Mining viral protease data to extract cleavage knowledge

Ajit Narayanan, Xikun Wu and Z. Rong Yang

School of Engineering and Computer Sciences (Bioinformatics Laboratory), Old Library, University of Exeter, Exeter, EX4 4PT, UK

Received on January 19, 2002; revised and accepted on March 28, 2002

ABSTRACT

Motivation: The motivation is to identify, through machine learning techniques, specific patterns in HIV and HCV viral polyprotein amino acid residues where viral protease cleaves the polyprotein as it leaves the ribosome. An understanding of viral protease specificity may help the development of future anti-viral drugs involving protease inhibitors by identifying specific features of protease activity for further experimental investigation. While viral sequence information is growing at a fast rate, there is still comparatively little understanding of how viral polyproteins are cut into their functional unit lengths. The aim of the work reported here is to investigate whether it is possible to generalise from known cleavage sites to unknown cleavage sites for two specific viruses—HIV and HCV. An understanding of proteolytic activity for specific viruses will contribute to our understanding of viral protease function in general, thereby leading to a greater understanding of protease families and their substrate characteristics.

Results: Our results show that artificial neural networks and symbolic learning techniques (See5) capture some fundamental and new substrate attributes, but neural networks outperform their symbolic counterpart.

Availability: Publicly available software was used (Stuttgart Neural Network Simulator—http://www-ra.informatik.uni-tuebingen.de/SNNS/ and See5—http://www.rulequest.com). The datasets used (HIV, HCV) for See5 are available at: http://www.dcs.ex.ac.uk/~anarayan/bioinf/ismbdatasets/

Keywords: protease cleavage; protease inhibitors; machine learning; neural networks; decision trees.

Contact: a.narayanan@ex.ac.uk; wuxikun@yahoo.com; z.r.yang@ex.ac.uk

INTRODUCTION

Intact human immunodeficiency HIV and Hepatitis C (HCV) virions are endocytosed (inserted into a cell) via specific cellular receptors on human cells. For ‘retroviruses’, a single stranded RNA sequence (typically between 8–12 kilobases and containing at least 9 genes, including genes for producing core protein precursors (gag), envelope proteins (env) and pol (reverse transcriptase, integrase and protease)) is then transcribed by one of the enzymes accompanying the RNA sequence into double stranded DNA (by the reverse transcriptase enzyme) and integrated with the host genome (by the integrase enzyme). The DNA provirus (originally reverse transcribed from RNA or single stranded DNA, or simply the original double stranded inserted viral DNA), when expressed, is transcribed into messenger RNA (mRNA) and translated into a protein chain (viral polyproteins), giving rise to new viral molecules which then reassemble to form complete virions that are then released and infect more cells.

Viral protease is the third enzyme typically accompanying viral DNA or RNA into the cell, although protease can also self-cleave itself naturally from the viral polyprotein if it isn’t introduced through endocytosis. It cleaves the precursor viral polyproteins (the substrate) at specific cleavage-recognition sites when they emerge from the ribosomes of the host cell as one long sequence. This cleavage step is essential in the final maturation step of HIV and HCV. That is, protease is responsible for the post-translational processing of the viral gag and gag-pol polyproteins to yield the structural proteins and enzymes of the virus for further infection (Figure 1). If viral protease action can be inhibited by drugs so that such cleavage-recognition sites cannot be identified, viral replication can be stopped.

Currently, four families of protease have been identified in cells and viruses: serine, cysteine, aspartic and metalloprotease, each of which may have ‘clans’ or subfamilies (e.g. Rawlings and Barrett, 1994; Paetzel and Strynadka, 1999). Normal animal tissue contains mostly serine, cysteine and metalloproteases, whereas plant cells contain serine and cysteine proteases. Bacterial cells contain mostly serine and metalloproteases. The HIV and HCV proteases belong to the aspartic and serine families, respectively. Examples of viral cysteine proteases can be found in human enteric caliciviruses such as the Norwalk-like viruses (NLVs) and the Sapporo-like...
When the HIV viral polyprotein emerges from the CD4+T cell’s ribosomes (a), potential recognition sites (amino acid sequences of length 8) become available to the viral protease for cleavage. If a binding site is found the protease cleaves the polyprotein (b) after locking on to the polyprotein (substrate) and cutting (c) at the active site. Protease inhibitors (d) are a relatively new form of anti-viral agent which, through competitive inhibition, prevent the protease from further functioning.

Viruses (SLVs) (Liu et al., 1999). Retroviruses need a protease from the aspartic family during their lifecycle, as described above. Proteases unite with their substrates mainly but not exclusively through non-covalent forces, such as hydrogen bonds, ionic interactions and hydrophobic interactions, which permit binding between an ‘active site’ or ‘binding site’ of the protease and a ‘recognition site’ of the substrate over a broad front (several residues). There are two current methods for inhibiting such binding. Competitive inhibition consists of identifying an inhibitor that will lock on to the protease’s active site and thereby prevent that protease from binding to any further substrate (Figure 1). These inhibitors are used only once (one inhibitor—one protease). Such inhibitors can be overcome, however, if the presence of substrate in comparison to inhibitor is sufficiently large to overcome the presence of the inhibitor. Non-competitive inhibition on the other hand works by identifying a regulatory site rather than an active site of the protease so that the inhibitor, when bound to the regulatory site, distorts the
structure of the protease and thereby prevents it from binding to its substrate.

HIV protease inhibitors (e.g. Saquinivir, Ritonavir, Indinavir) are a relatively recent type of antiviral agent which are considered more powerful than first generation reverse transcriptase inhibitors (e.g. AZT), with provisional results indicating that these protease inhibitors can reduce the amount of virus by as much as 99%. For HIV-1, protease inhibitors have been used since 1996, and between 1996 and 1998 the number of deaths caused by AIDS fell by over 70%. These inhibitors are compounds which permit competitive inhibition. However, a protease inhibitor will not eradicate a virus from an infected person’s body. Also, the high degree of viral mutation caused, it is believed, by the error-proneness of the reverse transcriptase enzyme as it transcribes single stranded RNA into double stranded DNA, has led researchers to think that protease inhibitors may have to ‘co-evolve’ with their protease targets. While protease appears relatively unaffected by significant mutations in their target polyprotein, probably due to these mutations being in non-recognition substrate regions, research into protease inhibitors will need to take into account that proteases appear to be able to evolve quickly and efficiently in response to such mutations. Given the rate at which viruses mutate, this could lead to a rapid proliferation of inhibitors and increased cost. The task here is to increase our understanding of protease functionality as well as substrate characteristics, especially in the area of polyprotein functional unit lengths.

Another fundamental requirement is that a protease inhibitor must not interfere with normal human protease function (it is estimated that the human genome contains a hundred or more genes coding for proteases of various types). The more specific the recognition substrate, the more likely it is that a particular inhibitor can be found which targets the highly specific substrate or the active/binding site of the protease. The focus of the research described here is on attempting to identify, through machine learning approaches, the key features of HIV and HCV substrate (polyprotein) attributes which permit, as well as increase the probability of, cleavage through the identification of recognition sites of highest specificity. Given that other viruses, such as cytomegalovirus, rhinovirus and picornaviruses, also involve protease (all picornaviruses encode one or more serine proteases), there is a need to explore ways in which techniques from artificial intelligence can be usefully applied to direct future protease inhibitor experimental design.

Viral protease binds to n-residue amino acid recognition sequences in the polyprotein, where n can vary typically between 8 and 15 from virus to virus (the approximate length of recognition sequences is determined through in vitro experiments), and then cleaves the polyprotein wherever it finds these n-residue amino acid sequences.

A viral polyprotein can be cleaved 8 or 9 times or more by one or more proteases. These n-residue recognition regions vary in content within the same polyprotein, yet the protease binds to different cleavage points despite this variation. For HIV (both HIV-1 and HIV-2) and retroviruses in general, there is a growing body of evidence that suggests that cleavage typically takes place when medium to large hydrophobic residues such as phenylalanine (F), tryptophan (W) or tyrosine (Y) are present on either side of the cleavage point in the substrate (Pettit et al., 1991), but it is not known whether other residues help or hinder cleavage.

**SYSTEMS AND METHODS**

Laboratory researchers have started to produce data on what represents cleavage sites. Given a set of specific recognition regions for a particular viral protease (typically, 8 or 9 such recognition regions in one polyprotein sequence will be identified for a particular virus), oligopeptide regions (that is, artificially produced n-residue amino acid sequences) can be constructed which vary in some way from the real recognition regions in polyproteins to represent mutations to the viral DNA/RNA. The behaviour of protease on these oligopeptide regions can then be observed in vitro and recorded. A significant amount of potential cleavage site data for HIV and HCV has been produced in this manner, constituting data sets for pattern recognition and machine learning applications. Another way to produce negative cleavage sites is to assume that regions between known cleavage sites are non-cleavage. That is, as well as trying to produce cleavage and non-cleavage oligopeptide sequences in vitro, the full polyprotein sequence of the virus can be analysed by a computer and fixed length sequences (either 8 or 10 amino acids long for the HIV and HCV polyproteins, respectively) which are currently not known to be cleavage sites are extracted as negative cleavage sequences. This assumes, falsely, that all 8 or 10 amino acid sequences within the polyprotein are equally likely candidates for cleavage, whereas the real recognition sites are positioned near gene boundaries. Nevertheless, this method can help constrain the machine learning method to be more specific (in the technical sense of identifying true negatives—see Table 1 for a definition of specificity).

Identified by molecular cloning in 1989 (Choo et al., 1989; Kuo et al., 1989), the Hepatitis C virus (HCV) is the major etiological agent of non-A and non-B hepatitis. Infection by HCV causes chronic liver disease, which is a serious health problem worldwide (Jenny-Avital, 1998). During the next few years, HCV-related liver disease and hepatocellular carcinoma may kill more people than AIDS (Cohen, 1999). There is currently no effective therapy against HCV (Alter, 1997), except through the use of general anti-viral agents (e.g. interferons and interleukins).
sites on the enzyme corresponding to residues P6-P5-

The binding or active S2-S1-S1 P4-P3-P2-P1-P1

P4-P3-P2-P1-P1

S8

family Flaviviridae (Francki

which are not fully effective. HCV is a member of the

See5

ANN

ANN

ANN

ANN

See5

See5

Table 1. The results of running ANNs and See5 on the 10 sets of test data. The upper table gives the overall performance on the HIV data, and the lower table on the HCV data. The means show that ANNs outperform See5 on both test data sets. Accuracy (AC), specificity (SP, or true negative rate) and sensitivity (SE, or true positive rate) are defined as follows: AC=(TP+TN)/(TP+TN+FP+FN); SE=TP/(TP+FN); SP=TN/(TN+FN), where TP=true positive, TN=true negative, FP=false positive, and FN=false negative.

HIV protease

<table>
<thead>
<tr>
<th></th>
<th>set1</th>
<th>set2</th>
<th>set3</th>
<th>set4</th>
<th>set5</th>
<th>Set6</th>
<th>set7</th>
<th>set8</th>
<th>set9</th>
<th>set10</th>
<th>mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY (%)</td>
<td>88.6</td>
<td>95.4</td>
<td>91.3</td>
<td>92.2</td>
<td>91.8</td>
<td>90.3</td>
<td>95.0</td>
<td>89.0</td>
<td>95.9</td>
<td>95.4</td>
<td>92.50</td>
<td>2.76</td>
</tr>
<tr>
<td>SENSITIVITY (%)</td>
<td>82.6</td>
<td>95.7</td>
<td>78.3</td>
<td>83.2</td>
<td>82.6</td>
<td>87.0</td>
<td>78.3</td>
<td>69.6</td>
<td>82.6</td>
<td>87.0</td>
<td>82.19</td>
<td>6.94</td>
</tr>
<tr>
<td>SPECIFICITY (%)</td>
<td>86.0</td>
<td>94.0</td>
<td>92.0</td>
<td>94.0</td>
<td>90.0</td>
<td>80.0</td>
<td>94.0</td>
<td>84.0</td>
<td>100</td>
<td>94.0</td>
<td>90.80</td>
<td>5.90</td>
</tr>
<tr>
<td>See5</td>
<td>82.2</td>
<td>84.9</td>
<td>80.0</td>
<td>90.4</td>
<td>86.3</td>
<td>80.8</td>
<td>89.0</td>
<td>86.3</td>
<td>87.7</td>
<td>86.3</td>
<td>85.47</td>
<td>3.31</td>
</tr>
</tbody>
</table>

HCV NS3 protease

<table>
<thead>
<tr>
<th></th>
<th>set1</th>
<th>set2</th>
<th>set3</th>
<th>set4</th>
<th>set5</th>
<th>Set6</th>
<th>set7</th>
<th>set8</th>
<th>set9</th>
<th>set10</th>
<th>mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY (%)</td>
<td>96.0</td>
<td>96.0</td>
<td>95.3</td>
<td>95.5</td>
<td>97.3</td>
<td>98.0</td>
<td>94.9</td>
<td>96.7</td>
<td>96.4</td>
<td>96.4</td>
<td>96.25</td>
<td>0.93</td>
</tr>
<tr>
<td>SENSITIVITY (%)</td>
<td>88.2</td>
<td>88.2</td>
<td>79.4</td>
<td>88.2</td>
<td>94.1</td>
<td>88.2</td>
<td>82.4</td>
<td>88.2</td>
<td>85.3</td>
<td>85.3</td>
<td>86.77</td>
<td>3.98</td>
</tr>
<tr>
<td>SPECIFICITY (%)</td>
<td>96.0</td>
<td>96.7</td>
<td>96.0</td>
<td>96.7</td>
<td>96.7</td>
<td>98.0</td>
<td>96.0</td>
<td>98.0</td>
<td>97.3</td>
<td>96.80</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>See5</td>
<td>94.0</td>
<td>90.2</td>
<td>90.2</td>
<td>88.0</td>
<td>87.0</td>
<td>93.5</td>
<td>96.2</td>
<td>87.5</td>
<td>88.6</td>
<td>93.5</td>
<td>90.87</td>
<td>3.21</td>
</tr>
</tbody>
</table>

HIV protease set1 set2 set3 set4 set5 Set6 set7 set8 set9 set10 mean STD

HIV protease set1 set2 set3 set4 set5 Set6 set7 set8 set9 set10 mean STD

HSC NS3 protease set1 set2 set3 set4 set5 Set6 set7 set8 set9 set10 mean STD

HSC NS3 protease set1 set2 set3 set4 set5 Set6 set7 set8 set9 set10 mean STD

Due to the flat and featureless substrate binding sites, NS3 requires a relatively large, at least decamer, peptide substrate (Steinkühler et al., 1996; Urbani et al., 1997; Zhang et al., 1997). The nomenclature of Schechter and Berger (Schechter and Berger, 1967) is applied here to designate the recognition sites on the polypeptides, P0-P5-P6-P7-P8-P9-P10-P11-P12-P13-P14 (P sequence), the cleavage point being between P1 and P1, (i.e. not exactly in the centre of the P sequence). The binding or active sites on the enzyme corresponding to residues P0-P5-P6-P7-P8-P9-P10-P11-P12-P13-P14 are indicated as S0-S5-S6-S7-S8-S9-S10-S11-S12-S13-S14 (S sequence). Unlike other serine protease such as trypsin and chymotrypsin that can cleave small peptide substrates, the NS3 protease has an unusual substrate recognition mechanism that makes it difficult to find small molecule inhibitors. Also due to the number of possible decapeptides formed from 20 amino acids \( (20^{10} = 1.024 \times 10^{13}) \), testing them one by one in the laboratory is currently impossible. Instead, a restricted number of oligopeptide sequences have been generated and the behaviour of NS3 on these has been observed and recorded. The machine learning and pattern recognition task is to use this restricted set to identify patterns in the P sequence to which the S sequence binds, to express these patterns in a way which allows generalisations to be made, and therefore to predict the behaviour of NS3 on all the remaining sequences for in vitro experiments. So far, apart from our own preliminary studies, there has been no attempt to use neural networks or symbolic machine learning/pattern recognition techniques to help predict polyprotein cleavability through the identification of recognition patterns in HCV P sequences.

HIV is the main causative agent of AIDS (acquired immunodeficiency syndrome) (Gallo et al., 1984). HIV protease is responsible for the post-translational processing of the viral gag and gag-pol polyproteins to yield the structural proteins and enzymes of HIV (Kantor et al., 2001). Immature but non-infectious viral particles are produced when there is loss of protease cleavage ability. Therefore, HIV protease has become a major target for the rational design of drugs against AIDS. HIV protease is
a member of the aspartic proteases. Its catalytic apparatus is made up of carboxyl groups from two aspartyl residues in the N- and C-terminal halves of the enzyme (Toh et al., 1985; Pearl and Talyor, 1987). There are eight different naturally occurring cleavage regions in the gag-pol polyprotein (Ridky et al., 1996), i.e. eight naturally occurring cleavage sites. Each region is symbolized by an eight amino acid sequence $P_4-P_3-P_2-P_1-P'_1-P_2'-P_3'-P'_4'$, whose counterparts in HIV protease are symbolized $S_4-S_3-S_2-S_1-S'_1-S'_2-S'_3-S'_4$. The catalytic site of the protease is between $S_1$ and $S_1'$, and the scissile bond (the peptide bond cleaved) is between $P_1$ and $P_1'$ (Weber and Harrison, 1996, following the active site model described by Berger and Schechter, 1970). Identifying general protease inhibitors for HIV protease is important (Chou, 1996). Various prediction methods have been applied recently to HIV polyprotein fragments. Yu-Dong Cai and Kuo-Chen Chou produced a first attempt at using a neural network in 1998 (Cai and Chou, 1998) to predict cleavability, with some success. However, the knowledge used by the neural network to make these predictions is locked within the neural network, due to the form of input representation used (to be described below). So far, apart from our own preliminary studies, we are not aware of any attempt to use symbolic machine learning/pattern recognition techniques to help predict HIV polyprotein cleavability.

**ALGORITHM**

The Hepatitis C Virus NS3 protease has several known cleavage sites in the polyprotein precursor. Previous work indicates that if one or more residues is changed or modified, cleavage ability will not always be damaged. In this way, many oligopeptides can be created as potential cleavage (‘positive’) and noncleavage (‘negative’) sequences, some of which are then tried in vitro. We collected 168 positive sequences from published papers, while the negative sequences come from both published papers and the HCV proteome. The region between two cleavage sites of NS3 protease can be considered as negative polyprotein recognition sites (i.e. non-cleavage), and a 10-residue window is applied through these regions to create further negative sequences. While this method of generating negative samples does not take into account gene boundaries (where cleavage sites are more likely to be found), the inclusion of these samples provides a useful specificity benchmark against which any conclusions concerning positive samples can be measured (see Table 1).

Also, for the HCV dataset, besides the normal 20 amino acids, there are five other molecules which usually appear in the decapeptide substrate: alpha-aminobutyric acid; norleucine; 3,3-diphenylalanine; beta-cyclohexylalanine; and 1,2,3,4-tetrahydroisoquinoline-3-L-carboxylic acid (25 residues altogether). Altogether, 752 negative sequences (consisting of oligopeptides as well as negative sequences obtained from the 10-residue window method described above) were collected to supplement the 168 positive sequences (920 sequences in total for HCV). For the HIV dataset, Yu-Dong Cai and Kuo-Chen Chou (Cai and Chou, 1998) described 114 positive oligopeptides and 248 negative oligopeptides (362 sequences in total for HIV).

Yu-Dong Cai and Kuo-Chen Chou (Cai and Chou, 1998) used 299 (60 are positive) sequences as the training set, 63 sequences (54 are positive) as the test set, for a feedforward backpropagation (FFBP) neural network. The prediction rate on the test set was reported to be $58/63=92.06\%$. We replicated their experiments using a different selection criterion for both the training and test sets: 20% positive and 20% negative oligopeptides are randomly selected to build the test set, with the remaining 80% of each being selected as the training set. This was done to test whether Cai and Chou’s (1998) results could have been affected by their selection criteria which led to an imbalance between the ratios of positive to negative samples in the training and test sets (60 positives out of 299 samples in the training set and 54 positives out of 63 samples in the test set). Furthermore, we constructed 10 different training and test sets, through random roulette wheel selection, preserving the 20% criterion. An identical neural network was then run on all ten training and test sets, three times each (to overcome any bias in initialisation), and the averages taken for the test tests. This again was done to test whether Cai and Chou’s (1998) results, which were obtained from what appears to be one run only, were repeatable. A roulette wheel selection method, again using the 20% criterion, was adopted to generate ten training and ten corresponding test sets for HCV data. All the data sets were input to both FFBP networks (constructed on the SNNS neural network package) and the symbolic machine learning package, See5 (Quinlan, 2000), which uses decision trees (identification trees) to identify the most important attributes (residue positions) to arbitrary levels to determine classification. The use of See5 was motivated by the need to extract, in symbolic form, the critical aspects of pattern recognition which allowed the otherwise non-symbolic neural networks to achieve high prediction rates.

**IMPLEMENTATION**

Recognition sequences are usually represented as character strings where each character represents an amino acid. Different methods are applied to change these strings to the right format for the two models. Symbolic input to See5 is straightforward and consists of a string of characters followed by the class of that sample. So, for example, for the HIV dataset, each octapeptide is given 9 attributes:
position1, position2, position3, position4, position5, position6, position7, position8, and cleavage-ability. The first eight attributes are just the same characters as in the octapeptide, and the ninth attribute, cleavage-ability, will be either cleavage (1) or non-cleavage (0), indicating whether this octapeptide can or cannot be cleaved by HIV protease. For HCV NS3 protease, each decapeptide is given 11 attributes: from position1 to position10, and cleavage-ability. For instance, a typical ‘record’ for the HIV dataset is: G,Q,V,N,Y,E,E,F,1, where the eight positions in the substrate are occupied by G through F and 1 signifies that this substrate is cleaved. A typical record for the HCV dataset is: D,L,E,V,V,R,S,T,W,V,0, where the ten positions in the substrate are encoded D through V and 0 signifies non cleavage. Five extra characters have been used to code for the five extra molecules found in HCV substrates (i.e. the HCV ‘amino acid alphabet’ consists of 25 characters rather than 20.)

The situation for neural networks is more complex, since inputs to neural networks are usually binary (0/1) or real valued rather than character. There is some evidence to support the view that neural network learning may be improved if input patterns are binary strings (Fausett, 1994). Since there are 20 amino acids, each amino acid is represented by a 20-bit binary string consisting of nineteen 0s and one 1, where the position of the 1 identifies the amino acid. An octapeptide is converted into an \(8 \times 20 = 160\)-bit binary string, as shown in Figure 2. This is the method adopted by Cai and Chou (1998). This representation method then leads to the use of a FFBP neural network with 160 input units (one for each bit of the input binary string), 8 hidden units, and one output unit. Training continued until the output node produced 0.9 or greater for positive samples and 0.1 or less for negative samples. The weights were then clamped and the test set run through the trained neural network. For the HIV dataset, given that there are 25 characters to encode as noted earlier, a 25-bit binary string is used to represent each of the 25 characters. Thus a decapptide is converted into a \(10 \times 25 = 250\)-bit binary string. Our FFBP neural network used 250 input units, 20 hidden units (after several trial runs which indicated that 20 hidden units produced good results), and one output unit, using the same training regime as for the HIV dataset. No physical properties are encoded at all. Previous work (Brusic et al., 1994) showed that neural networks could give the same result no matter whether physical properties are represented or not.

Backpropagation with momentum (Fausett, 1994) is used in the training, where initialization weights are randomly selected between \(-1,1\). In the training set, the output of positive sequences is 1 and the output of negative sequences is 0. In the test set, the sequences are predicted as positive if the output is greater than 0.5 and predicted as negative if the output is less than 0.5. 300 training cycles were sufficient for the HIV training, while for the HCV NS3 training 500 training cycles were used. In both cases, the root mean square error of the neural networks was below 0.001. Neural networks are data sensitive, that is, different results are produced when same coding principle has been applied to the HIV and HCV datasets, following Cai and Chou (1998), except that each amino acid has its own 20 bit representation for HIV sequences (bottom half of diagram) and 25 bit representation for HCV ‘residues’. The ANN for HCV consisted of 250 input units (10 amino acids by 25 bits each), 20 hidden units and one output unit (0 for noncleavage, 1 for cleavage). The ANN for HIV consisted of 160 input units (8 amino acids by 20 bits each), 8 hidden units and one output unit.
different data is given to the model. For that reason, ten randomly selected training and test sets were used to avoid idiosyncratic results. Neural networks are also initially state sensitive. So three random initializations were given to each training data set, from which three prediction results were obtained for each test data set subsequently.

**DISCUSSION AND CONCLUSION**

Tables 1A and 1B show all the results of the two FFBP neural networks for the HIV and HCVNS3 test (unseen) samples. The accuracy, sensitivity and specificity values (see Table 1 for an explanation of these terms) in each column of Table 1 are the mean of these three. In the calculation of sensitivity and specificity value, a sequence was considered as correctly predicted only if it was correctly predicted all three times, otherwise, it was considered wrongly predicted.

The accuracy of the FFBP neural network for HIV protease is 92.50%, very similar to that of previous work (92.06%) (Cai and Chou, 1998). For the HCV NS3 protease, the FFBP network produced high accuracy, 96.25%, and for the ten test sets, the standard deviation is 0.93%. This demonstrates that artificial neural networks provide powerful and reliable techniques for predicting HCV NS3 protease cleavage sites.

Unfortunately, the reasons for neural network predictions are locked within the ANN. We ran See5 on the same ten training and test datasets (just once, since there is no initialisation problem) to see how a symbolic machine learning approach compared with the ANN approach and if rule-based knowledge could be derived. Overall, See5 was not as accurate or sensitive as the ANN but did demonstrate comparable specificity (i.e. See5 identified good rules for non-cleavage). For the HIV data, See5’s accuracy of 85.47% was not as good as the FFBP ANN’s accuracy of 92.5%. Similarly, for the HCV data, the FFBP ANN’s accuracy of 96.25% was better than See5’s 90.87%. Also, the standard deviation for the ANNs was better than for See5, indicating fewer variations in performance across runs.

When See5 was run on the total HCV data set (i.e. all data were training data with no test data) to maximise extracted knowledge, the following rules emerged (true positives/false positives in parentheses; positions 1–6 represent the left hand sequence in front of the cleavage point and positions 7–10 the right hand sequence): (a) If position6 is Cysteine then cleavage. (133/27). (b) If position6 is Threonine and position4 is Valine then cleavage. (28/5) (c) If position6 is Cysteine and position7 is Serine then cleavage. (100/33) (d) If position1 is Aspartate then cleavage (122/41). (e) If position10 is Tyrosine then cleavage (98/22). (f) If position10 is Leucine then cleavage (70/27). The importance of residues in positions 1, 6 and 7 can be found in previously published literature (Urbani et al., 1997; Kolykhalov et al., 1994; Bartenschlager et al., 1995). However, the importance of residues in positions 4 and 10 has not been previously reported, to our knowledge.

For the total HIV dataset, the following rules were derived: (a) If position4 is Phenylalanine then cleavage. (35/5) (b) If position4 is Leucine then cleavage. (38/9) (c) If position4 is Serine then non-cleavage. (26/1) (d) If position4 is Tyrosine and position5 is Proline then cleavage. (32/5) Other minor rules covering fewer cases tended to reflect the importance of positions 4 and 5 (on either side of the cleavage site). However, none of the rules were successful in capturing the majority of cases (114 positive sequences in total). One interesting piece of new knowledge extracted by See5 was the relative importance of position 6 (If position6 is Glutamate then cleavage (44/8)). Also, the above rules provide evidence that hydrophobic residues phenylalanine and tyrosine are involved in cleavage site prediction (rules (a) and (d)). Interestingly, the importance of serine and leucine close to the cleavage point is shared with the Rous sarcoma virus (RSC), which also has an octapeptide recognition substrate. However, for RSC, it appears that cleavage requires both serine and leucine immediately on either side of the cleavage point, in that order (Weber and Harrison, 1996), whereas HIV, according to rules (b) and (c) above, requires just leucine immediately in front of the cleavage point.

See5 fared much better when asked to produce rules with subsets of values for conditions rather than individual values. For HCV, the following rule was produced: If position6 in {T,U,C} and position7 in {Q,W,E,Y,U,I,O,A,S,D,H,J,Z,X,V,B,N,M} then cleavage (131/6). (See the table in Figure 2 for the mapping from characters to amino acids.) For the HIV data, the following rule was produced: If position4 in {L,M,F,Y} and position5 in {A,R,E,G,H,I,L,M,F,P,T,W,Y,V} then cleavage (114/18), which covers all cleavage cases but with 18 false positives.

The results demonstrate the need for machine learning algorithms to take into account combined influence of attributes (all residue positions) on classification. ANNs allow for this by adjusting the weights of all the input nodes (through backpropagation) for each sample to ensure more accurate classification, whereas See5 uses the classic decision tree algorithm which looks for just one attribute (the most important residue position) at each decision point in the tree on which to split the samples. The problem with ANNs is that the knowledge they contain after successful training and testing is locked within the neural network. See5 produces rule-based output but is not as accurate in this domain. Nevertheless, the results produced here are promising, in that they provide a pointer towards future in vitro experiments.
which take into account various patterns of residues and identify the most important positions to mutate in the n-residue regions to test for cleavability.

Finally, the results were obtained with no physical information present in the datasets. That is, both the FFBP networks and See5 returned their classification success rates with no additional information on the biochemistry of amino acids or where the cleavage points were located. See5 found important information on either side of the cleavage point as well as at other sites on the basis of ‘pure’ symbolic pattern recognition, and neural networks returned high predictive rates on the basis of arbitrary binary representations of amino acids. Future research will focus on including more physical information in the representations presented to such machine learning systems and broadening the scope of viral data to include cysteine proteases.

ACKNOWLEDGEMENTS

We would like to thank Dr Andy Dalby of the Bioinformatics Laboratory and the anonymous referees for helpful comments which led to major improvements in the content and presentation of this paper.

REFERENCES


Liu,B.I., Viljoen,G.J., Clarke,I.N. and Lambden,P.R. (1999) Identification of further proteolytic cleavage sites in the Southampton calicivirus polyprotein by expression of the viral


