DETECTION OF CLEAVAGE SITES FOR HIV-1 PROTEASE IN NATIVE PROTEINS

Liwen You*
Computational Biology and Biological Physics Group
Department of Theoretical Physics, Lund University
Sölvegatan 14A, SE-22362, Lund, Sweden
**Email: liwen@thep.lu.se*

Intelligent Systems Lab
School of Information Science, Computer and Electrical Engineering, Halmstad University
Box 823, SE-30118, Halmstad, Sweden

Predicting novel cleavage sites for HIV-1 protease in non-viral proteins is a difficult task because of the scarcity of previous cleavage data on proteins in a native state. We introduce a three-level hierarchical classifier which combines information from experimentally verified short oligopeptides, secondary structure and solvent accessibility information from prediction servers to predict potential cleavage sites in non-viral proteins. The best classifier using secondary structure information on the second level classification of the hierarchical classifier is the one using logistic regression. By using this level of classification, the false positive ratio was reduced by more than half compared to the first level classifier using only the oligopeptide cleavage information. The method can be applied on other protease specificity problems too, to combine information from oligopeptides and structure from native proteins.

1. INTRODUCTION

Within the HIV-1 virion genome, *gag* and *pol* are two main genes. It is known that the *gag* gene encodes four separate proteins which form the building blocks for the viral core (i.e. matrix protein, capsid protein, and nucleocapsid protein) and the *pol* gene encodes four replication related proteins (i.e. protease, reverse transcriptase and integrase). Translation of *gag* and *gag/pol* transcript results in Gag and GagPol polyproteins. During the HIV-1 virion maturation process, HIV-1 protease cleaves viral Gag and GagPol polyproteins into structural and other replication proteins and make it possible to assemble into an infectious virion. Therefore, the cleavage of the polyproteins by HIV-1 protease plays an important role in the final stage of the HIV virion maturation process. Efficiently hindering the cleavage process is one way of blocking the viral life cycle. HIV-1 protease inhibitors are therefore part of the therapy arsenal against HIV/AIDS today. Efficiently cleaved substrates are excellent templates for the synthesis of tightly binding chemically modified inhibitors [1]. However, the difficulty is that the protease cleaves at different sites with little or no sequence similarities. In the last two decades, several studies, including wet-lab experiments on HIV-1 protease cleavage of oligopeptides, have been performed to study cleavage specificity [2–5].

On the other hand, little is known about what happens to the protease after its mutation and post-maturation phases of the viral life cycle. It raises questions with regard to the involvement of the protease in breakdown of host proteins related to the immune system, the protein synthesis process, gene regulatory pathways and so on. So far, it has been discovered that the protease acts on more than 20 variant non-viral proteins, such as Actin [6] and Vimentin [7]. However, there is lack of comprehensive information about the interaction between the protease and non-viral proteins. Therefore, the study of the susceptibility of host proteins in native states to hydrolysis by the protease is important to understand the role of HIV-1 protease in its host cell.

The two cleavage problems, cleaving of short oligopeptides and cleaving of native proteins, are related but different. Short oligopeptides or denatured proteins do not have folded structures. It is known that the protease has an active site with eight subsites, where eight corresponding residues can be bound. There are quite a lot of oligopeptides that have been experimentally verified as substrates of

*Corresponding author.

HIV-1 protease. The cleavage specificity of the protease is both sensitive to its context and broad. We have in our previous work collected an extended data set with 746 octamers [8] and built a predictor with 92% sensitivity and specificity for predicting cleavage of short oligopeptides. In contrast to short oligopeptides are proteins in their native states folded with complex structures. There are only around 20 tested native protein substrates reported in the literature. In total, there are around 42 cleaved sites in those native proteins. On average, a protein with a length of about 400 amino acids has only one or two cleavage sites. In other words, the cleavages in native proteins are rare cases. Due to the rarity and complex structure, this cleavage problem is much harder to attack than the one on short oligopeptides. The two problems are related in the sense that cleavage sites in short oligopeptides are very likely to be cleaved in native proteins if it is located at surface exposed regions. On the other hand, cleavage sites in native proteins might not be cleaved since the local environment makes it recognize some specific structures.

The aim of the present work was to predict cleavage sites in native proteins by combining information from short oligopeptides and native proteins. This is complicated since the information from short oligopeptides is difficult to transfer to native proteins, and vice versa, but this is important to do since experiments on oligopeptides are much easier to perform, and they are more abundant in the literature than experiments on native proteins.

2. Systems and methods

There are about 42 experimentally verified cleavage sites within 21 proteins with a total length of 8212 amino acids. Native protein cleavage sites are rarely observed, which implies that the cleavage sites are in a tiny region of the whole protein sequence and structure space.

As mentioned before, the two cleavage problems are different. A predictor based on short oligopeptides should discover all true cleavage sites but with lots of false positive ones. Taking the Bcl2 protein [9] as an example, it has 205 amino acids, but only one cleavage site. The predictor predicts 55 cleavage sites including the true one. Therefore, the predictor based only on short oligopeptides does not work well on native proteins. This is not surprising since some predicted cleavage sites might not be exposed to the protease or their local secondary structures may prevent the binding with the protease.

Tyndall *et al.* [10, 11] have targeted the recognition of substrates and ligands by proteases based on PDB files. They found that proteases generally recognize the extended beta strand conformation in the active sites. Peptidic compound's structure can be defined by their ϕ and ψ angels. But strictly speaking, short oligopeptides do not contain much structure information. Therefore it is not possible to build a predictor based on short oligopeptides for native proteins, as the example with Bcl2 shows.

Is it possible to get structure information for proteins? We know that as far as ligands go, unless they are peptidic compounds, then secondary structure cannot be readily defined. Although there are lots of PDB files describing ligand structure information, it is almost impossible to find structure information for a whole protein, except a short part of it. So, lack of experimental structure information is a problem. We use structure predictors to get secondary structure information. Many research groups have developed secondary structure predictors. Today, some predictors can reach around 80% correct prediction performance. In this way, secondary structure information can be accessed. The risk here is that it contains noise. However, if it contains more information than noise, then it should still improve the prediction. The same goes for solvent accessibility information.

Due to little information about the cleavage of native proteins and insufficient structure information on proteins, it would be hard to directly work on the native protein level. As mentioned in the above section, there are much more data available for short oligopeptides. Fortunately, our predictor based on short oligopeptides contains information about cleavage specificity of the protease, but predicts too many false positives on native proteins since it is not possible to take structure into account on short oligopeptides. By accessing some prediction servers to get secondary structure and solvent accessibility information, we can combine them with the information from short oligopeptides to build a predictor to determine the cleavage of native proteins.

2.1. Hierarchical classifier

Boyd *et al.* [12] have built a publicly accessible bioinformatics tool to build computational models of protease specificity which could be based on amino acid sequences, expert knowledge, and secondary/tertiary structure of substrates. However, their way to build prediction models was mainly based on protease specificity profiles, which is too flexible to tune. In addition, they extracted accessibility surface area information from PDB files, which might not be available for interesting proteins. Furthermore, they used a rule based method to use secondary structure information, instead of a data-driving one.

Here we used a three-level hierarchical classifier to combine the information from oligopeptides and native proteins. Figure 1 illustrates the structure of the hierarchical classifier.

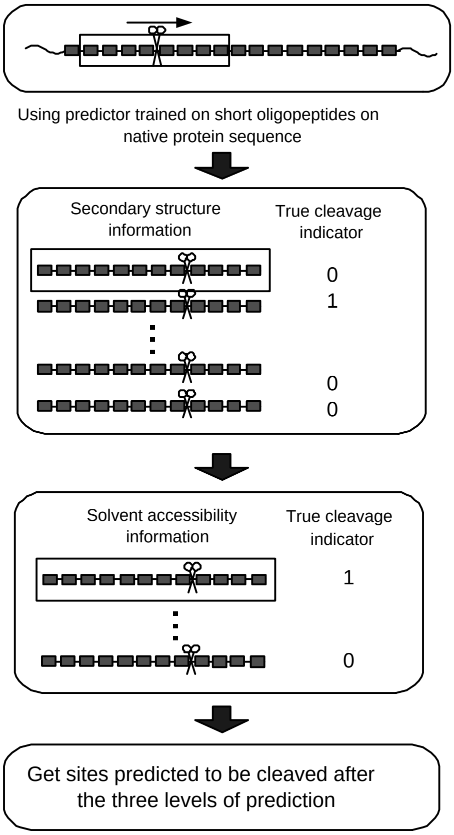
Using predictor trained on short oligopeptides on native protein sequence
Secondary structure information
True cleavage indicator
0
1
0
0
Solvent accessibility information
True cleavage indicator
1
0
Get sites predicted to be cleaved after the three levels of prediction

Fig. 1. A three-level hierarchical classifier, which combines information from short oligopeptides and native proteins.

(1) At the first classification level, the predictor trained using short oligopeptides and denatured proteins with a window size of 8 amino acids, meaning 4 residues at both sides of the cleavage sites, moves along a protein sequence and predicts all possible cleavage sites. Only sites predicted to be cleaved are collected with true cleavage indicators (class labels). The predictor works like a filter on the protein sequence level that only removes a part of true non-cleaved sites.

(2) At the second level, secondary structure information around these predicted cleavage sites are collected with a larger window size to include residue interaction and the window does not need to be symmetric around its cleavage site. A predictor trained using the secondary structure information is used to check the cleaving at those sites.

(3) At the final step, solvent accessibility information with the same window size as the second step around the remaining cleaved sites is collected. If residues inside the window are not exposed to the protease, it should not be cleaved, and thus removed from the cleaving list. Only those claimed to be cleaved at this step are classified as to be cleaved by the whole hierarchical classifier.

With regards to the first classification level, our previous work [8, 13] has discussed how to build a classifier based on short oligopeptides. The first classification level should never miss a true cleaving site. In the sense that it should be tuned to never produce false negatives, at the cost of some more false positives. At the last step, in order to measure the fraction of exposed volume of all residues inside that window, a conservative measurement is taken in such way that if 90% residues inside the window are buried, then the whole fragment is not accessible to the protease.

2.2. Data

Cleavage sites in 21 native proteins were collected from the literature [6, 7, 9, 14–23]. Similar sequences with minor mutations sites were not included since they contain redundant information. These protein sequences were submitted to a structure and solvent accessibility prediction server [24] (http://www.predictprotein.org/), where PROFsec and PROFacc were used to get secondary structure and solvent accessibility information individually.

3. Algorithms

Two generative models, a naive Bayes classifier and a Bayesian inference model, and two discriminative models, logistic regression and support vector machines (SVM), were tested for the cleavage prediction. For a generative model, the data distribution is either known or assumed to be close to a well known distribution. For a discriminative model, the density estimation is not needed. It directly works on the model to find optimum values for its parameters.

3.1. Rare case detection

The cleavage site prediction problem is a rare case detection. For rare case detection with an imbalanced data set, there is a majority class and a minority class. Classifiers tend to be biased towards the majority class but sampling methods (i.e., undersampling of majority class and over-sampling of minority class) can compensate for this to some extent. We use the synthetic minority over-sampling technique (SMOTE) introduced by Chawla *et al.* [25]. It introduces new data by randomly choosing a sample and interpolating new samples between it and its K-nearest samples.

Classification accuracy is a common measure for evaluating model performance. However, the data set is very imbalanced and a classifier that always predicts uncleaved is correct in more than 97% of the cases. Therefore accuracy metric is not a suitable one for this problem. Good metrics for this problem are sensitivity, specificity, geometric mean (G), which is square root of the product of sensitivity and specificity, or area under ROC (receiver operating characteristic) curve. We use all of them to evaluate and compare models.

3.2. Naive Bayes

The secondary structure predictor outputs probabilities. We use the notation $\pi_j^i = (\pi_{E,j}^i, \pi_{H,j}^i, \pi_{L,j}^i)$ for the secondary structure probabilities for position j of sample i. The numbers are provided by the secondary structure predictor and are normalized so that $\pi_{E,j}^i + \pi_{H,j}^i + \pi_{L,j}^i = 1$. The index j runs from 1 to J (the size of the input window). We assume that the data set π_j^i, where $i = 1, \ldots, N$ (the number of cleaved samples), in cleaved class has dirichlet distribution at position j. It is the same for the non-cleaved class. In addition, we assume that all positions inside the window are independent. In total, there are $3 \times J$ parameters for each class and we use maximum likelihood to estimate those parameters. The posterior probability needed for the classification decision is computed using Bayes' theorem. The Fastfit MATLAB toolbox was used to estimate dirichlet distribution parameters.

3.3. Bayesian inference

Each amino acid residue has a specific structure in native proteins and HIV-1 protease recognizes specific structures. We can interpret each helix, strand and loop probability set at each position as the probability to observe H, E and L at that position if we randomly draw new samples from unseen but possible structure character sequence space around cleaved sites. In other words, we can draw new samples representing possible structure patterns (i.e. HHHHL…LLL) from the structure probability data set. Using the drawn new structure character sequences, we can estimate the parameters on the dirichlet distributions at different positions. When we predict a new structure probability data, we draw a set of structure sequences from it and use the dirichlet distributions to calculate the probability for observing those structure character sequences and average them to get its posterior probability. We used Gibbs-sampling method to implement it.

3.4. Logistic regression

Logistic regression has the form: $\log \frac{P(\theta=1|\pi)}{P(\theta=0|\pi)} = \mathbf{w} \cdot \pi + \mathbf{b}$, where π is the secondary structure probability for residues inside a window and θ denotes class label. The parameters, $\mathbf{w}$ and $\mathbf{b}$, are fitted using maximum likelihood with MATLAB StatBox toolbox (version 4.2).

3.5. SVM

To train SVM with a very imbalanced data set makes the decision boundary biased towards majority class. Randomly under-sampling, over-sampling method (i.e. SMOTE) were used in our experiments to remove and add secondary structure probability data respectively. The problem with this method is that there are quite a lot of parameters to tune (constraints, kernel parameters for SVM, sampling rate and ratio between two classes after sampling and number of nearest neighbors in SMOTE). Cross-validation was used to find their optimum values for good generalization performance. We used the libSVM [26] MATLAB toolbox to train SVM.

4. Experiments and results

4.1. Exploring the data set

We explored the structure sequence data from the secondary structure predictor output to see if the structure data set contains any information to separate cleaved and non-cleaved class. Figure 2 displays the probabilities for observing L, H and E at each position for non-cleaved (upper part) and cleaved (bottom part) class.

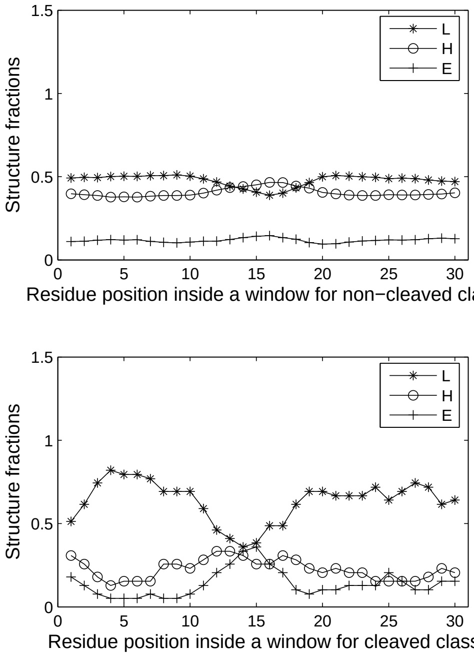

Fig. 2. The probabilities for observing loop, helix and strand structures at each site for non-cleaved (upper part) and cleaved (bottom part) class. We use window (15,15) to demonstrate it.

There is almost no structure difference at different positions for the non-cleaved class. For each structure, it is uniformly distributed inside the window. For the cleaved class, loops are less likely to be observed around the active site. However, strands are more likely to be observed in the vicinity of the cleaved site, which agrees with Tyndall's conclusion that extended conformation is preferred at active sites. It is worth noting that the probability to observe helix structure increases somewhat when it is closer to the active site. This is probably due to the structure prediction performance on helix, strand and loop. From the secondary structure prediction server, it states that "PHD as well as other methods focus on predicting hydrogen bonds. Consequently, occasionally strongly predicted (high reliability index) helices are observed as strands and vice versa (expected accuracy of PHDsec)."

4.2. Experiments

After using the first level predictor on the 21 proteins having 42 true cleavage sites, the prediction results are shown in Table 1.

Table 1. Prediction results after using the first level predictor on the 21 native protein sequences. Sensitivity=100%; false positive rate=16%; precision=2.5%.

	True cleavage sites	True non-cleavage sites
Predicted to be cleaved sites	42	1613
Predicted to be non-cleaved sites	0	6557

We can see that after this step, all true cleavage sites were kept with 100% sensitivity (TP/(TP+FN)) and 16% (false positive rate= FP/(FP+TN)) non-cleaved sites were predicted to be cleaved. The precision (TP/(TP+FP)) is 2.5%, which means if there is one true cleavage site, the predictor predicts 38.4 non-cleaved sites to be cleaved. In total, there are 1655 sites predicted to be cleaved and fed into the second level predictor.

The next experiment was to try the four different classifiers, two generative and two discriminative methods, with different window sizes using secondary structure information on the 1655 sites and estimate the generalization performance using cross-validation. AUC (area under ROC curve) was used to compare the four classifiers. The cross-validation was done in the following way. For each classifier, the whole data set was randomly divided into two parts, 80% of the data set was used to train the classifier, and the remaining 20% was used to test its prediction performance. The process was repeated 100 times for each window size. For SMOTE over-sampling, 5-nearest neighbors were used to interpolate new samples. The cleaved samples were over-sampled 3 times and the uncleaved samples were under-sampled to get the same number of cleaved samples after its over-sampling process.

Table 2. The best three classification generalization performances, area under ROC curve, of the four different classifiers only based on the secondary structure information. It reports its mean value and its standard deviation.

	Logistic regression	SVM	Naive Bayes	Bayesian inference
1	0.706 (0.095)	0.701 (0.068)	0.685 (0.079)	0.67 (0.079)
2	0.702 (0.083)	0.698 (0.081)	0.684 (0.081)	0.67 (0.080)
3	0.701 (0.083)	0.697 (0.075)	0.682 (0.080)	0.66 (0.078)

Table 2 lists the three largest AUC values for each classifier. Although the performance variance is around 8%, we can see that, on average, the best classifier is the one with logistic regression method and SVM with sampling methods performs as good as logistic regression method. Naive Bayes classifier is little inferior to them and Bayesian inference is almost the same as naive Bayes.

The third experiment is to estimate the influence of the cutoff value on the sensitivity, specificity and precision performance of the best classifier, logistic regression. In this experiment, during training, cross-validation was used to tune the cutoff value over the outputs of the classifier in such way that it gives the best geometric mean value. The generalization performances were estimated by using this tuned cutoff value over held out test data set. Table 3 lists the sensitivity, specificity and precision using the tuned cutoff value on the test data set. If only the predictor based on short oligopeptides is used, the precision is around 2.5%. After using the second predictor based on secondary structure information, the precision increases to 4.4%, which increases 1.7 times. In other words, for each true cleavage site there are 38 false ones from the first level classifier. After using the second classifier, for each correct cleavage site, it predicts 22 false cleavage sites.

Table 3. Sensitivity, specificity and precision performance were estimated with the tuned cutoff value with logistic regression classifier.

	Mean	Standard deviation
Sensitivity	0.64	0.21
Specificity	0.65	0.10
Precision	0.044	0.013
TP	5.1	1.7
FP	112.4	30.5
FN	2.9	1.7
TN	204.6	30.5

There are no exact criteria to choose the best cutoff value. For this cleavage site prediction, if it is required to reach 90% sensitivity, we can lower the cutoff value, but then get more false positive ones. Generally, the ROC curve can give a good idea to pick the cutoff value for a specific requirement. Figure 3 displays the ROC curve of the best classifier with the logistic regression method.

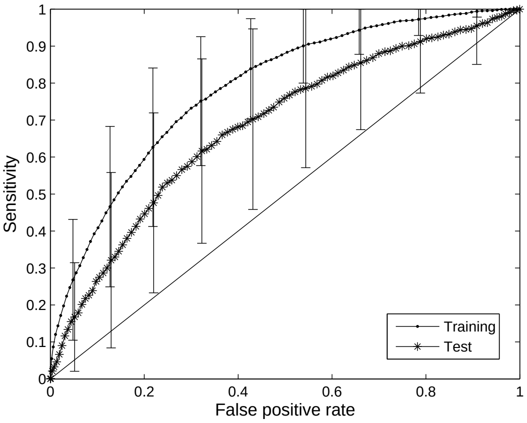

Fig. 3. ROC curve of the best classifier with logistic regression method. The upper curve is ROC measured on the training data set and the lower curve is for the test data set. Error bars with standard deviation are also displayed.

5. Discussion

Two discriminative (logistic regression and SVM) and two generative models (naive Bayes and Bayesian inference) were used to build classifiers with secondary structure information. From our experiments, there is no major difference between them. The logistic regression, is the best among them on average. For Bayesian methods, a bad data and model parameter distribution assumption could affect their performance quite a lot. With Bayesian method, there are $2 \times 3 \times J$ parameters needed to estimate for dirichlet distributions. Logistic regression has only $2 \times J + 1$ parameters in the model. While, SVM with sampling methods has around $2 \times J + 5$ in the model. Due to the lack of data, discriminative approaches are better than generative ones in general. Since logistic regression has few parameters and is fast to train, it is the method of choice in this case.

During the experiments, the secondary structure and solvent accessibility information were predicted only by one prediction sever. It has not been tested how sensitive the classifiers are to the predicted structure and accessibility information. In addition, the hierarchical classifier does not consider cleaving ordering in a protein if there are more than one cleavage site. If a protein is cleaved at the first cleavage site, the protein is cleaved into two fragments and their secondary structures might change and previously buried parts can be exposed to the protease.

Another useful information is to use the absolute positions of predicted cleavage sites. Normally it is impossible to have cleavage sites at the very end of a native protein. Therefore, we can use this rule to directly rule out some false predicted cleaved sites.

To conclude, the hierarchical classifier, which combines protein sequences, experimentally tested short oligopeptides, protein secondary structure and solvent accessibility information, can be used to detect the cleavage sites on native proteins. By using the secondary level classification based on secondary structure information, the false positive ratio is more than halved compared to the classifier only using short oligopeptide information on the first level. Therefore structure and solvent accessibility data provide information to predict protease-substrate interactions. This method can also be used for other cleavage problems on native proteins.

Acknowledgments

The work was supported by the National Research School in Genomics and Bioinformatics hosted by Göteborg University, Sweden. The author also thanks Dr Martin Sköld, Dr Joel Tyndall, Prof Thorsteinn Rögnvaldsson and Dr Daniel Garwicz for discussions and pointers to important references.

References

1. Beck ZQ, Hervio L, Dawson PE, Elder JE, and Madison EL. Identification of efficiently cleaved substrates for HIV-1 protease using a phage display library and use in inhibitor development. *Virology*, 274:391–401, 2000.
2. Ridky TW, Bizub-Bender D, Cameron CE, Weber IT, Wlodawer A, Copeland T, Skalka AM, and Leis J. Programming the rous sarcoma virus protease to cleave new substrate sequences. *J Biol Chem*, 271:10538–10544, 1996.
3. Ridky TW, Cameron CE, Cameron J, Leis J, Copeland T, Wlodawer A, Weber IT, and Harrison RW. Human immunodeficiency virus, type 1 protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. *J Biol Chem*, 271:4709–4717, 1996.
4. Tözsér J, Bagossi P, Weber IT, Louis JM, Copeland TD, and Oroszlan S. Studies on the symmetry and sequence context dependence of the HIV-1 proteinase specificity. *J Biol Chem*, 272:16807–16814, 1997.
5. Tözsér J, Zahuczky G, Bagossi P, Louis JM, Copeland TD, Oroszlan S, Harrison RW, and Weber IT. Comparison of the substrate specificity of the human T-cell leukemia virus and human immunodeficiency virus proteinases. *Eur J Biochem*, 267:6287–6295, 2000.
6. Tomasselli AG, Hui JO, Adams L, Chosay J, Lowery D, Greenberg B, Yem A, Deibel MR, Zurcher-Neely H, and Heinrikson RL. Actin, troponin c, alzheimer amyloid precursor protein and pro-interleukin 1 beta as substrates of the protease from human immunodeficiency virus. *J Biol Chem*, 266(22):14548–53, 1991.
7. Shoeman RL, Honer B, Stoller TJ, Kesselmeier C, Miedel MC, Traub P, and Graves MC. Human immunodeficiency virus type 1 protease cleaves the intermediate filament proteins vimentin, desmin, and glial fibrillary acidic protein. *Proc Natl Acad Sci USA*, 87(16):6336–6340, 1990.
8. You L, Garwicz D, and Rögnvaldsson T. Comprehensive bioinformatic analysis of the specificity of

human immunodeficiency virus type 1 protease. *J Virol*, 79(19):12477–86, 2005.

9. Strack PR, Frey MW, Rizzo CJ, Cordova B, George HJ, Meade R, Ho SP, Corman J, Tritch R, and Korant BD. Apoptosis mediated by hiv protease is preceded by cleavage of bcl-2. *Proc Natl Acad Sci USA*, 93(18):9571–6, 1996.
10. Fairlie DP, Tyndall JD, Reid RC, Wong AK, Abbenante G, Scanlon MJ, March DR, Bergman DA, Chai CL, and Burkett BA. Conformational selection of inhibitors and substrates by proteolytic enzymes: implications for drug design and polypeptide processing. *J Med Chem*, 43(7):1271–81, 2000.
11. Tyndall JD, Nall T, and Fairlie DP. Proteases universally recognize beta strands in their active sites. *Chem Rev*, 105(3):973–99, 2005.
12. Boyd SE, Garcia de la Banda M, Pike RN, Whisstock JC, and Rudy GB. Pops: A computational tool for modeling and predicting protease specificity. *Proceedings of the IEEE Computer Society Bioinformatics Conference, Stanford, CA*, page 372, 2004.
13. Rögnvaldsson T and You L. Why neural networks should not be used for hiv-1 protease cleavage site prediction. *Bioinformatics*, 20(11):1702–1709, 2004.
14. Meier UC, Billich A, Mann K, Schramm HJ, and Schramm W. alpha 2-macroglobulin is cleaved by hiv-1 protease in the bait region but not in the c-terminal inter-domain region. *Biol Chem Hoppe Seyler*, 372(12):1051–6., 1991.
15. Oswald M and von der Helm K. Fibronectin is a non-viral substrate for the hiv proteinase. *FEBS Lett*, 292(1-2):298–300, 1991.
16. Riviere Y, Blank V, Kourilsky P, and Israel A. Processing of the precursor of nf-kappa b by the hiv-1 protease during acute infection. *Nature*, 350(6319):625–6, 1991.
17. Tomaszek TA Jr, Moore ML, Strickler JE, Sanchez RL, Dixon JS, Metcalf BW, Hassell A, Dreyer GB, Brooks I, Debouck C, and et al. Proteolysis of an active site peptide of lactate dehydrogenase by human immunodeficiency virus type 1 protease. *Biochemistry*, 31(42):10153–68, 1992.
18. Chattopadhyay D, Evans DB, Deibel MR Jr, Vosters AF, Eckenrode FM, Einspahr HM, Hui JO, Tomasselli AG, Zurcher-Neely HA, Heinrikson RL, and Sharma SK. Purification and characterization of heterodimeric human immunodeficiency virus type 1 (hiv-1) reverse transcriptase produced by in vitro processing of p66 with recombinant hiv-1 protease. *J Biol Chem*, 267(20):14227–32, 1992.
19. Freund J, Kellner R, Konvalinka J, Wolber V, Kraus-slich HG, and Kalbitzer HR. A possible regulation of negative factor (nef) activity of human immunodeficiency virus type 1 by the viral protease. *Eur J Biochem*, 223(2):589–93, 1994.
20. Mildner AM, Paddock DJ, LeCureux LW, Leone JW, Anderson DC, Tomasselli AG, and Heinrikson RL. Production of chemokines ctapiii and nap/2 by digestion of recombinant ubiquitin-ctapiii with yeast ubiquitin c-terminal hydrolase and human immunodeficiency virus protease. *Protein Expr Purif*, 16(2):347–354, 1999.
21. Alvarez E, Menendez-Arias L, and Carrasco L. The eukaryotic translation initiation factor 4gi is cleaved by different retroviral proteases. *J Virol*, 77:12392–400, 2003.
22. Tomasselli AG, Howe WJ, Hui JO, Sawyer TK, Reardon IM, DeCamp DL, Craik CS, and Heinrikson RL. Calcium-free calmodulin is a substrate of proteases from human immunodeficiency viruses 1 and 2. *Proteins*, 10(1):1–9, 1991.
23. Álvarez E, Castelló A, Menéndez-Arias L, and Carrasco L. Human immunodeficieny virus protease cleaves poly(a) binding protein. *Biochem. J.*, Immediate Publication, doi:10.1042/BJ20060108, 2006.
24. Rost B, Yachdav G, and Liu J. The predictprotein server. *Nucleic Acids Research*, 31(13):3300–3304, 2003.
25. Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
26. Chang CC and Lin CJ. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.