed in Section 3.7.

3.4 OETMAP encoding

Gök and Özcerit (2012) suggested the OETMAP encoding, after having tried several other encodings, and claimed that it gave slightly better results. LSVMs (and non-LSVMs) were trained using this OETMAP encoding instead of the orthogonal encoding. The results for LSVM are listed in Table 7, which should be compared with Table 4. There are few significant differences between results with the orthogonal encoding and the OETMAP encoding, and none of them are in favor of OETMAP (except the cross-validation result when the training and test data are the same).

3.5 GP encoding

Nanni and Lumini (2009) try several encodings and also combine models with different features. Two of their best encodings for single LSVM models are denoted GP1 and GP2, which were found using GP. They claim that these new features are better than the standard orthogonal encoding but the differences are small and most likely

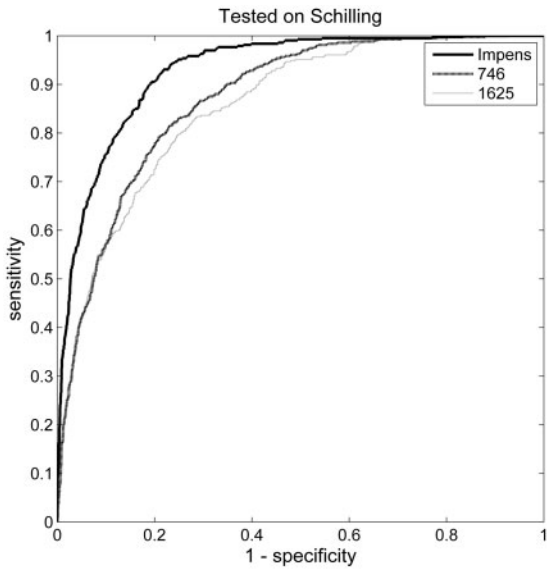

Fig. 1. ROC curves for LSVM using orthogonal coding, trained on the 746, 1625 or the Impens data and tested on the Schilling data

Table 5. AUC values when using LSVM and orthogonal encoding but where the sequences common between the training and test data have been removed from the test data. The numbers should be compared with those in Table 4. The cross-validation results are left out for clarity

Training\Test	746	1625	Schilling	Impens
746	–	0.955	0.870	0.833
1625	0.906	–	0.854	0.811
Schilling	0.933	0.936	–	0.906
Impens	0.902	0.894	0.927	–

Table 6. AUC values for LSVM and RBFSVM trained on the Schilling data and tested on the other data. The rows show the training data and the columns the test data. The underlined numbers for 'Schilling', when training and test data are the same, are the cross-validation AUC values

Training\test	746	1625	Schilling	Impens
Schilling (LSVM)	0.933	0.935	0.982	0.911
Schilling (RBFSVM)	0.937	0.936	0.968	0.908

insignificant. For example, the GP1 and GP2 features are identical when only one LSVM model is used (this is at least how their code works). However, the performance difference between GP1 and GP2 with one LSVM are similar to those between GP1/GP2 and the orthogonal encoding (Nanni and Lumini, 2009), i.e. the difference is probably due to random fluctuation.

Table 8 lists the AUC results when the GP1 features are used as encoding. The results are significantly worse than the orthogonal encoding in five cases and significantly better in one case. The significance level is set to 0.05 and it is not unlikely that there are occasional 'significant' results when doing several comparisons as in this case. Hence, there is no evidence that the features selected with GP are better than the standard orthogonal encoding, rather the opposite.

Table 7. AUC values when using LSVM and OETMAP coding. Compare with Table 4. Results that are significantly different from the results with orthogonal encoding, i.e. with P-value < 0.05, are marked with an asterisk (does not apply to the cross-validation results). The underlined numbers on the diagonal (i.e. when training and test data are the same) are the cross-validation AUC values

Training\test	746	1625	Schilling	Impens
746 (OETMAP)	0.980	0.982	0.874	0.827
1625 (OETMAP)	0.981	0.988	0.836*	0.793*
Schilling (OETMAP)	0.930	0.932	0.970	0.895*
Impens (OETMAP)	0.902	0.886	0.905*	0.934

Table 8. AUC values when using LSVM and GP1 coding. Compare with Table 4. Results that are significantly different from the results with orthogonal encoding, i.e. with P-value < 0.05, are marked with an asterisk (does not apply to the cross-validation results). The underlined numbers on the diagonal (i.e. when training and test data are the same) are the cross-validation AUC values

Training\test	746	1625	Schilling	Impens
746 (GP1)	0.981	0.979*	0.857*	0.840
1625 (GP1)	0.982	0.988	0.845*	0.830*
Schilling (GP1)	0.936	0.938	0.966	0.905*
Impens (GP1)	0.913	0.902	0.917*	0.921

It is worth noting that Nanni and Lumini (2009) used LSVMs 'without any kind of parameter optimization' (sic), i.e. they used the same default C value all the time, regardless of input encoding.

3.6 Comparison with HIVcleave

The four HIV-1 PR datasets were submitted to the cleavage server HIVcleave (Shen and Chou, 2008). The AUC values for the HIVcleave predictions were: 0.929 (on the 746 data); 0.891 (on the 1625 data); 0.753 (on the Schilling data) and 0.687 (on the Impens data). These results are not competitive with the best prediction results for any of the datasets. The 746 dataset overlaps significantly with the data used to construct the HIVcleave server.

3.7 Combining datasets

The factor that influences the prediction accuracy the most is the data used for training. The four HIV-1 PR datasets were therefore combined in the following seven ways:

1. The 746 data and the 1625 data were joined. The conflicts AEAMSQVT and FRSGVETT were set to cleaved. The conflicts GRINVALV, SGVFSVNG, SGVYQLSA, AAAMSSAI and ARVLAQAM were removed as 'uncertain'. This gave a dataset with 1707 unique octamers: 420 cleaved and 1287 non-cleaved.
2. The 1707 dataset (earlier) was joined with the Schilling dataset. The two conflicts EENFAVEA and QEEMLQRE were set to cleaved. This produced a dataset with 4959 unique octamers: 854 cleaved and 4105 non-cleaved.
3. The 1707 dataset (earlier) was joined with the Impens dataset. This gave 2654 unique octamers: 569 cleaved and 2085 non-cleaved.
4. The 746 data were joined with the Impens data. This produced a dataset with 1693 unique octamers: 551 cleaved and 1142 non-cleaved.
5. The 746 data were joined with the Schilling data. This gave 4018 unique octamers: 836 cleaved and 3182 non-cleaved.

Table 9. AUC test values when the combined datasets are used for training. The second column shows whether a LSVM or a non-LSVM was used. The AUC values should be compared with those in Table 4

Training\test		Schilling	Impens
746 + 1625	(LSVM)	0.865	0.817
746 + 1625	(RBFSVM)	0.866	0.819
746 + Schilling	(LSVM)	–	0.920
746 + Schilling	(RBFSVM)	–	0.911
1625 + Schilling	(LSVM)	–	0.902
1625 + Schilling	(RBFSVM)	–	0.898
746 + 1625 + Schilling	(LSVM)	–	0.898
746 + 1625 + Schilling	(RBFSVM)	–	0.900
746 + Impens	(LSVM)	0.938	–
746 + Impens	(RBFSVM)	0.934	–
1625 + Impens	(LSVM)	0.926	–
1625 + Impens	(RBFSVM)	0.934	–
746 + 1625 + Impens	(LSVM)	0.931	–
746 + 1625 + Impens	(RBFSVM)	0.934	–

6. The 1625 data were joined with the Impens data. This gave 2572 unique octamers: 524 cleaved and 2048 non-cleaved.
7. The 1625 data were joined with the Schilling data. The two conflicts EENFAVEA and QEEMLQRE were set to cleaved. This produced a dataset with 4877 unique octamers: 809 cleaved and 4068 non-cleaved.

The test results when training LSVMs and non-LSVMs with these datasets are listed in Table 9, which should be compared with Table 4. The results show two things. First, they confirm that there is no need to use a non-LSVM instead of a LSVM. Secondly, adding the 746 data to the Schilling or Impens datasets tends to improve the out-of-sample prediction, whereas adding the 1625 data tends to deteriorate the out-of-sample prediction.

The problem with the 1625 dataset is visible also in Table 4 and Figure 1. The main difference between the 746 and the 1625 data are the non-cleaved octamers so this is probably where the problem lies.

3.8 Cleaving human proteins

The cleavage server PROSPER is designed to handle full-length proteins. The majority of the human proteins in Impens *et al.* (2012) are not in the Schilling dataset and are not referenced in the PROSPER publication, these proteins could therefore serve as out-of-sample test for PROSPER. The proteins listed in Impens *et al.* (2012) were submitted, in FASTA format, to the PROSPER cleavage server (in September 2014). The same proteins were also cut up into octamers and submitted to the LSVM predictors described earlier: one trained with the 746 + Schilling data and another trained with the 746 + 1625 + Schilling data. The predictions were then compared.

The PROSPER server is restrictive with predicting cleavage. For a fair comparison were the LSVM predictors thresholded to yield a lowest specificity that equaled the PROSPER lowest specificity on the Impens dataset, i.e. they should have specificities above 0.89.

The results from predicting cleavage in the 148 cleaved peptides listed in Impens *et al.* (2012) are shown in Table 10. The LSVM predictor trained with the 746 + Schilling data is better than the other two. This model predicts the peptides to be cleaved in 29% of the cases and gets the cleavage position perfectly right in almost 20% of the cases; it is also less off than the other methods in those cases when the cleavage position is not perfectly right. The PROSPER server is second best. Combining the 1625 data with the other data worsens the performance.

Table 10. Results for predicting cleavage in the proteins listed in Impens *et al.* (2012). The first column shows which model that was used. The second column, labeled 'Perfectly corr.', shows how often the peptides were predicted to be cleaved and the predicted cleavage site matched the experimental location exactly. The third column, labeled 'Ave. pos. off', shows how many positions off, on average, the predicted cleavage site was when a peptide was predicted to be cleaved but the location did not match exactly. The fourth and rightmost column shows how often no cleavage was predicted in the peptides

Predictor	Perfectly corr.	Ave. pos. off	No cleavage
PROSPER	13.5%	4.1	75.0%
LSVM 746 + Schilling	19.6%	2.7	70.9%
LSVM 746 + 1625 + Schilling	12.2%	2.8	84.5%

4 Conclusions and discussion

The experiments show that the LSVM with standard orthogonal encoding is the hitherto best model for predicting cleavage by HIV-1 protease. They also show that the training data is the most important factor for the performance, and that there seem to be some problems in the 1625 data published by Kontijevskis *et al.* (2007). We also suggest a few corrections to the data published by Schilling and Overall (2008).

The best training data were those that had been derived from human proteins, e.g. the data collected by Schilling and Overall (2008) and Impens *et al.* (2012). This is not surprising; the earlier data were to a large extent collected by point mutations in short peptides, which is a biased and limited sampling of the sample space, as discussed in Rögnvaldsson *et al.* (2007). Future predictors of HIV-1 protease cleavage should build on data collected like the Schilling and the Impens data.

No evidence was found that more advanced feature encoding and selection schemes yield better out-of-sample results than using the standard orthogonal encoding without feature selection. Unfortunately, many published results on this topic are difficult to reproduce. Software was downloaded or requested in the cases this was offered but it turned out to be problematic to reproduce the results also in some of those cases, for various reasons. In two cases, the OETMAP and the GP1 encodings, the methods were available for straightforward use. In those cases, using the new feature encoding was not better than the standard orthogonal encoding. The facts that the feature selection papers build on cross-validation, which is shown to be a poor indicator of out-of-sample performance in this case, and the observed differences are small, also supports the conclusion that standard orthogonal encoding is at least as good as other schemes.

There may be significant effects from using different feature encodings and feature selection methods when evaluated on a particular dataset. However, these effects are then small and drown in comparison with the effect of using a different dataset that has been collected in a different way.

The data used in this study are available at the UCI machine learning data repository (http://archive.ics.uci.edu/ml/) and researchers are recommended to use this data when doing proper out-of-sample testing and method development for this problem.

The prediction from the LSVM with standard orthogonal encoding was compared with the best currently available cleavage servers: HIVcleave and PROSPER. The LSVM outperformed HIVcleave by a wide margin. The comparison with PROSPER was done using only full-length substrates, since this is what it is designed for, but also in this case was the LSVM better.

The finding that the LSVM is the best predictor is in line with the discussion in Rögnvaldsson *et al.* (2007). The character of a protease cleavage problem agrees well with a linear classifier with orthogonal encoding. This could also mean that other viral and non-viral protease cleavage problems are linear (or close to linear) when orthogonal encoding is used.

Conflict of interest: none declared.

References

Álvarez,E. *et al.* (2006) HIV protease cleaves poly(A)-binding protein. *Biochem J.*, **396**, 219–226.

Beck,Z.Q. *et al.* (2001) Molecular basis for the relative substrate specificity of human immunodeficiency virus type 1 and feline immunodeficiency virus proteases. *J. Virol.*, **75**, 9458–9469.

Bejerano,G. and Yona,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **17**, 23–43.

Boross,P. *et al.* (1999) Effect of substrate residues on the P20 preference of retroviral proteinases. *Eur. J. Biochem.*, **264**, 921–929.

Cameron,C.E. *et al.* (1992) Mechanism of inhibition of the retroviral protease by a Rous sarcoma virus peptide substrate representing the cleavage site between the *gag* p2 and p10 proteins. *J. Biol. Chem.*, **267**, 23735–23741.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM TIST*, **2**, 27:1–27:27.

Darke,P.L. *et al.* (1988) HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins. *Biochem. Bioph. Res. Co.*, **156**, 297–303.

DeLong,E.R. *et al.* (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Devroe,E. *et al.* (2005) HIV-1 incorporates and proteolytically processes human NDR1 and NDR2 serine-threonine kinases. *Virology*, **331**, 181–189.

Gerenčer,M. and Burek,V. (2004) Identification of HIV-1 protease cleavage site in human C1-inhibitor. *Virus Res.*, **105**, 97–100.

Gök,M. and Özcerit,A.T. (2012) A new feature encoding scheme for HIV-1 protease cleavage site prediction. *Neural Comput. Appl.*, **22**, 1757–1761.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18.

Hughes,P.J. *et al.* (2011) Protease inhibitors for patients with HIV-1 infection: a comparative overview. *P&T*, **36**, 332–345.

Impens,F. *et al.* (2012) A catalogue of putative HIV-1 protease host cell substrates. *Biol. Chem.*, **393**, 915–931.

Jaeger,S. and Chen,S.-S. (2010) Information fusion for biological prediction. *J. Data Sci.*, **8**, 269–288.

Kádas,J. *et al.* (2004) Narrow substrate specificity and sensitivity toward ligand-binding site mutations of human T-cell Leukemia virus type 1 protease. *J. Biol. Chem.*, **279**, 27148–27157.

Kim,G. *et al.* (2010) An MLP-based feature subset selection for HIV-1 protease cleavage site analysis. *Artif. Intell. Med.*, **48**, 83–89.

Kontijevskis,A. *et al.* (2007) Computational proteomics analysis of HIV-1 protease interactome. *Proteins*, **68**, 305–312.

Li,X. *et al.* (2010) Predicting human immunodeficiency virus protease cleavage sites in nonlinear projection space. *Mol. Cell. Biochem.*, **339**, 127–133.

MATLAB. (2013) *MATLAB Release 2013b*. The MathWorks Inc., Natick, MA.

Nanni,L. and Lumini,A. (2009) Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins. *Amino Acids*, **36**, 409–416.

Newell,N.E. (2011) Cascade detection for the extraction of localized sequence features; specificity results for HIV-1 protease and structure–function results for the Schellman loop. *Bioinformatics*, **27**, 3415–3422.

Nie,Z. *et al.* (2007) Human immunodeficiency virus type 1 protease cleaves procaspase 8 in vivo. *J. Virol.*, **81**, 6947–6956.

Niu,B. *et al.* (2013) HIV-1 protease cleavage site prediction based on two-stage feature selection method. *Protein Pept. Lett.*, **20**, 290–298.

Oğul,H. (2009) Variable context Markov chains for HIV protease cleavage site prediction. *Biosystems*, **96**, 246–250.

Öztürk,O. *et al.* (2013) A consistency-based feature selection method allied with linear SVMs for HIV-1 protease cleavage site prediction. *PLoS One*, **8**, e63145.

Poorman,R.A. *et al.* (1991) A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. *J. Biol. Chem.*, **266**, 14554–14561.

Prabu-Jeybalan,M. *et al.* (2002) Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure*, **10**, 369–381.

Ridky,T.W. *et al.* (1998) Drug-resistant HIV-1 proteases identify enzyme residues important for substrate selection and catalytic rate. *Biochemistry*, **37**, 13835–13845.

Rögnvaldsson,T. *et al.* (2007) Bioinformatic approaches for modeling the substrate specificity of HIV-1 protease: an overview. *Expert Rev. Mol. Diagn.*, **7**, 435–451.

Rögnvaldsson,T. *et al.* (2009) How to find simple and accurate rules for viral protease cleavage specificities. *BMC Bioinformatics*, **10**, 149.

Rögnvaldsson,T. and You,L. (2004) Why neural networks should not be used for HIV-1 protease cleavage site prediction. *Bioinformatics*, **20**, 1702–1709.

Schilling,C. and Overall,C.M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat. Biotechnol.*, **26**, 685–694.

Shen,H.-B. and Chou,K.-C. (2008) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Anal. Biochem.*, **375**, 388–390.

Shoeman,R.L. *et al.* (1990) Human immunodeficiency virus type 1 protease cleaves the intermediate filament proteins vimentin, desmin, and glial fibrillary acidic protein. *Proc. Natl. Acad. Sci. USA.*, **87**, 6336–6340.

Song,J. *et al.* (2012) PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One,* **7**, e50300.

Sundquist,W.I. and Kräusslich,H-G. (2012) HIV-1 assembly, budding and maturation. *Cold Spring Harb. Perspect. Med.*, **2**, a006924.

Tözsér,J. *et al.* (1991) Comparison of the HIV-1 and HIV-2 proteinases using oligopeptide substrates representing cleavage sites in Gag and Gag-Pol polyproteins. *FEBS Lett.*, **281**, 77–80.

World Health Organization (2014) *World Health Statistics 2014*. World Health Organization.

You,L. *et al.* (2005) Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. *J. Virol.*, **79**, 12477–12486.